Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems

Cristina España-Bonet and Josef van Genabith
[UdS & DFKI]

MLP–MomenT Workshop, Miyazaki, Japan

12th May 2018

Motivation

Goal1 Improve MT specially for low-resourced languages

Goal2 (future) Translation with monolingual data
and/or parallel from other pairs

Motivation

Goal1 Improve MT specially for low-resourced languages

Goal2 (future) Translation with monolingual data and/or parallel from other pairs

Framework (Multilingual) NMT

Approach Interlinguality

Resources Multilingual Knowledge Graphs

Outline

1 Introduction

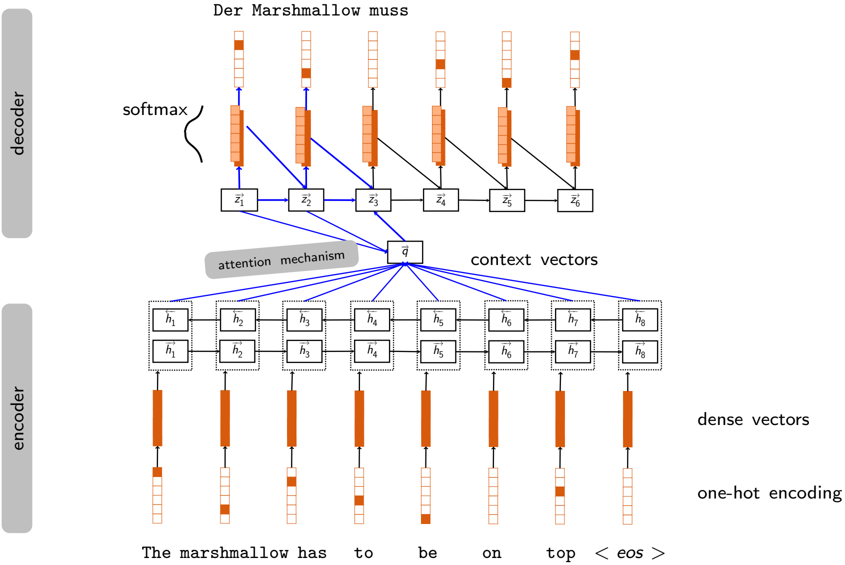
2 Approach

3 Current Results

4 Conclusions

Introduction

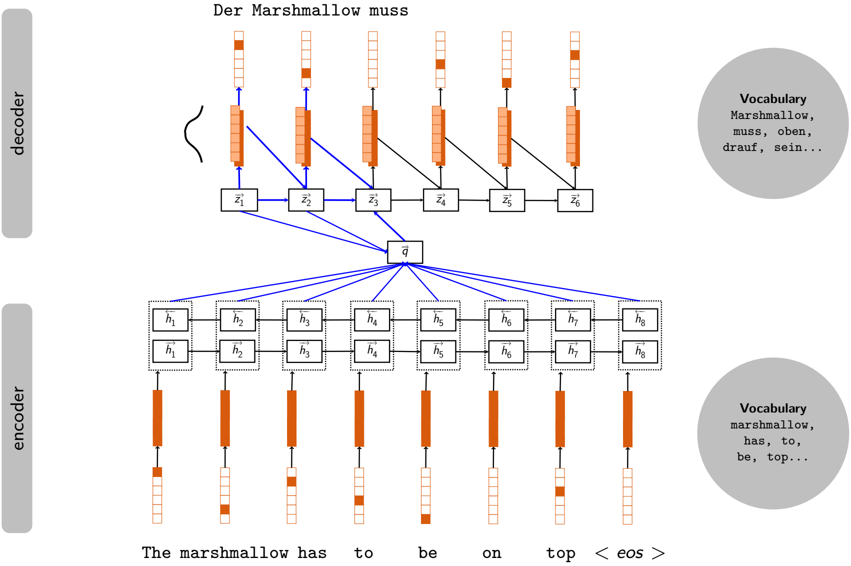
(Multilingual) NMT



decoder
Der Marshmallow muss
softmax
$\vec{z}_1$ $\vec{z}_2$ $\vec{z}_3$ $\vec{z}_4$ $\vec{z}_5$ $\vec{z}_6$
$\vec{q}$
attention mechanism
context vectors
encoder
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$ $\overleftarrow{h}_6$ $\overleftarrow{h}_7$ $\overleftarrow{h}_8$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$ $\overrightarrow{h}_5$ $\overrightarrow{h}_6$ $\overrightarrow{h}_7$ $\overrightarrow{h}_8$
dense vectors
one-hot encoding
The marshmallow has to be on top $< eos >$

Introduction

(Multilingual) NMT



Der Marshmallow muss

Vocabulary
Marshmallow,
muss, oben,
drauf, sein...

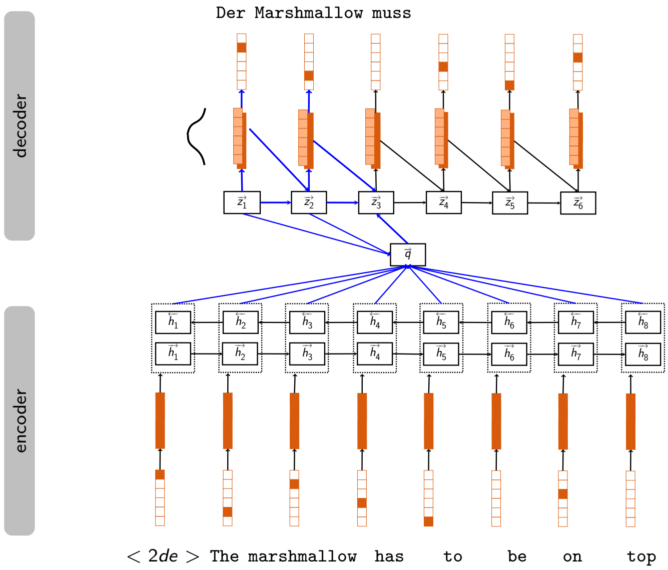
Vocabulary
marshmallow,
has, to,
be, top...

The marshmallow has to be on top $< eos >$

decoder
encoder
$\vec{z}_1$ $\vec{z}_2$ $\vec{z}_3$ $\vec{z}_4$ $\vec{z}_5$ $\vec{z}_6$
$\vec{q}$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$ $\overleftarrow{h}_6$ $\overleftarrow{h}_7$ $\overleftarrow{h}_8$
$\vec{h}_1$ $\vec{h}_2$ $\vec{h}_3$ $\vec{h}_4$ $\vec{h}_5$ $\vec{h}_6$ $\vec{h}_7$ $\vec{h}_8$

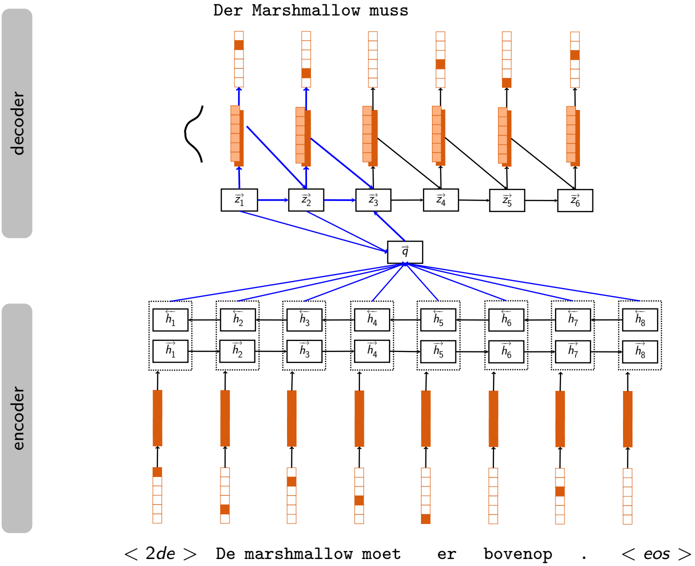
Introduction

(Multilingual) NMT


Der Marshmallow muss
decoder
$\vec{z}_1$ $\vec{z}_2$ $\vec{z}_3$ $\vec{z}_4$ $\vec{z}_5$ $\vec{z}_6$
$\vec{q}$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$ $\overleftarrow{h}_6$ $\overleftarrow{h}_7$ $\overleftarrow{h}_8$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$ $\overrightarrow{h}_5$ $\overrightarrow{h}_6$ $\overrightarrow{h}_7$ $\overrightarrow{h}_8$
encoder
$< 2de >$ The marshmallow has to be on top

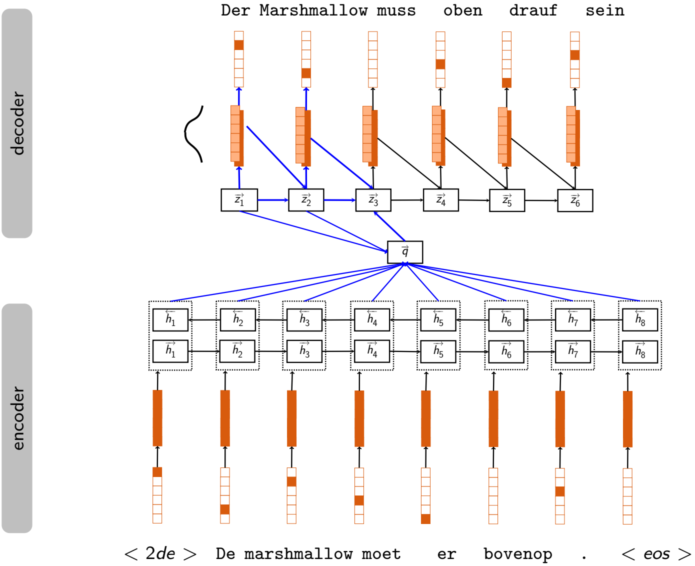
Introduction

(Multilingual) NMT



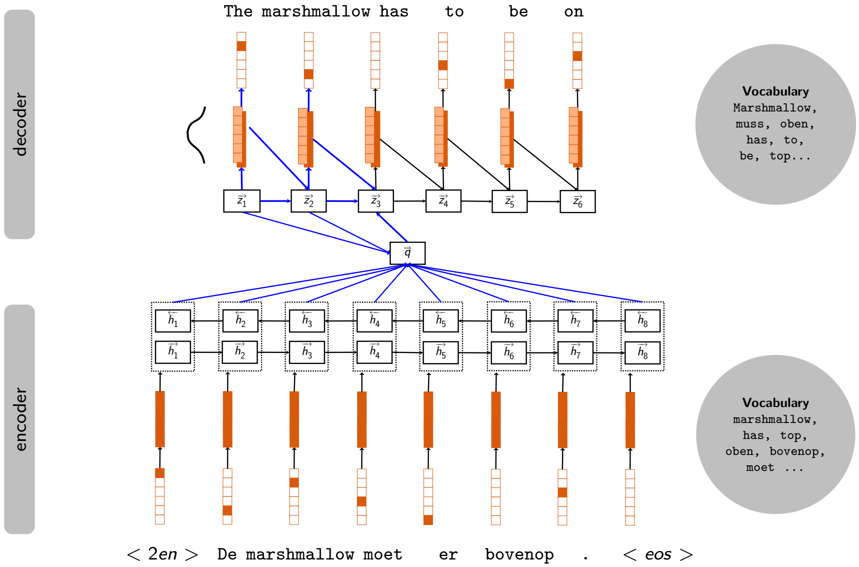
Introduction

(Multilingual) NMT



Introduction

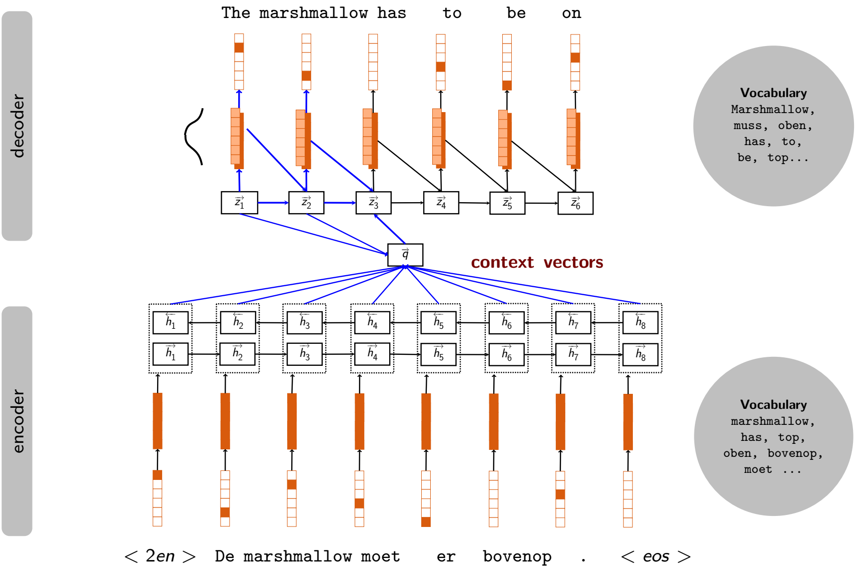
(Multilingual) NMT



decoder
encoder
The marshmallow has to be on
$\vec{z}_1$ $\vec{z}_2$ $\vec{z}_3$ $\vec{z}_4$ $\vec{z}_5$ $\vec{z}_6$
$\vec{q}$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$ $\overleftarrow{h}_6$ $\overleftarrow{h}_7$ $\overleftarrow{h}_8$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$ $\overrightarrow{h}_5$ $\overrightarrow{h}_6$ $\overrightarrow{h}_7$ $\overrightarrow{h}_8$
$< 2en >$ De marshmallow moet er bovenop . $< eos >$
Vocabulary
Marshmallow, muss, oben, has, to, be, top...
Vocabulary
marshmallow, has, top, oben, bovenop, moet ...

Introduction

(Multilingual) NMT



The marshmallow has to be on

context vectors

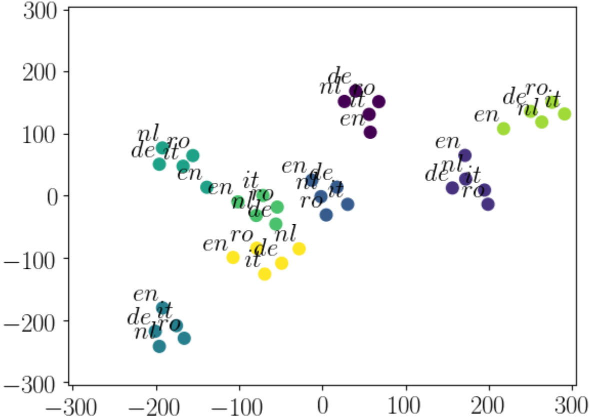
Vocabulary
Marshmallow,
muss, oben,
has, to,
be, top...

Vocabulary
marshmallow,
has, top,
oben, bovenop,
moet ...

$< 2en >$ De marshmallow moet er bovenop . $< eos >$

Introduction

Interlingual Semantic Space



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

Introduction

Purposes of MSNs in NMT

- Help to cluster by semantics

  Improve general translation

- Reduce UNKs with help of other languages

  Help translation of under-resourced languages

- Translation from unseen languages

  Beyond-zero-shot translation

- (semi-)Abstract representation for sentences

  Data-driven “Interlingual” translation

Outline

1 Introduction

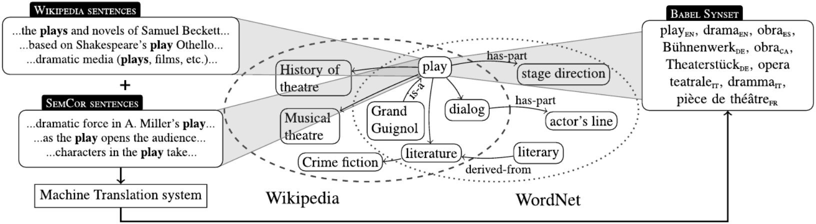
2 Approach

3 Current Results

4 Conclusions

Approach

Multilingual Semantic Network: BabelNet



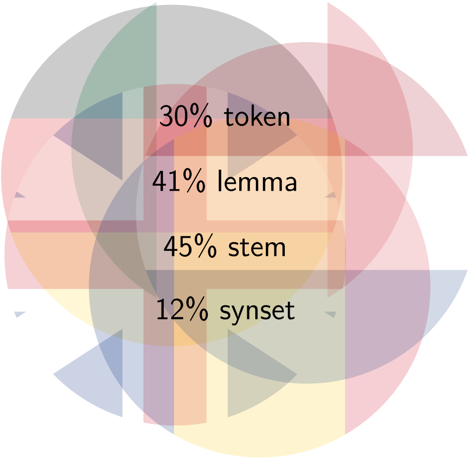
Wikipedia sentences
...the **plays** and novels of Samuel Beckett...
...based on Shakespeare's **play** Othello...
...dramatic media (**plays**, films, etc.)...
+
SemCor sentences
...dramatic force in A. Miller's **play**...
...as the **play** opens the audience...
...characters in the **play** take...
Machine Translation system
History of theatre
play
has-part
stage direction
is-a
Musical theatre
Grand Guignol
dialog
has-part
actor's line
literature
literary
Crime fiction
derived-from
Wikipedia
WordNet
Babel Synset
play$_{EN}$, drama$_{EN}$, obra$_{ES}$, Bühnenwerk$_{DE}$, obra$_{CA}$, Theaterstück$_{DE}$, opera teatrale$_{IT}$, dramma$_{IT}$, pièce de théâtre$_{FR}$

[Navigli & Ponzetto, 2012]

1. Synsets as cross-lingual lexicons (for 271 languages!)
2. Relations among synsets

Approach

Multilingual Corpus: en–de–nl–ro–it TED



30% token

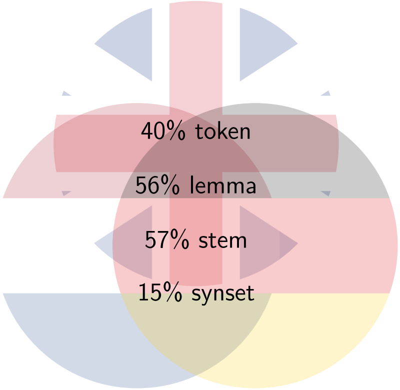
41% lemma

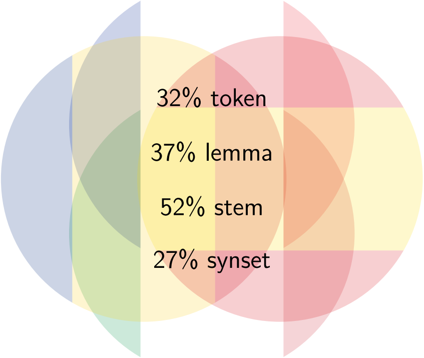
45% stem

12% synset

Approach

Multilingual Corpus: TED, Germanic & Latin Languages





Approach

Simple Approaches: Source Sentence Encodings

- **word2word translation** *w:*

  $< 2en >$ es war ein riesiger Erfolg

  $< 2en >$ è stato un enorme successo

Approach

Simple Approaches: Source Sentence Encodings

- **word2word translation** *w:*

  $< 2en >$ es war ein riesiger Erfolg

  $< 2en >$ è stato un enorme successo

- **synset2word translation** *b:*

  PRONOUN VERB DETERMINER ADJECTIVE bn:15350982n

  bn:00083181v bn:00083181v DETERMINER bn:00102268a bn:00078365n

Approach

Simple Approaches: Synset Annotation

- Enrich content words (nouns —including NEs, foreign words and numerals—, adjectives, adverbs and verbs) with synsets
- Retrieve a synset according to the lemma and PoS of a word
- Select the BabelNet ID according to BabelNet's odering
- Mark negation particles with a tag NEG
- Keep (coarse) PoS for the remaining tokens

Approach

Simple Approaches: Synset Annotation

- Enrich content words (nouns —including NEs, foreign words and numerals—, adjectives, adverbs and verbs) with synsets
- Retrieve a synset according to the lemma and PoS of a word
- Select the BabelNet ID according to BabelNet's odering
- Mark negation particles with a tag NEG
- Keep (coarse) PoS for the remaining tokens **(73% of the corpus!)**

Approach

Simple Approaches: Source Sentence Encodings

- **word2word translation** *w:*

  $< 2en >$ es war ein riesiger Erfolg

  $< 2en >$ è stato un enorme successo

- **synset2word translation** *b:*

  PRONOUN VERB DETERMINER ADJECTIVE bn:15350982n

  bn:00083181v bn:00083181v DETERMINER bn:00102268a bn:00078365n

Approach

Simple Approaches: Source Sentence Encodings

- **word2word translation** *w:*

  $< 2en >$ es war ein riesiger Erfolg

  $< 2en >$ è stato un enorme successo

- **synset2word translation** *b:*

  PRONOUN VERB DETERMINER ADJECTIVE bn:15350982n

  bn:00083181v bn:00083181v DETERMINER bn:00102268a bn:00078365n

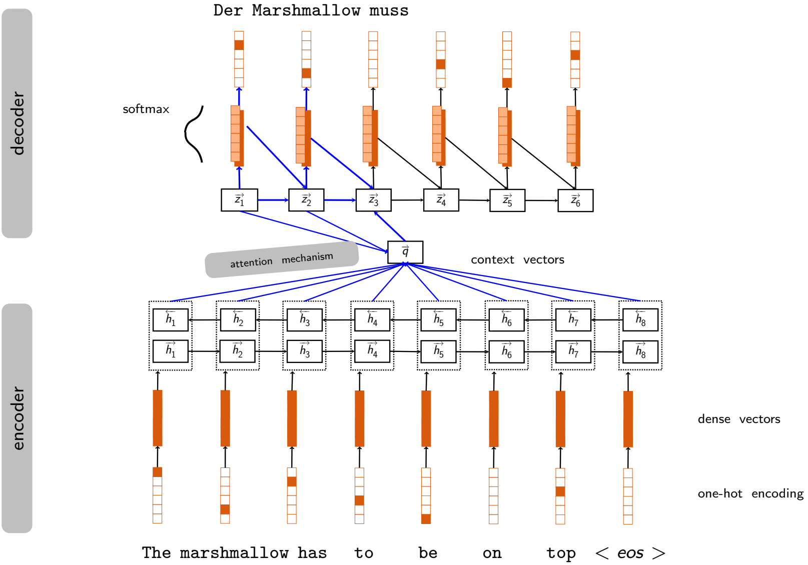
- **factored translation** *wb:*

  $< 2en >$|- es|- war|- ein|- riesiger|- Erfolg|bn:15350982n

  $< 2en >$|- è|bn:00083181v stato|bn:00083181 un|- enorme|bn:00102268a successo|bn:00078365n

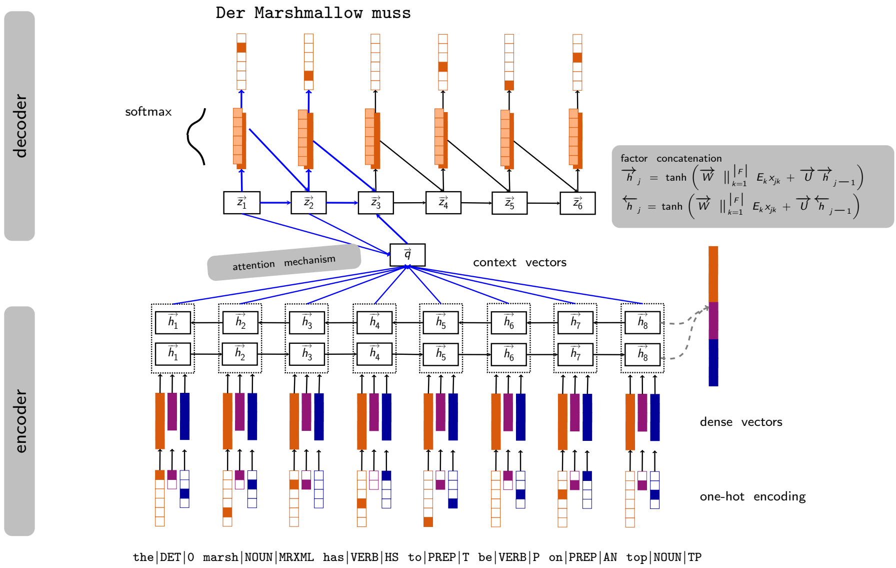
Approach

(Multilingual) NMT



Approach

Factored NMT



factor concatenation

$$\overrightarrow{h}_j = \tanh\left(\overrightarrow{W} \;\|_{k=1}^{|F|}\; E_k x_{jk} + \overrightarrow{U}\,\overrightarrow{h}_{j-1}\right)$$

$$\overleftarrow{h}_j = \tanh\left(\overrightarrow{W} \;\|_{k=1}^{|F|}\; E_k x_{jk} + \overrightarrow{U}\,\overleftarrow{h}_{j-1}\right)$$

Simple Approaches: Source Sentence Encodings

- **word2word translation** *w:*

  $< 2en >$ es war ein riesiger Erfolg

  $< 2en >$ è stato un enorme successo

- **synset2word translation** *b:*

  PRONOUN VERB DETERMINER ADJECTIVE bn:15350982n bn:00083181v bn:00083181v DETERMINER bn:00102268a bn:00078365n

- **factored translation** *wb:*

  $< 2en >$|- es|- war|- ein|- riesiger|- Erfolg|bn:15350982n

  $< 2en >$|- è|bn:00083181v stato|bn:00083181 un|- enorme|bn:00102268a successo|bn:00078365n

Current Results

ML-NMT System Characteristics

- **Data**

  Corpus for *w* and *wb*, *all2all* (2,113,917 parallel fragments)

  Corpus for *b*, *bn2en* (868,226 parallel fragments)

Current Results

ML-NMT System Characteristics

- **Data**

  Corpus for *w* and *wb*, *all2all* (2,113,917 parallel fragments)

  Corpus for *b*, *bn2en* (868,226 parallel fragments)

- **Architecture**

  Nematus with 150k Vocab. (+2k BPE for *w* and *wb*)

  Adadelta, 506D embeddings, 800 hidden units, mini-batch size 100

  Robust results with 4-ensemble

Current Results

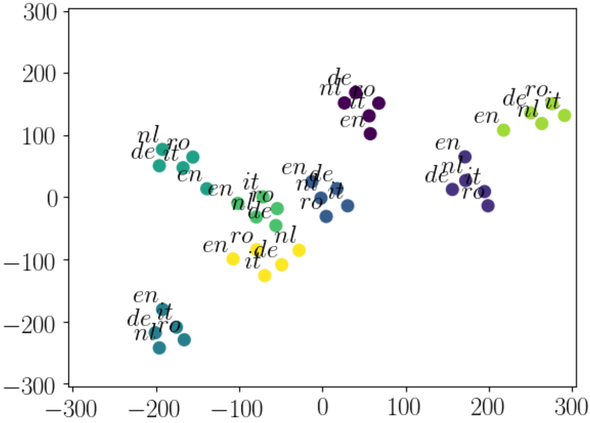
Automatic Evaluation, ML System

TED ML-NMT test set 2010

	BLEU			METEOR		
	w	*wb*	*b*	*w*	*wb*	*b*
de2en	**32.6**	**33.0**	17.5	33.1	**33.5**	24.2
it2en	**33.5**	**33.2**	21.4	**33.9**	**34.0**	27.4
nl2en	36.2	**36.6**	15.0	34.7	**34.9**	21.5
ro2en	**34.3**	**34.8**	19.6	**34.4**	**34.6**	25.9

Current Results

Automatic Evaluation, ML System



- *Too* ideal corpus!

Current Results

Automatic Evaluation, IL System

TED ML-NMT test set 2010

	BLEU			METEOR		
	w	*wb*	*b*	*w*	*wb*	*b*
de2en	**32.6**	**33.0**	17.5	33.1	**33.5**	24.2
it2en	**33.5**	**33.2**	21.4	**33.9**	**34.0**	27.4
nl2en	36.2	**36.6**	15.0	34.7	**34.9**	21.5
ro2en	**34.3**	**34.8**	19.6	**34.4**	**34.6**	25.9
fr2en	2.4	5.1	**7.3**	11.2	16.7	**17.5**
es2en	3.1	6.7	**11.3**	12.0	18.4	**20.7**

Current Results

Automatic Evaluation, IL System

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**

Current Results

Automatic Evaluation, IL System

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**

- Ambiguity issues (+ multilingual ambiguity!)
- Lot of room for improvement

Outline

1 Introduction

2 Approach

3 Current Results

4 Conclusions

Conclusions

Summary

- We have presented 2 ways of using **semantic information** (*synsets*) in a seq2seq architecture
- Improvements come specially in **beyond-zero-shot translation**
- The weakest detected point is the current need for **multilingual equivalence** and **multilingual WSD**
- The same methodology can be used in any **seq2seq** architecture (extension to CL-QA)

Conclusions

Future Work

Goal1 Improve MT specially for low-resourced languages

Goal2 (future) Translation with monolingual data
and/or parallel from other pairs

Conclusions

Future Work

Goal1 Improve MT specially for low-resourced languages

Goal2 (future) Translation with monolingual data
and/or parallel from other pairs

Data-driven "Interlingual" Translation

- Multilingual Word Sense Disambiguation
- Interlingual representation for non-content words

Thanks!

Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems

Cristina España-Bonet and Josef van Genabith
[UdS & DFKI]

MLP–MomenT Workshop, Miyazaki, Japan

12th May 2018

Extra Slides

BabelNet

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

Multilingual Corpus: TED in numbers

	West Germanic Languages			Latin Languages			
	en	*de*	*nl*	*ro*	*it*	*es*	*fr*
Sentences	545,270	303,668	444,287	225,980	513,693	151,631	140,717
Tokens	9,768,374	5,148,199	6,894,438	3,732,679	8,367,940	2,494,336	2,473,040
uToken	141,013	221,459	187,148	213,670	200,697	148,366	131,015
uLemma	73,048	101,003	85,846	72,535	52,525	52,052	53,088
uStem	50,128	94,126	85,560	54,227	44,691	35,307	40,504
uM3	57,630	79,029	60,534	30,576	32,828	31,840	32,234
uBN	28,445	34,022	27,720	24,375	27,172	23,567	23,856

Extra Slides

Multilingual Corpus: TED in numbers

	BabelNet				TED corpus	
Language (iso code)	Lemmas	Synsets	Senses	Synonym/Synset	Synsets	Coverage (%)
English (*en*)	11,769,205	6,667,855	17,265,977	2.59	28,445	27.25
French (*fr*)	5,301,989	4,141,338	7,145,031	1.73	–	–
German (*de*)	5,109,948	4,039,816	6,864,767	1.70	34,022	23.50
Spanish (*es*)	5,022,610	3,722,927	6,490,447	1.74	–	–
Dutch (*nl*)	4,416,028	3,817,696	6,456,175	1.69	27,720	26.25
Italian (*it*)	4,087,765	3,541,031	5,423,837	1.53	27,172	29.00
Romanian (*ro*)	3,009,318	2,697,720	3,384,256	1.25	24,375	27.25