Resolving OOVs with BWE in MT

Cristina España-Bonet

Universitat Politècnica de Catalunya
TALP Research Center

Saarland University, DFKI – Saarbrücken

29th April 2016

The talk

Content with keywords!

Resolving Out-of-Vocabulary Words with Bilingual Word Embeddings in Machine Translation

Joint Work with Pranava Swaroop Madhyastha

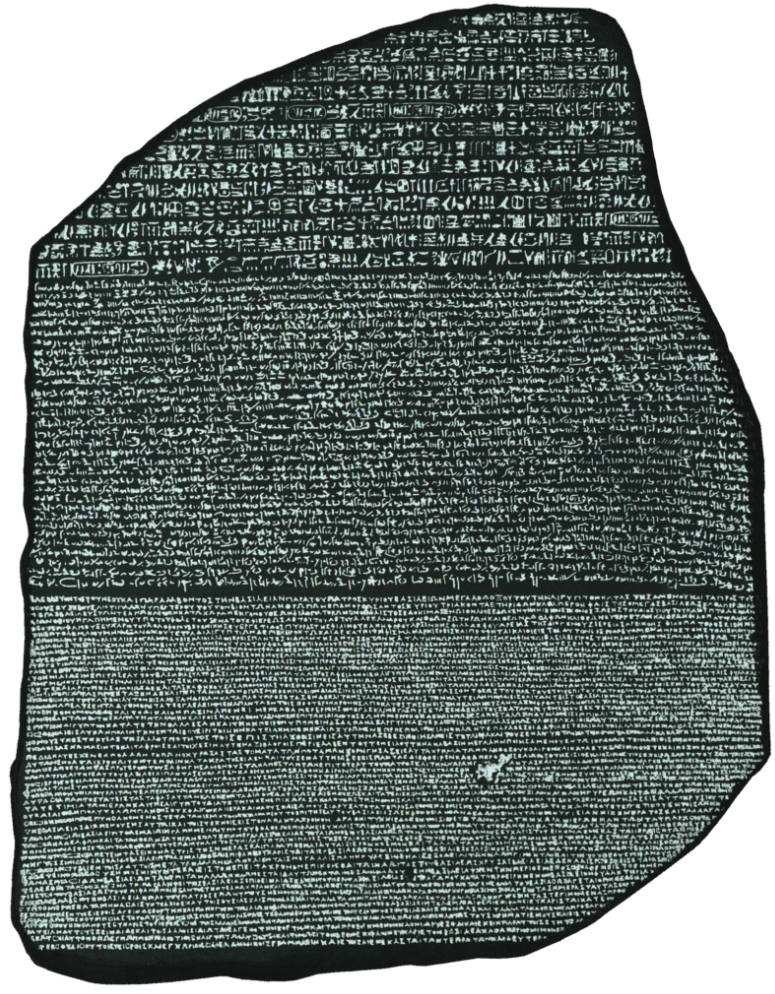
The talk

Contents

Motivation

Vocabulary in Empirical MT

Empirical MT relies on aligned corpora



Motivation

Looking for Corpora

It is known that astrocytes are plastic, enslaving their functions to the requirements of the neurons to which they are related.

Motivation

Looking for Corpora

It is known that astrocytes are plastic, enslaving their functions to the requirements of the neurons to which they are related.

- Biomedical corpus: **OOV** – enslaving

Motivation

Looking for Corpora

`It is known that astrocytes are plastic, enslaving their functions to the requirements of the neurons to which they are related.`

- Biomedical corpus: OOV – enslaving
- "General" corpus: **OOV** – astrocytes

Motivation

Example: En2Es, a rich resourced pair

Corpus	**Vocab**	**enslaving**	**astrocytes**
Biomedical	$0.3 \cdot 10^6$	0	67
Quest	$0.5 \cdot 10^6$	34	0

Quest: EP+UN+NC
`http://www.quest.dcs.shef.ac.uk/wmt13_qe.html`

Motivation

Example: En2Es, a rich resourced pair

Corpus	**Vocab**	**enslaving**	**astrocytes**
Biomedical	$0.3 \cdot 10^6$	0	67
Quest	$0.5 \cdot 10^6$	34	0
En–Wikipedia	$2.0 \cdot 10^6$	487	606
Es–Wikipedia	$0.8 \cdot 10^6$	62^+	102

Quest: EP+UN+NC
`http://www.quest.dcs.shef.ac.uk/wmt13_qe.html`

Motivation

Moral

Take advantage of monolingual corpora

- Extract lexicons (word pairs)
- Extract parallel corpora (sentence pairs)

Motivation

Moral

Take advantage of monolingual corpora

- Extract lexicons (word pairs)
- Extract parallel corpora (sentence pairs)

And in this talk

- Use (bilingual) embeddings
- Is it always possible? No, but Wikipedia is a multilingual, multidomain and continuously growing corpus!

Motivation

Moral

Take advantage of monolingual corpora

- Extract lexicons (word pairs) **99% of the talk**
- Extract parallel corpora (sentence pairs) **1% of the talk**

And in this talk

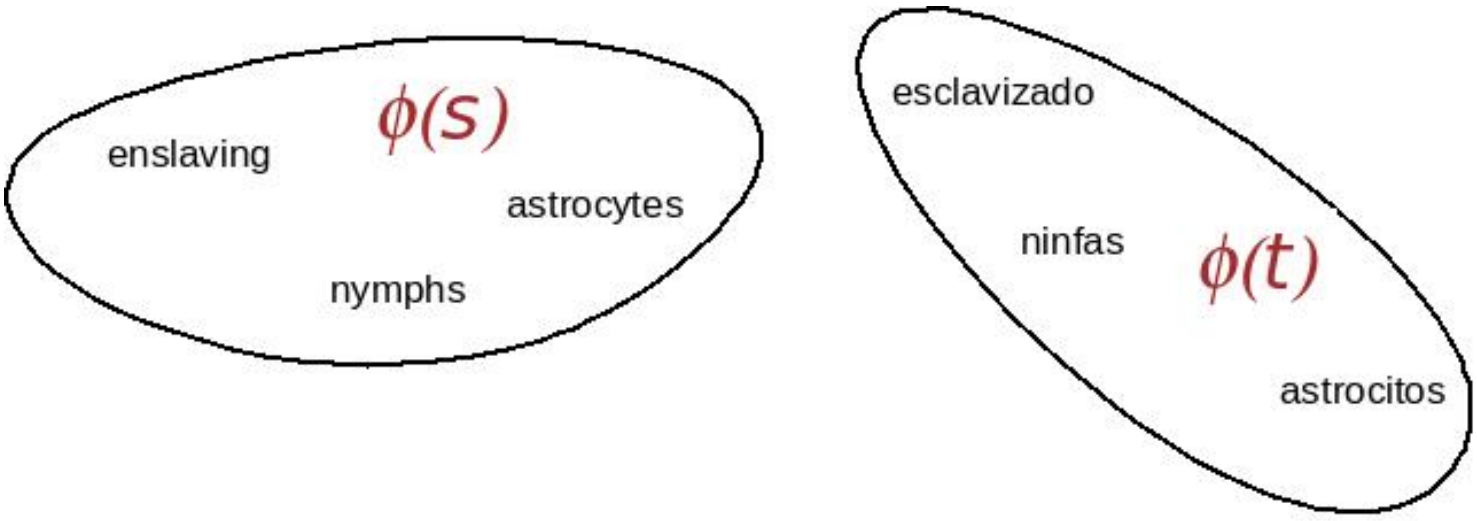
- Use (bilingual) embeddings
- Is it always possible? No, but Wikipedia is a multilingual, multidomain and continuously growing corpus!

Embeddings for Machine Translation

Contents

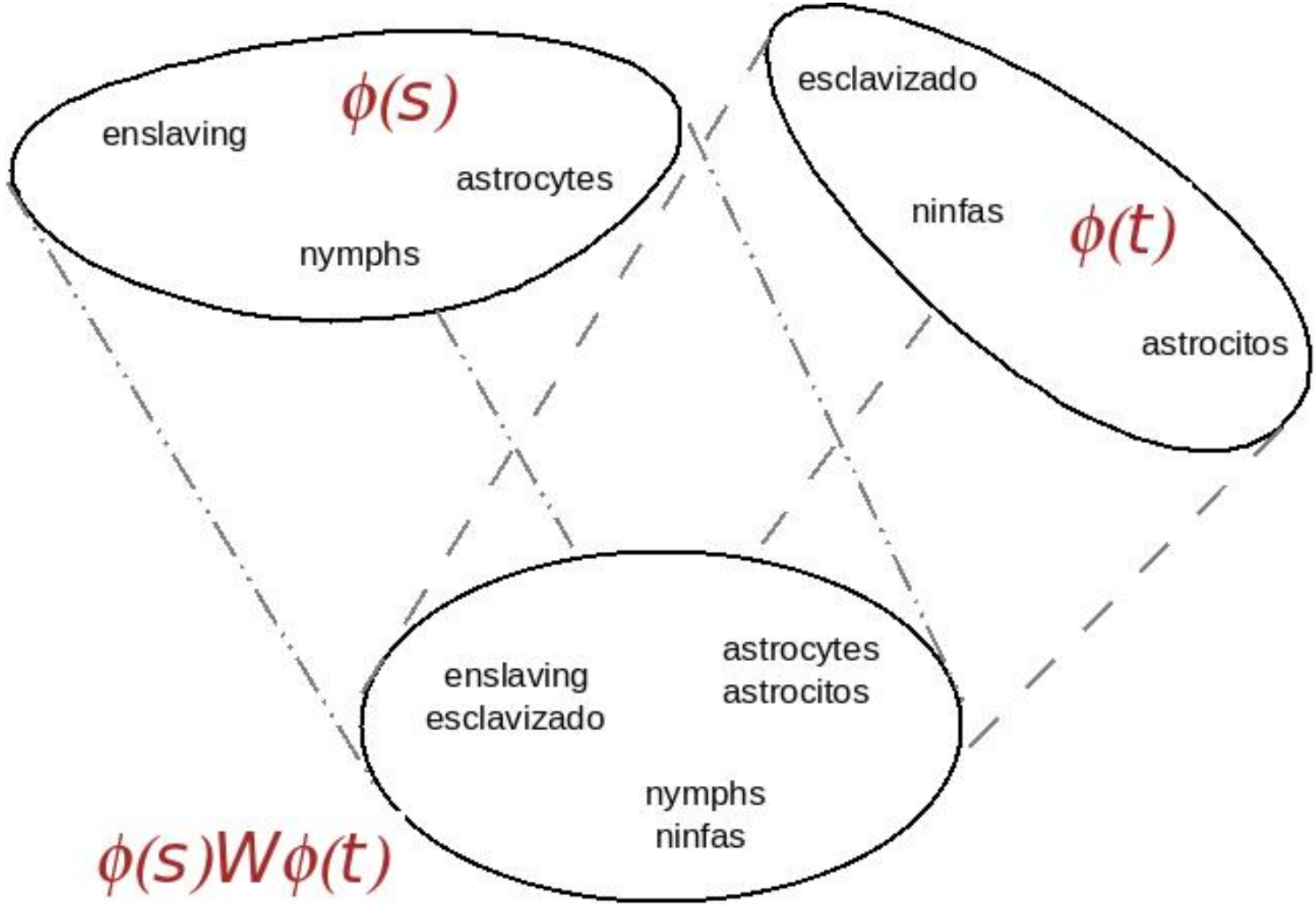
Embeddings for Machine Translation

Key: Bilingual Embeddings



Embeddings for Machine Translation

Key: Bilingual Embeddings



Embeddings for Machine Translation

The Probabilistic Model

SMT, log-linear model

$$P(t|s) \sim \exp\left\{\sum \lambda_m \, h_m(s,t)\right\}$$

$$\lambda_m \rightarrow \text{MERT}$$

Embeddings for Machine Translation

The Probabilistic Model

SMT, log-linear model

$$P(t|s) \sim \exp\left\{\sum \lambda_m \, h_m(s,t)\right\}$$

$$\lambda_m \rightarrow \text{MERT}$$

BWE, bilinear model

$$P(t|s) \sim \exp\left\{\phi(s)^\top W \phi(t)\right\}$$

$$W \rightarrow \text{FOBOS}$$

Embeddings for Machine Translation

Bilinear Model for BWE

$$\Pr(t|s; W) = \frac{\exp\left\{\phi(s)^\top W \phi(t)\right\}}{\sum_{t'} \exp\left\{\phi(s)^\top W \phi(t)\right\}}$$

Objective function: log-likelihood with regularisation

$$-\sum_{t,s} \log \Pr(t|s; W) + \lambda \|W\|_*$$

Dataset: (small) bilingual dictionary

Nice properties

- Nuclear norm $\|W\|_*$ used so
  - dimensionality reduction and,
  - projection to the reduced space

Nice properties

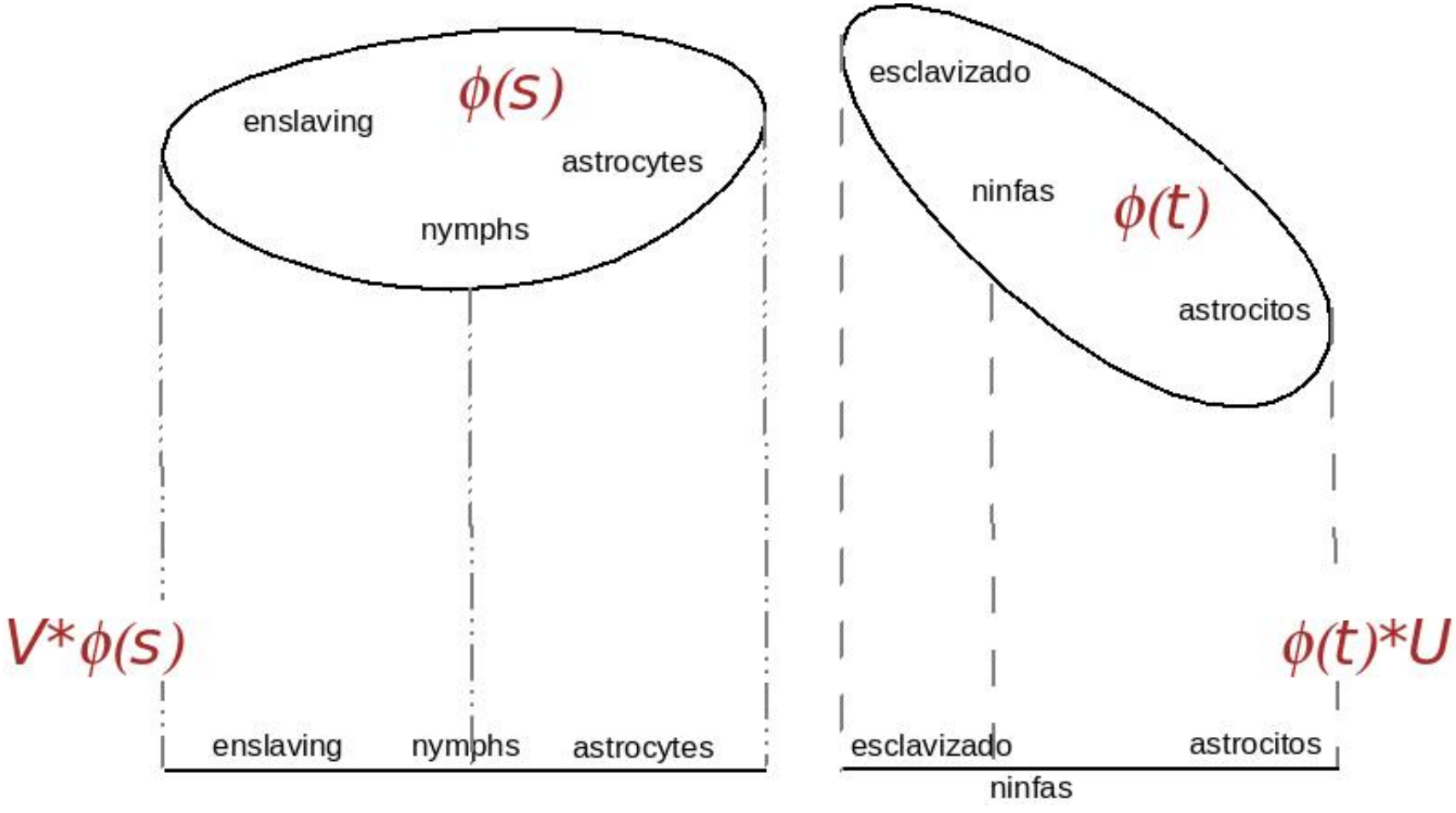
- Nuclear norm $\|W\|_*$ used so
  - dimensionality reduction and,
  - projection to the reduced space

$$W = U\Sigma V^\top \qquad \text{(Singular Value Decomposition)}$$

$$\phi(t)^\top W \phi(s) = [\phi(t)^\top U]\ \Sigma\ [V^\top \phi(s)]$$

Embeddings for Machine Translation

Compressed Embeddings



Embeddings for Machine Translation

Contents

Embeddings for Machine Translation

Resolving OOVs, a First Experiment

Assumption:

- Translation pairs will be better coming from parallel corpora than from monolingual ones

Consequence:

- Use bilingual embeddings for unseen words

Embeddings for Machine Translation

Resolving OOVs, a First Experiment

Assumption:

- Translation pairs will be better coming from parallel corpora than from monolingual ones

Consequence:

- Use bilingual embeddings for unseen words

Confirmed with experiments ✓

Embeddings for Machine Translation

The WE Setting

English & Spanish

Monolingual corpora:	Wikipedia + Quest
Number of words:	$2.3 \cdot 10^9$ English; $0.8 \cdot 10^9$ Spanish
Embeddings:	$2.0 \cdot 10^6$ English; $0.8 \cdot 10^6$ Spanish (word2vec, cbow, 300 dimensions)
Dictionary:	Apertium bilingual dictionary
Number of words:	34,806 (training and validation)

Embeddings for Machine Translation

The Output

For a source word s, a list of words t ranked according to $\phi(t)^T W \phi(s)$

galaxy $\Rightarrow$ galaxia $*$ planeta $*$ galaxias $*$ enana $*$ planetas $*$ universo $*$ supernova $*$ nebulosa $*$ Láctea $*$ Galaxia $*$ Gliese $*$ constelación $*$ NGC $*$ galáctico $*$ cúmulo $*$ cúmulos $*$ asteroide $*$ orbitando $*$ interestelar $*$ orbita $*$ Centauri $*$ estrella $*$ órbita $*$ espectral $*$ meteorito $*$ alienígena $*$ Tierra $*$ estelar $*$ nube $*$ Universo $*$ Decepticons $*$ mundos $*$ neutrinos $*$ asteroides $*$ Nebulosa $*$ extraterrestre $*$ alienígenas $*$ extrasolar $*$ ...

Embeddings for Machine Translation

The Analysis

Things we can do

galaxy	nymphs
galaxia	**ninfas**
planeta	ninfa
universo	crías
planetas	diosa
galaxias	dioses
...	...

Embeddings for Machine Translation

The Analysis

Things we can do

galaxy	nymphs
galaxia	**ninfas**
planeta	ninfa
universo	crías
planetas	diosa
galaxias	dioses
...	...

Things we cannot do

Stuart	folksong
William	música
Henry	**folclore**
John	literatura
Charles	himno
Thomas	folklore
Estuardo (#48)	**canción**

Embeddings for Machine Translation

Test Sets & Coverage

	Seg.	**Tokens**
NewsDev	3003	72988
NewsTest	3000	64810
WikiTest	500	11069

Embeddings for Machine Translation

Test Sets & Coverage

	Seg.	**Tokens**
NewsDev	3003	72988
NewsTest	3000	64810
WikiTest	500	11069

Our bilingual models cover 97% of the words in the test sets used for MT

Embeddings for Machine Translation

Coverage & Percentage

	Seg.	**Tokens**	**OOV$_{\text{all}}$**	**OOV$_{\text{CW}}$**
NewsDev	3003	72988	1920 (2.6%)	378 (0.5%)
NewsTest	3000	64810	1590 (2.5%)	296 (0.5%)
WikiTest	500	11069	798 (7.2%)	201 (1.8%)

- OOV$_{\text{all}}$: words absent in the Europarl corpus ($0.1 \cdot 10^6$)

Embeddings for Machine Translation

Coverage & Percentage

	Seg.	**Tokens**	**OOV$_{\text{all}}$**	**OOV$_{\text{CW}}$**
NewsDev	3003	72988	1920 (2.6%)	378 (0.5%)
NewsTest	3000	64810	1590 (2.5%)	296 (0.5%)
WikiTest	500	11069	798 (7.2%)	201 (1.8%)

- OOV$_{\text{all}}$: words absent in the Europarl corpus ($0.1 \cdot 10^6$)
- OOV$_{\text{CW}}$: Content Words, excludes NEs

Embeddings for Machine Translation

The MT Setting

English-to-Spanish

Parallel corpus:	Europarl
Number of words:	$54 \cdot 10^6$ English; $57 \cdot 10^6$ Spanish
Monolingual corpus:	Wikipedia + Quest
Number of words:	$2.3 \cdot 10^9$ English; $0.8 \cdot 10^9$ Spanish (used for BLM)
Translation model:	PBSMT with Moses
Language model:	5-gram, interpolated Kneser-Ney discounting
Optimisation:	MERT against BLEU

Embeddings for Machine Translation

The MT Setting II

OOVs with Moses

- Default: verbatim output
- Alternative: remove OOVs in the output

Embeddings for Machine Translation

The MT Setting II

OOVs with Moses

- Default: verbatim output
- Alternative: remove OOVs in the output

Including translation options for the OOVs

- Via xml markup into the source test set
- Equivalent to add phrase-pairs in the translation table with equal probability for every feature
- No interaction with other phrases in the translation table

Embeddings for Machine Translation

System Definition

Systems

noOOV
verbatimOOV

Embeddings for Machine Translation

System Definition

Systems

noOOV
verbatimOOV
$\text{BWE}_{\text{all50}}$
BWE_{CW50}
BWE_{CW10}

Embeddings for Machine Translation

System Definition

Systems

noOOV
verbatimOOV
$\text{BWE}_{\text{all50}}$
BWE_{CW50}
BWE_{CW10}
BLM
$\text{BLM+BWE}_{\text{all50}}$
$\text{BLM+BWE}_{\text{CW50}}$
$\text{BLM+BWE}_{\text{CW10}}$

Embeddings for Machine Translation

Automatic Evaluation

NewsTest

	TER	BLEU	MTR
noOOV	58.21	21.94	45.79
verbatimOOV	57.90	22.89	47.06

Embeddings for Machine Translation

Automatic Evaluation

NewsTest

	TER	BLEU	MTR
noOOV	58.21	21.94	45.79
verbatimOOV	57.90	22.89	47.06
$\text{BWE}_{\text{all50}}$	58.33	22.23	45.76
BWE_{CW50}	57.66	23.09	47.14
BWE_{CW10}	57.85	23.06	47.11

Embeddings for Machine Translation

Automatic Evaluation

NewsTest

	TER	BLEU	MTR
noOOV	58.21	21.94	45.79
verbatimOOV	57.90	22.89	47.06
$\text{BWE}_{\text{all50}}$	58.33	22.23	45.76
BWE_{CW50}	57.66	23.09	47.14
BWE_{CW10}	57.85	23.06	47.11
BLM	55.37	25.83	**49.19**

Embeddings for Machine Translation

Automatic Evaluation

NewsTest

	TER	BLEU	MTR
noOOV	58.21	21.94	45.79
verbatimOOV	57.90	22.89	47.06
$\text{BWE}_{\text{all50}}$	58.33	22.23	45.76
BWE_{CW50}	57.66	23.09	47.14
BWE_{CW10}	57.85	23.06	47.11
BLM	55.37	25.83	**49.19**
BLM+$\text{BWE}_{\text{all50}}$	55.89	24.92	47.84
BLM+BWE_{50}	55.55	25.61	49.01
BLM+BWE_{10}	**55.31**	**25.86**	49.04

Embeddings for Machine Translation

Automatic Evaluation

0.5% OOV_{CW}

NewsTest

	TER	BLEU	MTR
noOOV	58.21	21.94	45.79
verbatimOOV	57.90	22.89	47.06
$\text{BWE}_{\text{all50}}$	58.33	22.23	45.76
BWE_{CW50}	57.66	23.09	47.14
BWE_{CW10}	57.85	23.06	47.11
BLM	55.37	25.83	**49.19**
BLM+$\text{BWE}_{\text{all50}}$	55.89	24.92	47.84
BLM+BWE_{50}	55.55	25.61	49.01
BLM+BWE_{10}	**55.31**	**25.86**	49.04

Embeddings for Machine Translation

Automatic Evaluation

	NewsTest			WikiTest		
	TER	BLEU	MTR	TER	BLEU	MTR
noOOV	58.21	21.94	45.79	61.26	16.24	38.76
verbatimOOV	57.90	22.89	47.06	58.55	21.90	45.77
$\text{BWE}_{\text{all50}}$	58.33	22.23	45.76	58.38	21.96	44.84
BWE_{CW50}	57.66	23.09	47.14	56.19	24.16	48.49
BWE_{CW10}	57.85	23.06	47.11	55.64	24.71	49.05
BLM	55.37	25.83	**49.19**	52.60	30.63	51.04
BLM+$\text{BWE}_{\text{all50}}$	55.89	24.92	47.84	51.02	32.20	52.09
BLM+BWE_{50}	55.55	25.61	49.01	49.50	33.94	54.93
BLM+BWE_{10}	**55.31**	**25.86**	49.04	**49.12**	**34.58**	**55.52**

Embeddings for Machine Translation

Automatic Evaluation

1.8% OOV_{CW}

	NewsTest			WikiTest		
	TER	BLEU	MTR	TER	BLEU	MTR
noOOV	58.21	21.94	45.79	61.26	16.24	38.76
verbatimOOV	57.90	22.89	47.06	58.55	21.90	45.77
$\text{BWE}_{\text{all50}}$	58.33	22.23	45.76	58.38	21.96	44.84
BWE_{CW50}	57.66	23.09	47.14	56.19	24.16	48.49
BWE_{CW10}	57.85	23.06	47.11	55.64	24.71	49.05
BLM	55.37	25.83	**49.19**	52.60	30.63	51.04
$\text{BLM+BWE}_{\text{all50}}$	55.89	24.92	47.84	51.02	32.20	52.09
BLM+BWE_{50}	55.55	25.61	49.01	49.50	33.94	54.93
BLM+BWE_{10}	**55.31**	**25.86**	49.04	**49.12**	**34.58**	**55.52**

Embeddings for Machine Translation

Manual Evaluation

Figures for BLM+BWE$_{\mathrm{CW50}}$ on WikiTest

- Accuracy in validation: $\sim$ 82% (on the dictionary)
- Accuracy@50: 68%
- OOVs translated correctly: 22%
- Absolute numbers: 45 OOVs / 11069 tokens improved

Embeddings for Machine Translation

Manual Evaluation

Figures for BLM+BWE$_{\mathrm{CW50}}$ on WikiTest

- Accuracy in validation: $\sim$ 82% (on the dictionary)
- Accuracy@50: 68%
- OOVs translated correctly: 22%
- Absolute numbers: 45 OOVs / 11069 tokens improved

That's 3.3 point of BLEU improvement?!

Embeddings for Machine Translation

Manual Evaluation II

`However, at the end of the movie there is a wedding where some girls sing an old` **folksong** `which reveals the` **murder-mystery**`.`

folksong

música (0.026) ∗ folclore (0.026) ∗ literatura (0.025) ∗ himno (0.024) ∗ folklore (0.023) ∗ poema (0.022) ∗ canción (0.022) ∗ lengua (0.022) ∗ poesía (0.021)

murder-mystery

thriller (0.026) ∗ comedia (0.026) ∗ novela (0.025) ∗ novelas (0.024) ∗ drama (0.023) ∗ película (0.023) ∗ ambientada (0.022) ∗ ficción (0.022) ∗ sátira (0.021)

Embeddings for Machine Translation

Manual Evaluation III

`However, at the end of the movie there is a wedding where some girls sing an old` **folksong** `which reveals the` **murder-mystery** `.`

BLM: Sin embargo, al final de la película hay una boda donde algunas niñas cantan un viejo **folksong** que revela el **murder-mystery**.

Embeddings for Machine Translation

Manual Evaluation III

`However, at the end of the movie there is a wedding where some girls sing an old` **folksong** `which reveals the` **murder-mystery** `.`

BLM: Sin embargo, al final de la película hay una boda donde algunas niñas cantan un viejo **folksong** que revela el **murder-mystery**.

BLM+BWE: Sin embargo, al final de la película hay una boda donde algunas niñas cantan una vieja **historia** que revela el **misterio**.

Embeddings for Machine Translation

Manual Evaluation III

`However, at the end of the movie there is a wedding where some girls sing an old` **folksong** `which reveals the` **murder-mystery** `.`

BLM: Sin embargo, al final de la película hay una boda donde algunas niñas cantan un viejo **folksong** que revela el **murder-mystery**.

BLM+BWE: Sin embargo, al final de la película hay una boda donde algunas niñas cantan una vieja **historia** que revela el **misterio**.

BLM+BWE$_{\mathrm{CW10}}$: Sin embargo, al final de la película hay una boda donde algunas niñas cantan una vieja **canción** que muestra la **película**.

Embeddings for Machine Translation

Observations

- Neighbouring words change with the choice of the OOV translation, that is a cause for the large improvement
- Language model (LM) plays a more important role than WE probabilities in the final translation
- Since LM is so important, top 50 gives too much freedom (*chelow kabab is doogh* $\rightarrow$ *pan pan es leche*)
- ...but accuracy is low even at top 50

WMT 2016 Participation

Contents

WMT 2016 Participation

The Task

Shared Task: Biomedical Translation Task

– First Conference on Machine Translation (WMT16) –

Translation of scientific publications
for the biological and health domains

WMT 2016 Participation

The Task II

Translation of scientific publications
for the biological and health domains in

English–French
English–Spanish
English–Portuguese

WMT 2016 Participation

The Task II

Translation of scientific publications
for the biological and health domains in

English–French
English–Spanish
English–Portuguese

- One language pair, different systems
- Language–dependant systems for Spanish

WMT 2016 Participation

Temptative Main Architectures

- Phrase-based SMT in-domain (PB-SMT)
- PB-SMT with morphology
- **PB-SMT with OOV**

WMT 2016 Participation

Temptative Main Architectures

- Phrase-based SMT in-domain (PB-SMT)
- PB-SMT with morphology
- **PB-SMT with OOV**

- Neural MT in-domain (NMT)
- Char-based NMT
- **NMT with OOV**

WMT 2016 Participation

Temptative Main Architectures

- Phrase-based SMT in-domain (PB-SMT)
- PB-SMT with morphology
- **PB-SMT with OOV**

- Neural MT in-domain (NMT)
- Char-based NMT
- **NMT with OOV**

- Rescoring with char-based neural LM

WMT 2016 Participation

The Team

TALP-UPC

Carlos Escolano

Cristina España-Bonet

José Adrián Rodríguez Fonollosa

Marta Ruiz Costa-jussà

Pranava Swaroop Madhyastha

WMT 2016 Participation

Contents

WMT 2016 Participation

OOVs in in-Domain Data

Best configuration of previous experiments:

- Top 10 translation options
- NEs excluded (verbatim output)

WMT 2016 Participation

OOVs in in-Domain Data

Best configuration of previous experiments:

- Top 10 translation options
- NEs excluded (verbatim output)

Difference from previous experiment:

- Generalisation vs. domain adaptation

WMT 2016 Participation

Training Data

Corpus	**Segments**	**Words**	**Vocab**
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$

WMT 2016 Participation

Training Data

Corpus	**Segments**	**Words**	**Vocab**
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$
Bio-mono/en	$0.1 \cdot 10^6$	$2 \cdot 10^6$	$0.1 \cdot 10^6$
Bio-mono/es	$0.01 \cdot 10^6$	$0.1 \cdot 10^6$	$0.01 \cdot 10^6$
Wikipedia/en	$92 \cdot 10^6$	$1900 \cdot 10^6$	$2.0 \cdot 10^6$
Wikipedia/es	$20 \cdot 10^6$	$465 \cdot 10^6$	$0.8 \cdot 10^6$

WMT 2016 Participation

Training Data

Corpus	**Segments**	**Words**	**Vocab**
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$
Bio-mono/en	$0.1 \cdot 10^6$	$2 \cdot 10^6$	$0.1 \cdot 10^6$
Bio-mono/es	$0.01 \cdot 10^6$	$0.1 \cdot 10^6$	$0.01 \cdot 10^6$
Wikipedia/en	$92 \cdot 10^6$	$1900 \cdot 10^6$	$2.0 \cdot 10^6$
Wikipedia/es	$20 \cdot 10^6$	$465 \cdot 10^6$	$0.8 \cdot 10^6$

Nomenclature
(**STT**), BTT, SLM, BLM

WMT 2016 Participation

Training Data

Corpus	Segments	Words	Vocab
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$
Bio-mono/en	$0.1 \cdot 10^6$	$2 \cdot 10^6$	$0.1 \cdot 10^6$
Bio-mono/es	$0.01 \cdot 10^6$	$0.1 \cdot 10^6$	$0.01 \cdot 10^6$
Wikipedia/en	$92 \cdot 10^6$	$1900 \cdot 10^6$	$2.0 \cdot 10^6$
Wikipedia/es	$20 \cdot 10^6$	$465 \cdot 10^6$	$0.8 \cdot 10^6$

Nomenclature

(STT), **BTT**, SLM, BLM

WMT 2016 Participation

Training Data

Corpus	Segments	Words	Vocab
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$
Bio-mono/en	$0.1 \cdot 10^6$	$2 \cdot 10^6$	$0.1 \cdot 10^6$
Bio-mono/es	$0.01 \cdot 10^6$	$0.1 \cdot 10^6$	$0.01 \cdot 10^6$
Wikipedia/en	$92 \cdot 10^6$	$1900 \cdot 10^6$	$2.0 \cdot 10^6$
Wikipedia/es	$20 \cdot 10^6$	$465 \cdot 10^6$	$0.8 \cdot 10^6$

Nomenclature

(STT), BTT, **SLM**, BLM

WMT 2016 Participation

Training Data

Corpus	Segments	Words	Vocab
Biomedical	$1 \cdot 10^6$	$20 \cdot 10^6$	$0.3 \cdot 10^6$
Quest	$13 \cdot 10^6$	$340 \cdot 10^6$	$0.5 \cdot 10^6$
Bio-mono/en	$0.1 \cdot 10^6$	$2 \cdot 10^6$	$0.1 \cdot 10^6$
Bio-mono/es	$0.01 \cdot 10^6$	$0.1 \cdot 10^6$	$0.01 \cdot 10^6$
Wikipedia/en	$92 \cdot 10^6$	$1900 \cdot 10^6$	$2.0 \cdot 10^6$
Wikipedia/es	$20 \cdot 10^6$	$465 \cdot 10^6$	$0.8 \cdot 10^6$

Nomenclature

(STT), BTT, SLM, **BLM**

WMT 2016 Participation

Test Data

English test sets

	Seg.	**Tokens**	**OOV$_{\mathrm{STT}}$**	**OOV$_{\mathrm{BTT}}$**
BioDev	1000	18967	16 (0.08%)	2 (0.01%)
BioTest	1000	26105	31 (0.11%)	19 (0.12%)
News2013	2997	64521	844 (1.31%)	215 (0.33%)
Biological	4344	115709	434 (0.37%)	333 (0.29%)
Health	5111	125624	133 (0.10%)	98 (0.08%)

WMT 2016 Participation

Test Data

Spanish test sets

	Seg.	**Tokens**	**OOV$_{\mathrm{STT}}$**	**OOV$_{\mathrm{BTT}}$**
BioDev	1000	19931	14 (0.07%)	6 (0.03%)
BioTest	1000	27651	25 (0.09%)	9 (0.03%)
News2013	2997	71404	1174 (1.64%)	102 (0.14%)
Biological	4344	126008	415 (0.33%)	254 (0.20%)
Health	5111	146368	160 (0.11%)	40 (0.03%)

WMT 2016 Participation

System Definition

Systems

BTToov
BTT
STToov
STT
STTSLMoov
STTSLM

WMT 2016 Participation

Automatic Evaluation: English-to-Spanish

BioTest

	TER	BLEU	MTR
BTT	44.33	43.10	62.68
STT	**43.37**	44.00	**63.42**
STTSLM	43.69	44.16	63.40

WMT 2016 Participation

Automatic Evaluation: English-to-Spanish

BioTest

	TER	BLEU	MTR
BTToov	44.85	43.32	62.80
BTT	44.33	43.10	62.68
STToov	44.38	43.74	63.16
STT	**43.37**	44.00	**63.42**
STTSLMoov	44.24	**44.46**	63.36
STTSLM	43.69	44.16	63.40

WMT 2016 Participation

Automatic Evaluation: English-to-Spanish

	BioTest			NewsTest		
	TER	BLEU	MTR	TER	BLEU	MTR
BTToov	44.85	43.32	62.80	**53.18**	**25.60**	**49.91**
BTT	44.33	43.10	62.68	53.39	24.90	49.26
STToov	44.38	43.74	63.16	57.97	21.53	45.11
STT	**43.37**	44.00	**63.42**	58.32	20.87	44.66
STTSLMoov	44.24	**44.46**	63.36	60.99	18.19	42.46
STTSLM	43.69	44.16	63.40	60.56	18.05	42.57

WMT 2016 Participation

Automatic Evaluation: Spanish-to-English

BioTest

	TER	BLEU	MTR
BTToov	46.77	40.97	37.29
BTT	45.98	41.97	**37.65**
STToov	46.68	40.82	37.12
STT	46.74	41.16	37.40
STTSLMoov	**45.62**	**42.16**	37.60
STTSLM	46.46	41.71	37.48

WMT 2016 Participation

Automatic Evaluation: Spanish-to-English

	BioTest			NewsTest		
	TER	BLEU	MTR	TER	BLEU	MTR
BTToov	46.77	40.97	37.29	**54.69**	**27.07**	33.62
BTT	45.98	41.97	**37.65**	55.14	26.86	**33.73**
STToov	46.68	40.82	37.12	60.15	22.15	30.82
STT	46.74	41.16	37.40	61.49	21.28	30.64
STTSLMoov	**45.62**	**42.16**	37.60	61.80	20.06	29.91
STTSLM	46.46	41.71	37.48	62.83	19.54	29.86

WMT 2016 Participation

Observations

- For out-of-domain data (OOV $>$ 1%?) results are consistently improved by using translation options coming from BWE
- For in-domain data the percentage of OOVs is very low especially when using large corpora (improvements are small and difficult to locate: OOVs, weights, data?)
- For in-domain data and small corpora, some OOVs correspond to common words (*enslaving*), frequent in monolingual corpora

WMT 2016 Participation

Submission to the Challenge

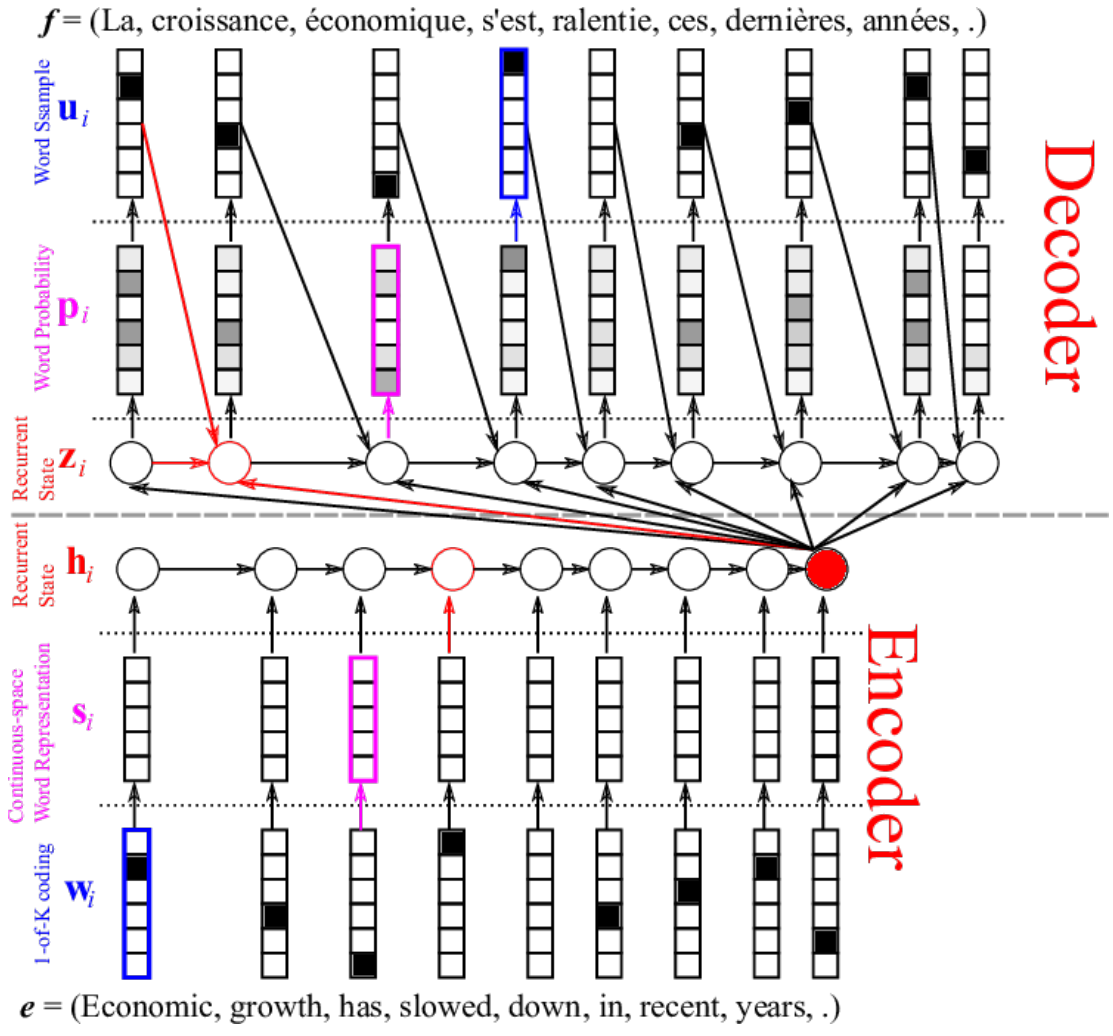
- SSTSLMoov in both directions
- SSTSLM submitted for comparison

and...

- Combination of 1000-best lists rescored using a char-based neural language model

WMT 2016 Participation

Left out... for now! Neural MT with BWE



Neural MT
Cho et al. 2014

(basic system
without attention)

WMT 2016 Participation

Left out... for now! Neural MT with BWE II

Interesting characteristics

- Embeddings projected into a low-dimensional space
  - From 500 to 100 for example
- The projection is specific for the translation problem
  - $[\phi(t)^\top U]\ \Sigma\ [V^\top \phi(s)]$

WMT 2016 Participation

Left out... for now! Neural MT with BWE III

Benefits of its integration

- Initialization or final monolingual embeddings
  - Can be pre-calculated
- Lower dimensional space
  - Faster execution
- Estimated for huge vocabularies
  - OOVs reduction

Recap & Comments

Contents

Recap & Comments

Summary

- BWE can be used to translate the otherwise OOVs
- They are a complement to the SMT translations, not a substitution (68% accuracy at top50, word-to-word)
- The language model plays a predominant role in the final choice (forcing the top1 as the translation is counter-productive)
- Even if the final translation of the OOV is not correct, the injection of new vocabulary is beneficial

Recap & Comments

Future Work

- Studying the relevance of the form of the loss function (log-probability vs. ranking –norm already studied–). The power of the LM makes less relevant than expected BWE probabilities
- Studying the effects of the seed lexicon: dictionary vs. IBM1 pairs, size and quality
- Using heuristics such as considering coocurrences of options in the monolingual corpus (or a better idea!) to deal with compound words

Recap & Comments

Other Applications

- BWE can have other applications in machine translation, such as the one sketched for neural machine translation
- A direct extension from the current work can also be used to extract parallel sentences from monolingual corpora
- Candidate sentences to be parallel would have a high word alignment score, were alignments can be established using BWE
- Framework: Wikipedia and WikiTailor

Recap & Comments

WikiTailor

WIKITAILOR

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

Joint work with Alberto Barrón-Cedeño

Recap & Comments

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

Recap & Comments

Integration in WikiTailor

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

- **Current Status**: Toy parallel sentence extractor, cosine similarities on *n*-grams, pseudo-cognates and translations
- **Aim**: Including scoring with BWE for widening the performance on more language pairs (with no external parallel data and/or from different families)

Resolving OOVs with BWE in MT

Cristina España-Bonet

Universitat Politècnica de Catalunya

TALP Research Center

Vielen Dank!