Fully Character-Level Machine Translation

Kyunghyun Cho
New York University
Courant Institute (Computer Science) and Center for Data Science
Facebook AI Research

Jason Lee, Kyunghyun Cho, Thomas Hoffman.
Fully Character-Level Neural Machine Translation without Explicit Segmentation. TACL 2017.
Junyoung Chung, Kyunghyun Cho, Yoshua Bengio.
Character-Level Decoding for Neural Machine Translation without Explicit Segmentation. ACL 2016.

What is a sentence?

- To a neural network, a sentence is just a sequence of one-hot vectors:

$$\left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \end{bmatrix}, \cdots \right)$$

  - See Lecture 2.
- What is the level of tokens with the minimal preprocessing?

What is a sentence?

- A sequence of words?
  - (Welcome, to, Montreal, !)
- A sequence of subwords?
  - (Wel, come, to, Mont, real, !)
- A sequence of sequences of letters?
  - ((W,e,l,c,o,m,e),(t,o),(M,o,n,t,r,e,al),(!))
- A sequence of characters?
  - (W,e,l,c,o,m,e, ,t,o, ,M,o,n,t,r,e,a,l,!)
- A sequence of bits…?

- <u>Research focus</u>
  - Subword-level translation (Sennrich et al., 2015)
  - Hybrid char/word translation (Luong et al., 2016)

What is a sentence to a neural network?

$$\left(\begin{bmatrix} 0 \\ 0 \\ \boxed{1} \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \end{bmatrix}, \cdots \right)$$

Vocabulary

Index	Word
1	the
2	a
3	,
4	and
5	of
6	.
7	to
⋮	

- A sentence is a sequence of one-hot vectors
- Underlying symbols are not visible to a neural network

What is a sentence to a neural network?

- A sequence of words?
  - (Welcome, to, Montreal, !)
- A sequence of subwords?
  - (Wel, come, to, Mont, real, !)
- A sequence of sequences of letters?
  - ((W,e,l,c,o,m,e),(t,o),(M,o,n,t,r,e,al),(!))
- A sequence of characters?
  - (W,e,l,c,o,m,e, ,t,o, ,M,o,n,t,r,e,a,l,!)
- A sequence of bits…?

- They are all just a sequence of one-hot vectors to a neural network…

$$\left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \end{bmatrix}, \ldots \right)$$

Why not (sub)word-level modelling?

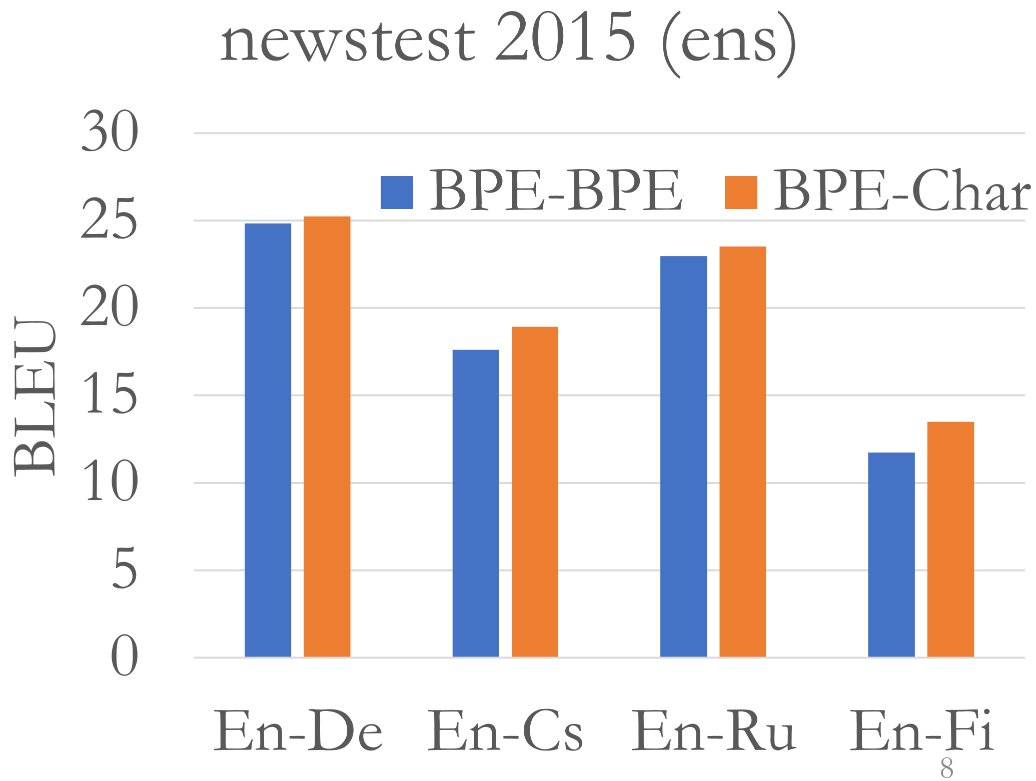
- Difficult to handle morphology
  - Morphologically rich languages → exploding vocabulary
  - Rare morphological variants: 과학/기술/정보/통신/부
- Difficult to handle misspelling
  - Especially serious with user-generated content (social media, reviews, blogs…)
  - Patterns of misspelling cannot be captured: could → cld, would → ?
- Modelling inefficiency
  - "kolmi/vaihe/kilo/watti/tunti/mittari": one vector?
  - "kolme": one vector???

Problems with character-level modelling

1. Can a neural network generate a long, coherent sequence?
2. Can a neural network capture highly nonlinear orthography?
3. Can character-level modelling be done efficiently?

Generating a long, coherent sequence

- Source: subword (BPE) sequence → Target: character sequence
- Language pairs: En→Cz, En→De, En→Fi, En→Ru
- Training sets: WMT'15
- Evaluation sets: newstest 2015
- Evaluation metric: BLEU
- BPE→Char ≅ BPE→BPE
- *Yes, a recurrent network can generate a long (100~300), coherent sequence*

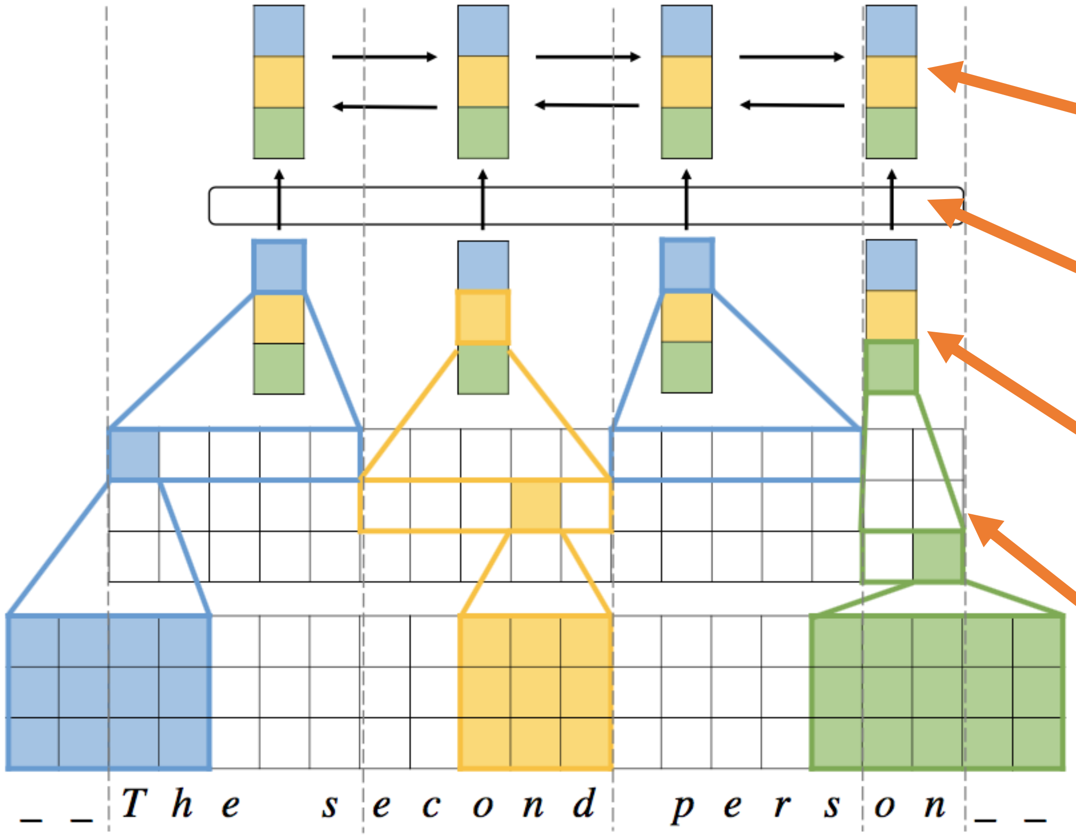


[Chung et al., 2016 ACL]

Going fully character-level

- Orthography is highly arbitrary without clear patterns
  1. Start: “quit”
  2. Insert “e” at the end: “quite”
  3. Swap the last two letters: “quiet”
- Complexity of attention grows quadratically w.r.t. the length
  - For each target symbol, all the source symbols must be considered
  - BPE-to-BPE: 30 x 30
  - BPE-to-Char: 30 x 120
  - Char-to-Char: 120 x 120

Going fully character-level

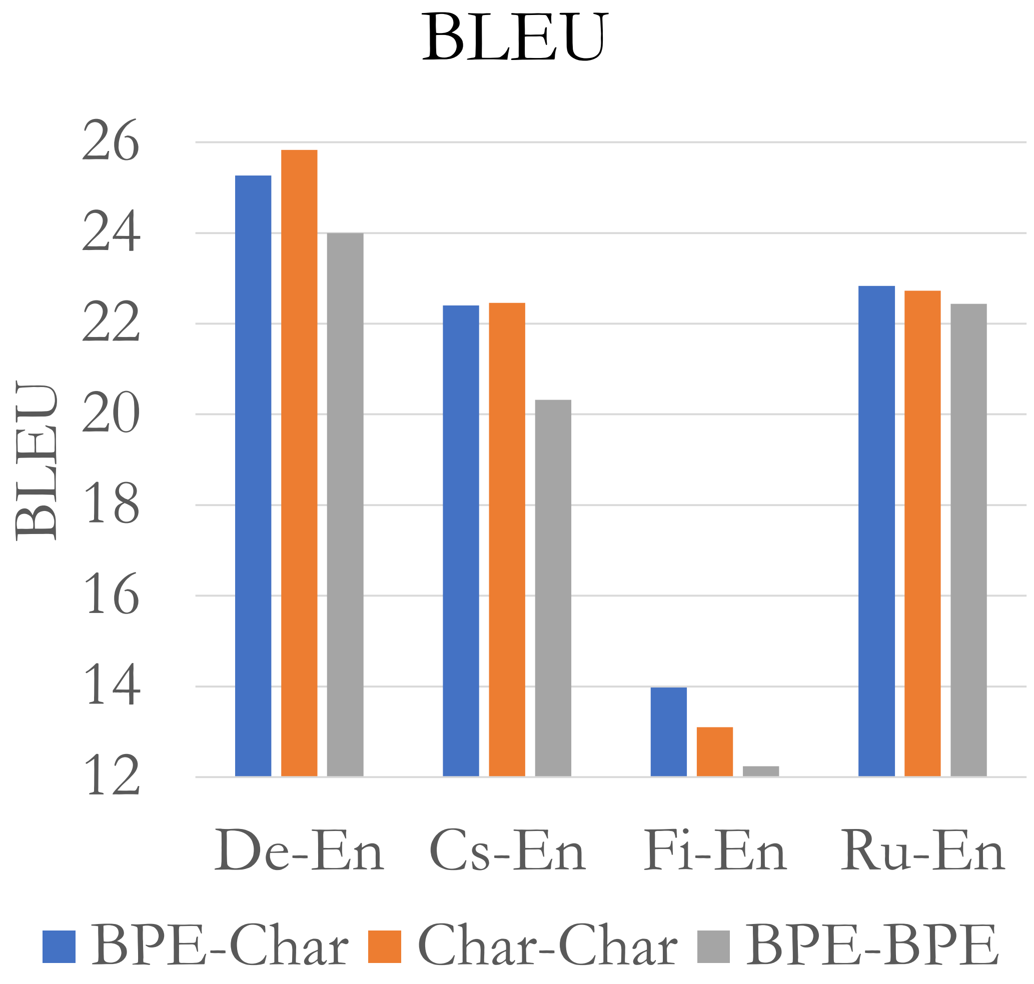


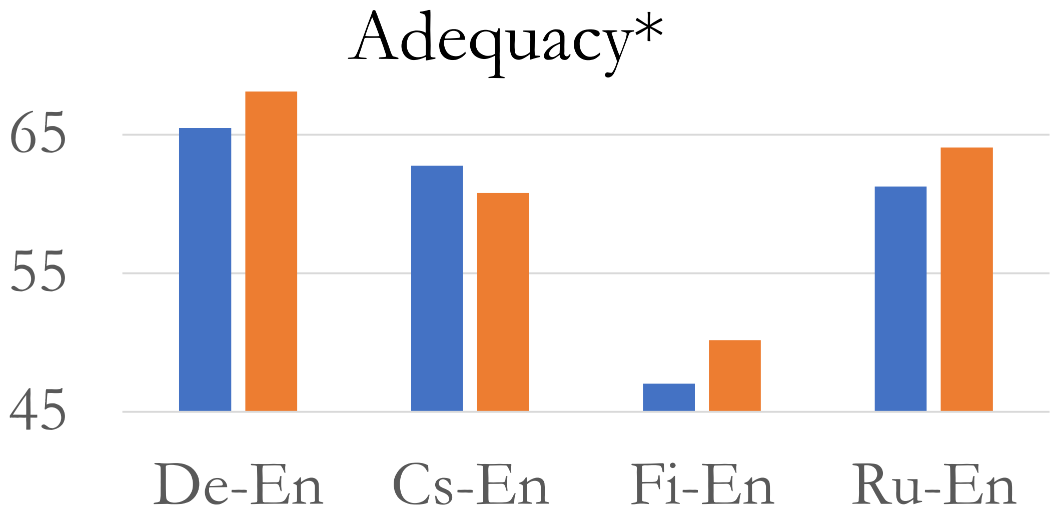
Character-Level Encoder

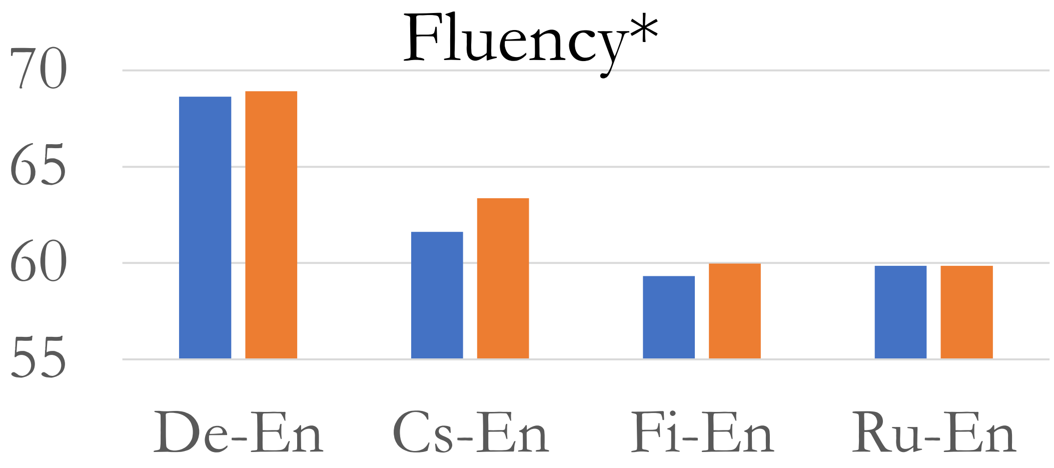
- **Attention** over a convolutional feature map [Xu et al., 2015]
- **High-way network** for nonlinear mapping [Srivastava et al., 2015]
- **Max pooling** for computational efficiency
- Multi-width **convolution** for a character sequence [Kim et al., 2015]

[Lee et al., 2017 TACL]

Going fully character-level







(*) Thanks to Yvette Graham for her help with human evaluation!

[Lee et al., 2017 TACL]

(a) Spelling mistakes

DE ori	Warum sollten wir nicht Freunde sei ?
DE src	Warum solltne wir nich Freunde sei ?
EN ref	Why should not we be friends ?
bpe2char	Why are we to be friends ?
char2char	Why should we not be friends ?

(b) Rare words

DE src	Siebentausendzweihundertvierundfünfzig .
EN ref	Seven thousand two hundred fifty four .
bpe2char	Fifty-five Decline of the Seventy .
char2char	Seven thousand hundred thousand fifties .

(c) Morphology

DE src	Die Zufahrtsstraßen wurden gesperrt , wodurch sich laut CNN lange Rückstaus bildeten .
EN ref	The access roads were blocked off , which , according to CNN , caused long tailbacks .
bpe2char	The access roads were locked , which , according to CNN , was long back .
char2char	The access roads were blocked , which looked long backwards , according to CNN .

(d) Nonce words

DE src	Der Test ist nun über , aber ich habe keine gute Note . Es ist wie eine Verschlimmbesserung .
EN ref	The test is now over , but i don’t have any good grade . it is like a worsened improvement .
bpe2char	The test is now over , but i do not have a good note .
char2char	The test is now , but i have no good note , it is like a worsening improvement .

[Lee et al., 2017 TACL]

Going fully character-level: Analysis

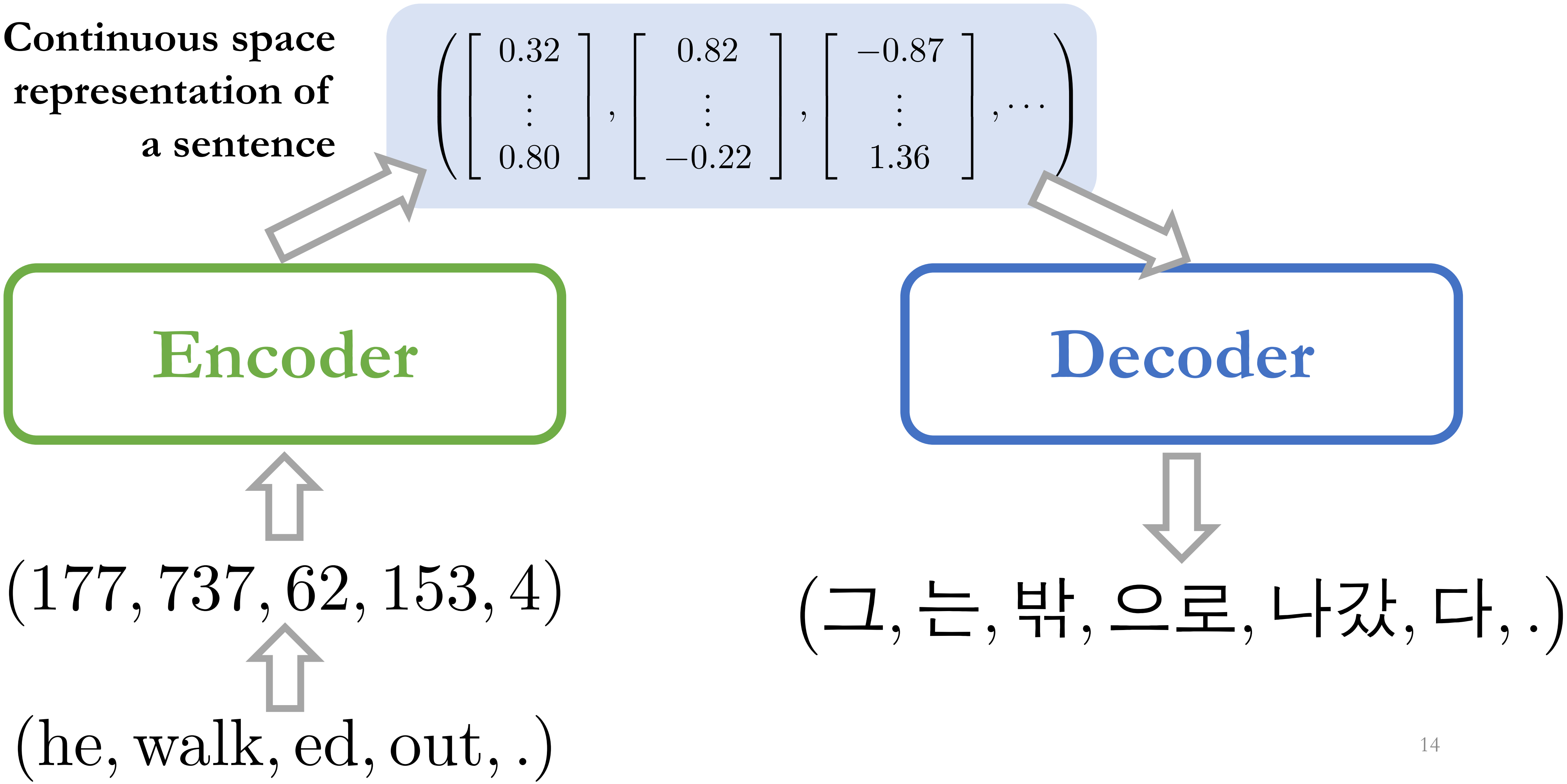
system (test set and size→)	2014 3003	2015 2169	2016 2999
BPE-to-BPE	20.1 (21.0)	23.2 (23.0)	26.7 (26.5)
BPE-to-char	19.4 (20.5)	22.7 (22.6)	26.0 (25.9)
char-to-char	19.7 (20.7)	22.9 (22.7)	26.2 (26.1)
(Sennrich et al., 2016a)	25.4 (26.5)	28.1 (28.3)	34.2 (34.2)

- Language pair/direction: En→De
- Similarly performing models (BLEU)
- Different properties:
    - Better transliteration with char-level modelling
    - Better syntactic properties with BPE-level

system	agreement		verb particle	polarity (negation)		transliteration
(category and size→)	noun phrase 21813	subject-verb 35105	2450	insertion 22760	deletion 4043	3490
BPE-to-BPE	**95.6**	**93.4**	**91.1**	97.9	**91.5**	96.1
BPE-to-char	93.9	91.2	88.0	**98.5**	88.4	**98.6**
char-to-char	93.9	91.5	86.7	**98.5**	89.3	**98.3**
(Sennrich et al., 2016a)	98.7	96.6	96.1	98.7	92.7	96.4
human	99.4	99.8	99.8	99.9	98.5	99.0

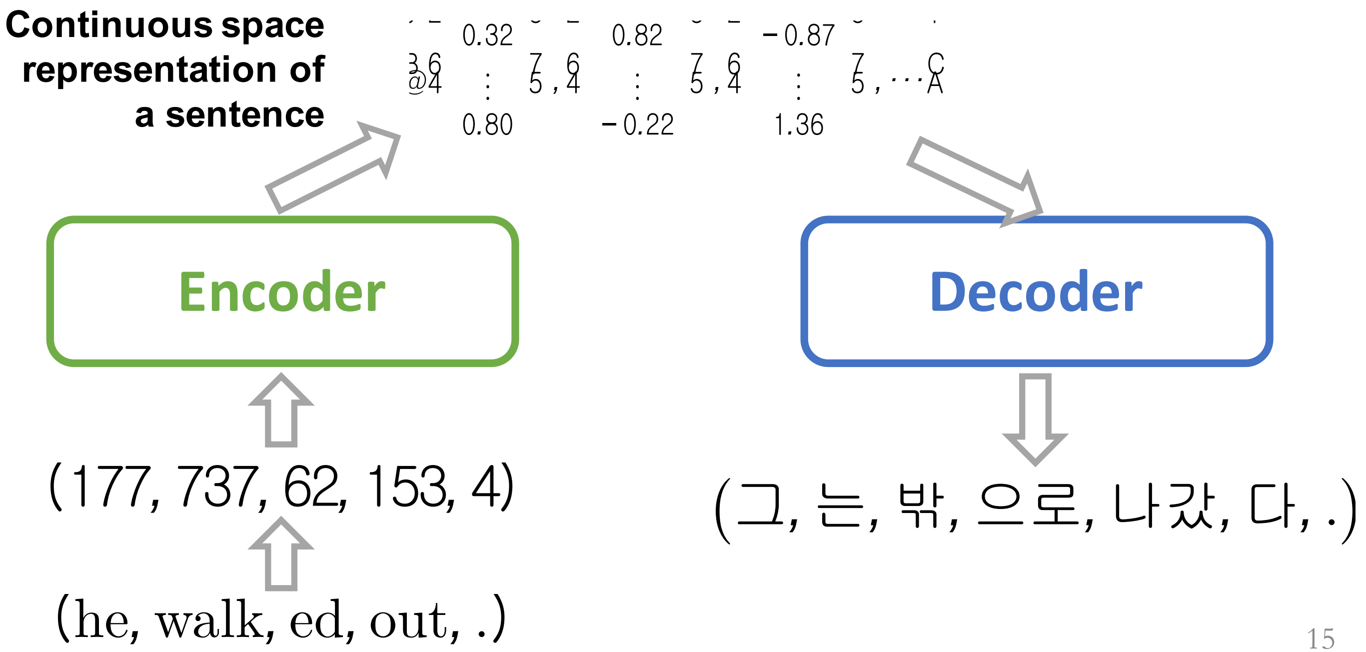
[Sennrich, 2017]

What does NMT do?

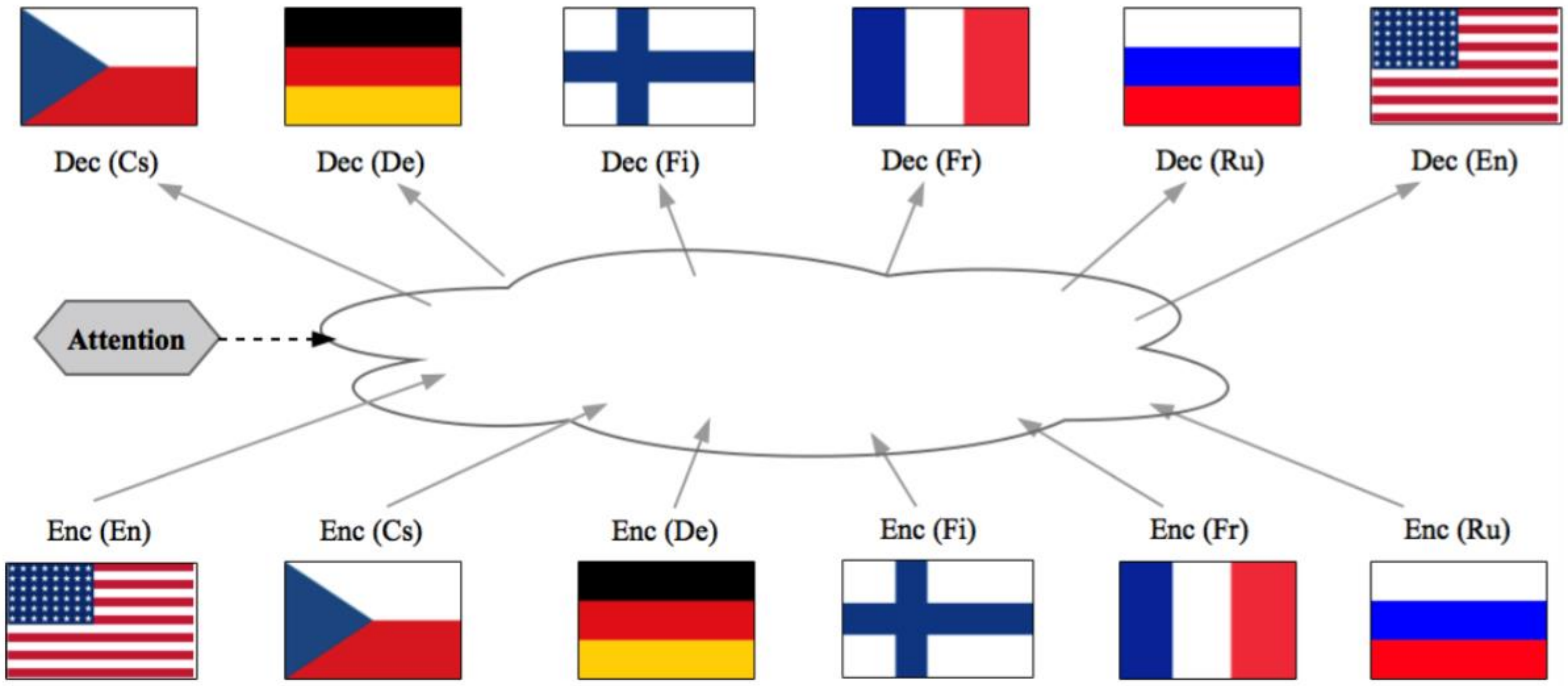


What does NMT do?

- Sequence of “discrete” symbols → Set of “continuous” vectors
- Continuous vectors *encode* semantics of discrete symbols
- Continuous vectors are *stripped* of hard, linguistic symbols
- *Can we map multiple languages on a single continuous space?*



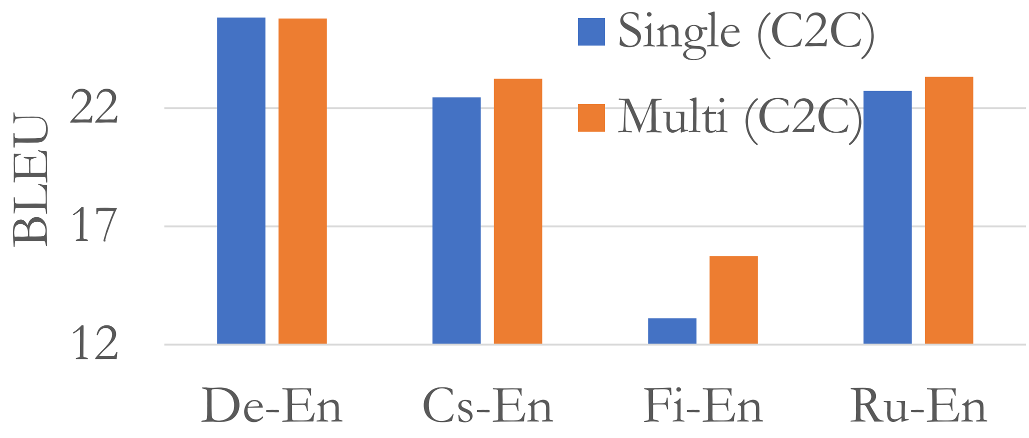
Multilingual Neural Machine Translation

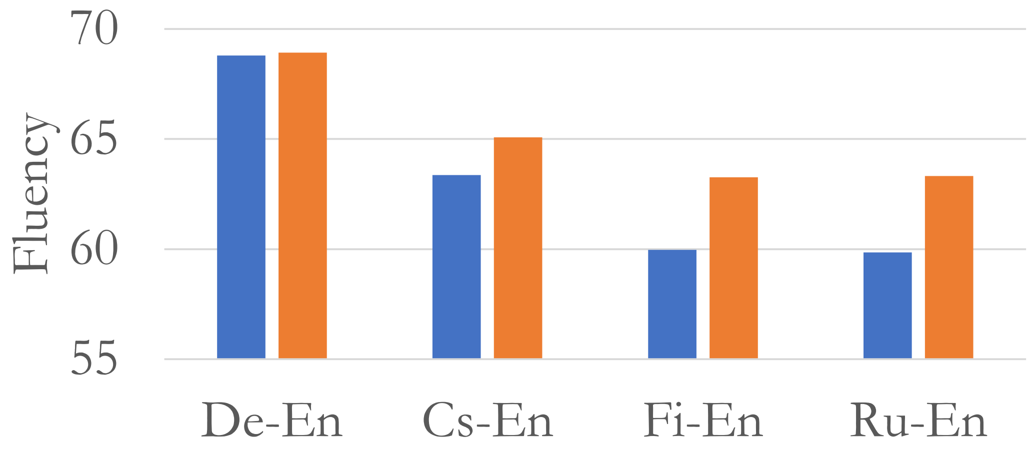


- Can this continuous vector space be shared across multiple languages?

(Lee et al., 2017; Ha et al., 2016; Johnson et al., 2016; Firat et al., 2016; Luong et al., 2015; Dong et al., 2015)

Character-Level, Multilingual Translation





- Characters are often shared across many languages
  - Latin alphabets for most of European languages
  - A sentence is given as a sequence of characters (inc alphabets, punctuation marks and blank spaces)
- {De, Cs, Fi, Ru} => En
- No language ID

[Lee et al., 2017 TACL] for character-level modelling
[John et al., 2016; Ha et al. ,2016] for subword-level modelling

Character-Level, Multilingual Translation

- Robust to intra-sentence code switching
- Huge saving in parameters: 4x less parameters without loss in BLEU

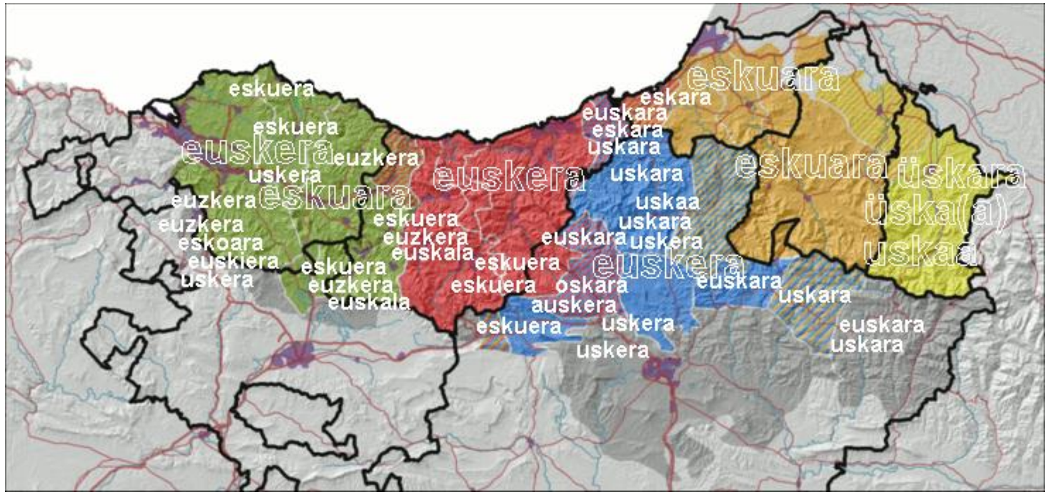
(e) Multilingual

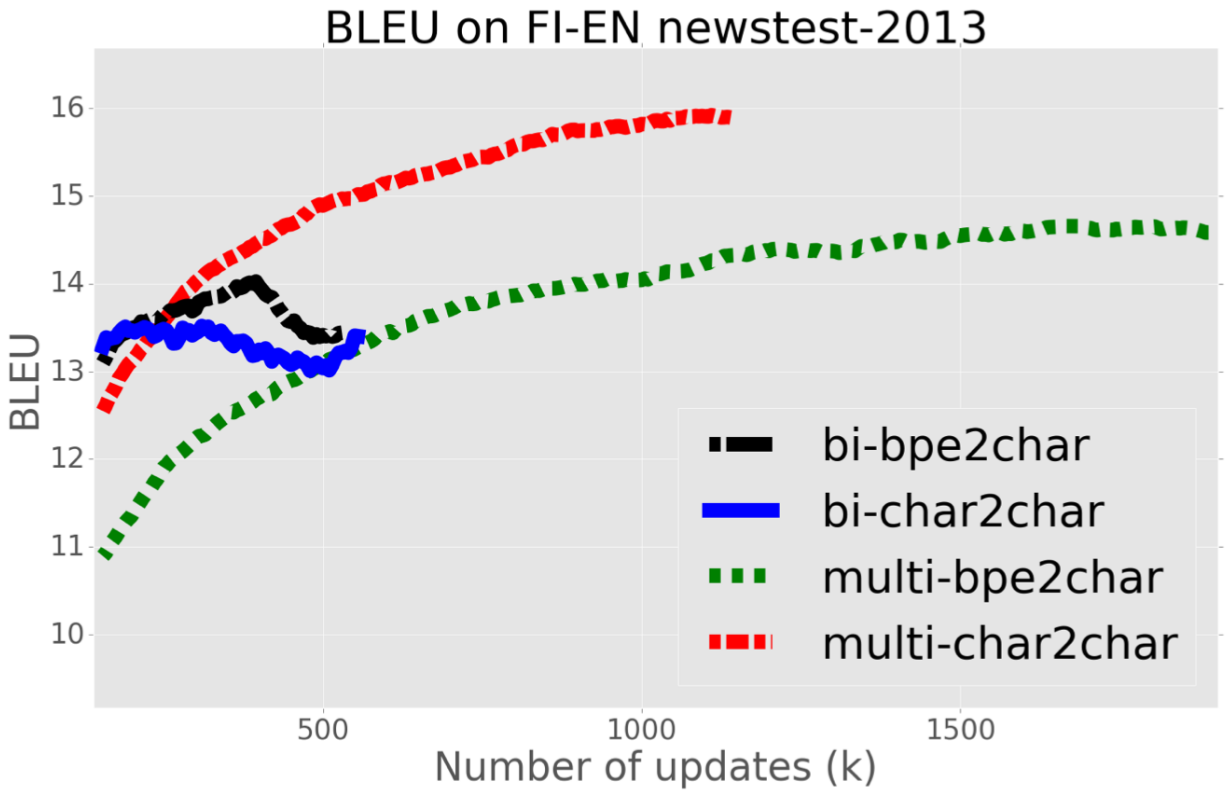
Multi src	Bei der Metropolitního výboru pro dopravu für das Gebiet der San Francisco Bay erklärten Beamte , der Kongress könne das Problem банкротство доверительного Фонда строительства шоссейных дорог einfach durch Erhöhung der Kraftstoffsteuer lösen .
EN ref	At the Metropolitan Transportation Commission in the San Francisco Bay Area , officials say Congress could very simply deal with the bankrupt Highway Trust Fund by raising gas taxes .
bpe2char	During the Metropolitan Committee on Transport for San Francisco Bay , officials declared that Congress could solve the problem of bankruptcy by increasing the fuel tax bankrupt .
char2char	At the Metropolitan Committee on Transport for the territory of San Francisco Bay , officials explained that the Congress could simply solve the problem of the bankruptcy of the Road Construction Fund by increasing the fuel tax .

[Lee et al., 2017 TACL] for character-level modelling
[John et al., 2016; Ha et al. ,2016] for subword-level modelling

Character-level, Multilingual Translation

- Prevents overfitting with low-resource language pairs
- Perhaps, a way to build a MT system for all the languages in the world?





[Lee et al., 2017, TACL]