Compositional Representation of Morphologically-Rich Input for Neural Machine Translation

Duygu Ataman[1,2] and Marcello Federico[1,3]

[1] *Fondazione Bruno Kessler, Trento, Italy*
[2] *Università degli Studi di Trento, Italy*
[3] *MMT Srl, Trento, Italy*
ataman@fbk.eu, federico@fbk.eu

ModernMT

Morphology

Analytic (Isolating) Languages
One word, one morpheme

เขา กำลัง เรียน ภาษา ไทย อยู่
Khaw **kamlang** rian phasaa thaai **yuu**
S/he PROG study language Thai at
*She **is** studying the Thai language.*

Synthetic Languages
One word, multiple morphemes

Fusional Morphology
Single inflectional morpheme to denote multiple grammatical, syntactic, or semantic features.

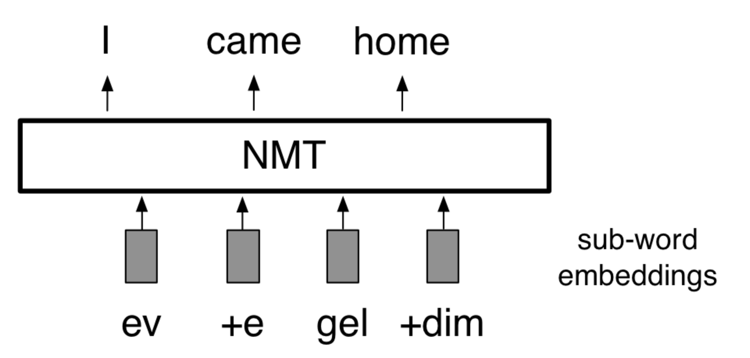
Я вижу при-дорож-н-ое кафе
Ya vizhu pri-dorozh-n-**oye** kafe.
I see.1Sg.Pres near-road-ADJ-**Acc+Sg+Neu** cafe.
I see a roadside cafe.

Agglutinative Morphology
Each morpheme corresponds to a separate semantic or syntactic feature.

Arkadaş-ım-ın aşk-ı-sı-n.
friend-my-of love-DET-Pres-2Sg
You are the love of my friend.

High morphological complexity leads to many rare surface forms in the vocabulary, that either
- do not fit in the limited NMT model dictionary, or,
- have poor internal representations

NMT with Sub-word Embeddings



Conventional Methods

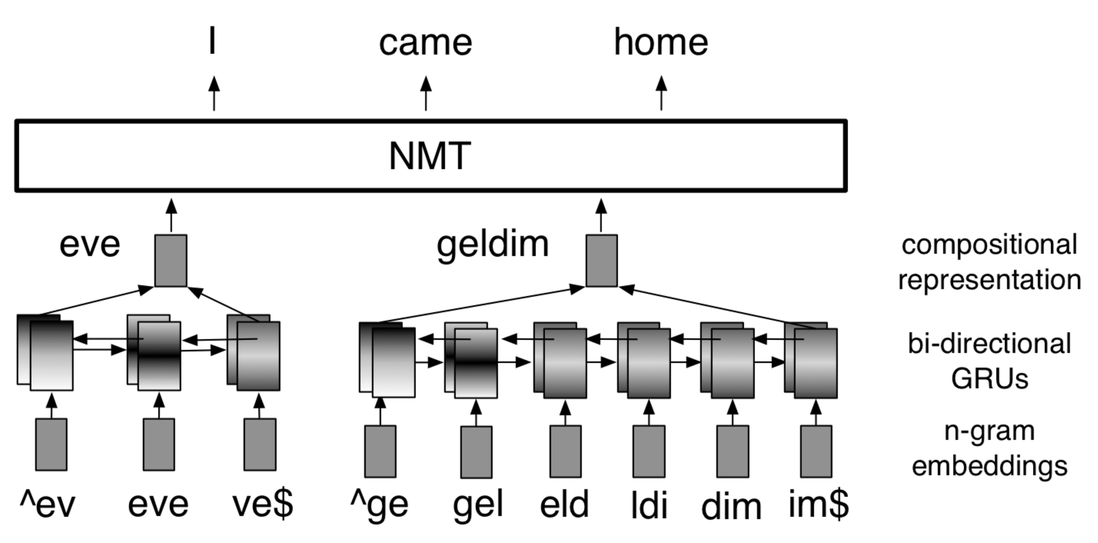
Byte-Pair Encoding (BPE)
(Sennrich et al. 2016)
- Iteratively merge character sequences based on their **frequency** to create new sub-words, until target number is reached
- Shows to cope with OOV loanwords and named entities

Linguistically-Motivated Vocabulary Reduction (LMVR)
(Ataman et al., 2017)
- Sub-words are split based on their **likelihood of being morphemes and morphological categories** through a Hidden Markov Model
- Useful for concatenative (agglutinative/templatic) morphology

Problems with Sub-word Segmentation

- Not optimized for the machine translation task
- No generic solution for different languages
- Translating sub-words requires remembering longer histories due to increased sentence lengths, increased complexity of alignments, loss of semantic/syntactic features due to morphological errors

NMT with Compositional Representations



Reduce data sparseness by integrating *morphology learning* into the model

Using bi-RNNs allows to
- predict a mapping from phonetic units to lexical context, an approximation to morphology
- encode relative positions of characters within words, providing cues on their functional roles (*i.e.* semantic/syntactic contribution to the word meaning)

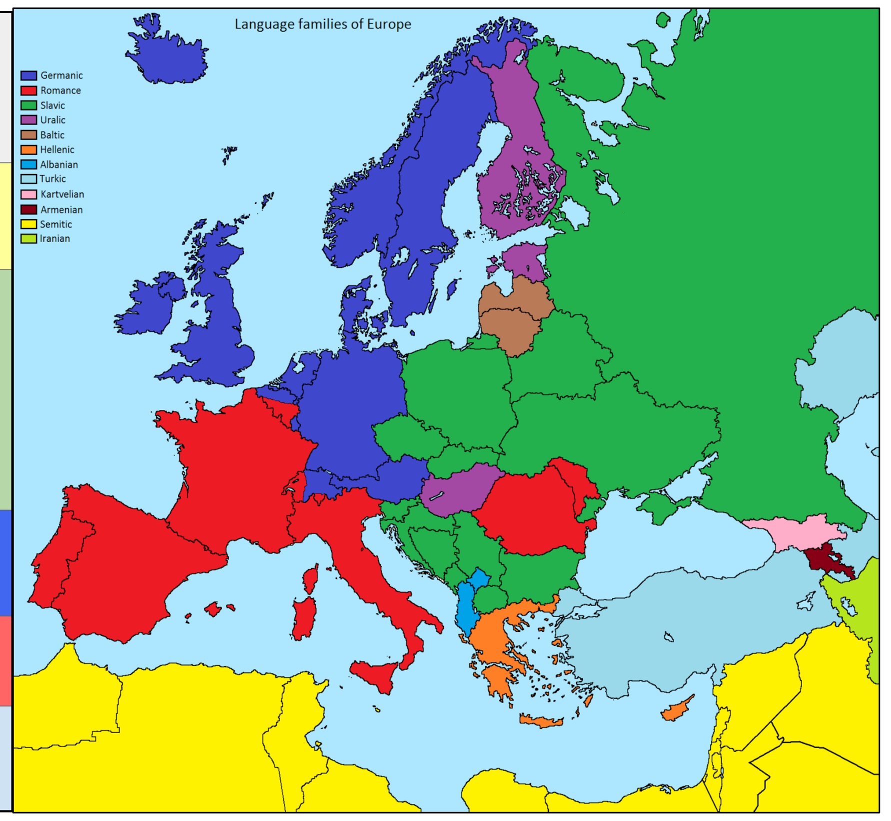
The final word representation is obtained by a linear combination of forward (f) and backward (b) final states of the bi-RNNs:

$$w = W_f h_f + W_b h_b + b$$

Parameters W, b jointly learned with rest of NMT model parameters to minimize overall network cost.

Evaluation

Language	Family	Morphological Complexity	Morphological Typology
Arabic	Semitic	High	Templatic
Czech	Slavic	High	Mostly Fusional, Partially Agglutinative
German	Germanic	Medium	Fusional
Italian	Italic	Low	Fusional
Turkish	Turkic	High	Agglutinative



Implementation
- Using Theano, integrated into NMT toolkit *Nematus*

Variables
- Levels of granularity for composition
- Morphological typology (*i.e.* lexical sparseness)

Data
- Training set: TED Talks (*150-200K* sentences)
- Dev and test: IWSLT (*3K* sentences each)

Hyper-parameters
- GRU: 512 hidden units, Embedding size: 512, Adagrad with lr=0.01
- Vocabulary size: 30,000 units (BPE, LMVR sub-words or character n-grams)

Results

Model	Vocabulary Units	Input Representations	BLEU TR-EN	AR-EN	CS-EN	DE-EN	IT-EN
NMT with Sub-word Embeddings	*Characters*	*Characters*	12.29	8.95	13.42	21.32	22.88
	Char Trigrams	*Char Trigrams*	16.13	11.91	20.87	25.01	26.68
	Sub-words (BPE)	*Sub-words (BPE)*	16.79	11.14	21.99	26.61	27.02
	Sub-words (LMVR)	*Sub-words (LMVR)*	17.82	12.23	22.84	27.18	27.34
NMT with Compositional Representations	*Char Trigrams*	*Sub-words (BPE)*	15.40	11.50	21.67	27.05	27.80
	Char Trigrams	*Sub-words (LMVR)*	16.63	13.29	23.07	26.86	26.84
	Char Trigrams	*Words*	**19.53**	**14.22**	**25.16**	**29.09**	**29.82**
	Subwords (BPE)	*Words*	12.64	11.51	23.13	27.10	27.96
	Subwords (LMVR)	*Words*	18.90	13.55	24.31	28.07	28.83

Examples

Input: *BPE Sub-words*	ama aslında bu resim tamamen , farklı yerlerin fotoğraf@@ larının birleştir@@ il@@ mesiyle meydana geldi .
NMT Output: *BPE Sub-words*	but in fact , this picture came up with a completely different place of photographs .
Input: *Compositional Model*	ama aslında bu resim tamamen , farklı yerlerin fotoğraflarının birleştirilmesiyle meydana geldi .
NMT Output: *Compositional Model*	but in fact , this picture came from collecting pictures of different places .
Reference	but this image is actually entirely composed of photographs from different locations .

Conclusions

- Compositional input representations compare favourably with sub-word embeddings
- Results suggest eliminating sub-word segmentation completely for morphologically-rich input for avoiding morphological errors
- Maintaining lexical boundaries allows to learn better syntax
- The compositional NMT approach provides a generic solution for machine translation that can generalize over different morphological typology or language families

The 56th Annual Meeting of the Association for Computational Linguistics, 15-20 July 2018, Melbourne.