Factored Neural Machine Translation Architectures

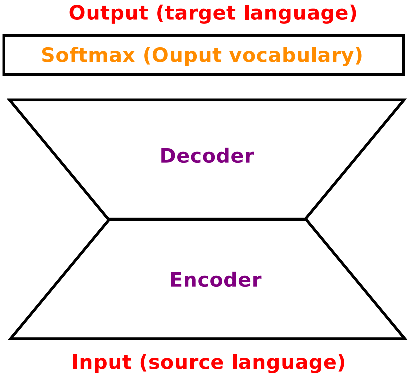
Mercedes García-Martínez,
Loïc Barrault and Fethi Bougares

LIUM - University of Le Mans

IWSLT 2016
December 8-9, 2016

Neural Machine Translation

- Sequence to sequence implemented in an encoder-decoder RNN



- Problems:
  - Softmax normalization has a high computational cost
  - Dealing with unknown words

Neural Machine Translation / Solutions

- Short-list: most frequent words are used and the rest are mapped to unknown
  → simple / model only few words, does generate many UNK
- Structured output layer LM / SOUL (short-list + subclasses) [Le,2011]
  → manage more vocabulary / more complex architecture
- Select the batches so that the softmax can be applied to only a subset of the output layer [Jean,2015]
  → architecture remains the same / mismatch between train and test modes
- Subword units extracted by BPE [Sennrich,2015] or characters [Chung,2016]
  → simple, no other resources required, unseen words can be generated → no UNK / less control on the output, incorrect words can be generated

  → BPE is the main method used in recent evaluation campaigns (WMT, IWSLT)

Motivation

- None of the methods presented before includes linguistics to solve this problem
- We propose to use a **factored word representation**
  - Based on the way humans learn how to construct inflected word forms

  $\rightarrow$ Handling larger effective vocabulary using same output size

  $\rightarrow$ Generation of words that did not appear in the vocabulary

Ex. EN: **be** $\rightarrow$ { am, is, are, was, were, been, being }
Ex. FR: **être** $\rightarrow$ { suis, es, est, sommes, êtes, sont, serai, ..., étais, .., été, ...}

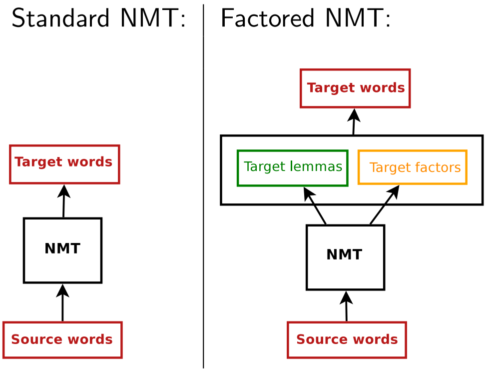
- Related factors works:
  - Factored NLM, in addition to words [Alexandrescu,2006; Wu,2012; Niehues,2016]
  - Additional information in the input side in NMT [Sennrich, 2016]

$\Rightarrow$ Our method uses **factors (<u>not words</u>)** at the **<u>output</u>** side of the NN

Overview

1. Factored NMT architectures
2. Experiments on TED talks IWSLT'15
3. Qualitative analysis
4. Conclusions and future work

Factored NMT approach

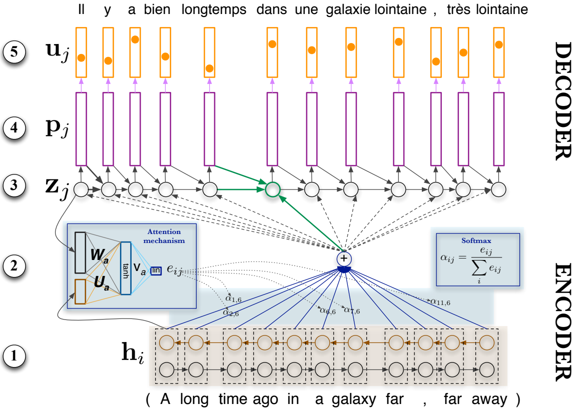


- Morphological analyser:

word	lemma	factors (POS+tense+person+gender+number)
devient	devenir	verb + present + 3rd person + # + singular

- Generate surface form using lemma and factors
- Same output size for lemma + small output for factors → larger vocabulary

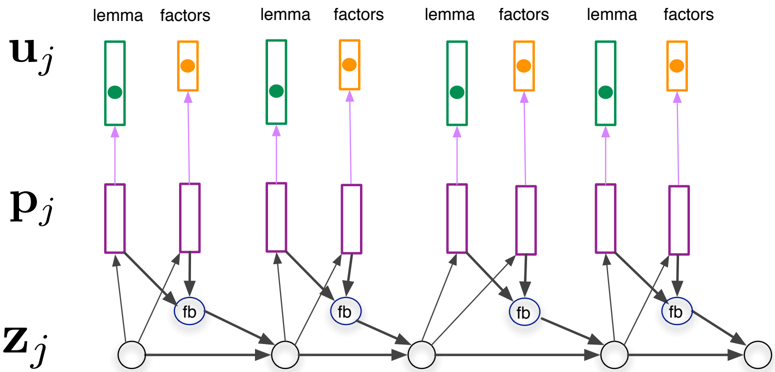
Base NMT model



- NMT by jointly learning to align and translate [Bahdanau et al.,2014]
- Conditional GRU
- 2015 DL4MT Winter School Code: https://github.com/nyu-dl/dl4mt-tutorial

Our Factored NMT model

→ Base NMT decoder extended to get 2 outputs:



- 2 symbols generated **synchronously**: (1) lemma and (2) factors
  - Factors sequence length = lemma sequence length
- For comparison: multiway multilingual setup from [Firat,2016]

Experiments

- EN-FR, IWSLT'15 corpus (data selection and filter out long sentences):

Sentences	2M	
EN unique words	147K	
FR unique words	266K	
Word input size	30K	
Lemma/word output size	30K	FNMT word vocabulary **172K**
Factors output size	142	

- Neural network settings:

RNN dim	1000
Embedding dim	620
Minibatch size	80
Gradient clipping	1
Weight initialization	Xavier [Glorot,2010]
Learning rate scheme	Adadelta
Beam size	12

First Results

Model	Feedback	%MET.	%BLEU word	%BLEU lemma	%BLEU factors	#UNK
FNMT	Lemma	56.96	34.56	37.44	42.44	798
NMT	Words	55.87	34.69	35.10	40.79	1841
BPE	Subwords	55.52	34.34	34.64	40.25	0
Multilingual	Lemma/Factors	52.48	28.70	37.72	45.81	871
Chain NMT	Lemma/Factors	56.00	33.82	37.38	90.54	773

- FNMT reduces #UNK compared to NMT

First Results

Model	Feedback	%MET.	%BLEU word	%BLEU lemma	%BLEU factors	#UNK
FNMT	Lemma	56.96	34.56	37.44	42.44	798
NMT	Words	55.87	34.69	35.10	40.79	1841
BPE	Subwords	55.52	34.34	34.64	40.25	0
Multilingual	Lemma/Factors	52.48	28.70	37.72	45.81	871
Chain NMT	Lemma/Factors	56.00	33.82	37.38	90.54	773

- FNMT reduces #UNK compared to NMT
- BPE does not generate UNK, but no impact on BLEU
- FNMT can apply UNK replacement to improve results

First Results

Model	Feedback	%MET.	%BLEU word	lemma	factors	#UNK
FNMT	Lemma	56.96	34.56	37.44	42.44	798
NMT	Words	55.87	34.69	35.10	40.79	1841
BPE	Subwords	55.52	34.34	34.64	40.25	0
Multilingual	Lemma/Factors	52.48	28.70	37.72	45.81	871
Chain NMT	Lemma/Factors	56.00	33.82	37.38	90.54	773

- FNMT reduces #UNK compared to NMT
- BPE does not generate UNK, but no impact on BLEU
- FNMT can apply UNK replacement to improve results
- Multilingual obtains higher BLEU in lemma and factors but lower in word

First Results

Model	Feedback	%MET.	%BLEU word	%BLEU lemma	%BLEU factors	#UNK
FNMT	Lemma	56.96	34.56	37.44	42.44	798
NMT	Words	55.87	34.69	35.10	40.79	1841
BPE	Subwords	55.52	34.34	34.64	40.25	0
Multilingual	Lemma/Factors	52.48	28.70	37.72	45.81	871
Chain NMT	Lemma/Factors	56.00	33.82	37.38	90.54	773

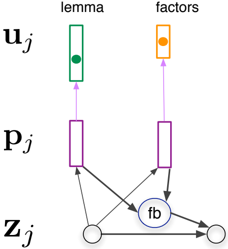
- FNMT reduces #UNK compared to NMT
- BPE does not generate UNK, but no impact on BLEU
- FNMT can apply UNK replacement to improve results
- Multilingual obtains higher BLEU in lemma and factors but lower in word
- Chain model: EN_words → NMT → FR_lemmas → NMT → FR_factors → FR_words
  → Low results can be explained by distinct training of two systems

First Results

Model	Feedback	%MET.	%BLEU word	%BLEU lemma	%BLEU factors	#UNK
FNMT	Lemma	56.96	34.56	37.44	42.44	798
NMT	Words	55.87	34.69	35.10	40.79	1841
BPE	Subwords	55.52	34.34	34.64	40.25	0
Multilingual	Lemma/Factors	52.48	28.70	37.72	45.81	871
Chain NMT	Lemma/Factors	56.00	33.82	37.38	90.54	773

- FNMT reduces #UNK compared to NMT
- BPE does not generate UNK, but no impact on BLEU
- FNMT can apply UNK replacement to improve results
- Multilingual obtains higher BLEU in lemma and factors but lower in word
- Chain model: EN_words → NMT → FR_lemmas → NMT → FR_factors → FR_words
  → Low results can be explained by distinct training of two systems
- Could expect a better Factor level BLEU? → only 142 tokens

Changing the feedback

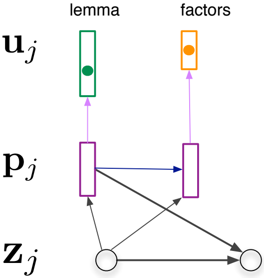


Model	Feedback	%BLEU word	%BLEU lemma	%BLEU factors
NMT	Words	34.69	35.10	40.79
FNMT	Lemma	34.56	37.44	42.44
FNMT	Factors	31.49	34.05	44.73
FNMT	Sum	34.34	37.03	44.16
FNMT	Concatenation	34.58	37.32	44.33

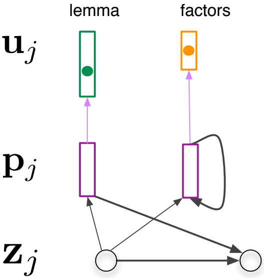
- Lemma embedding contains more information from the generated target word
- No big differences in word level BLEU
- Concatenation is good at lemma **and** factors level BUT no impact on word level BLEU

Dependency model

Lemma dependency:

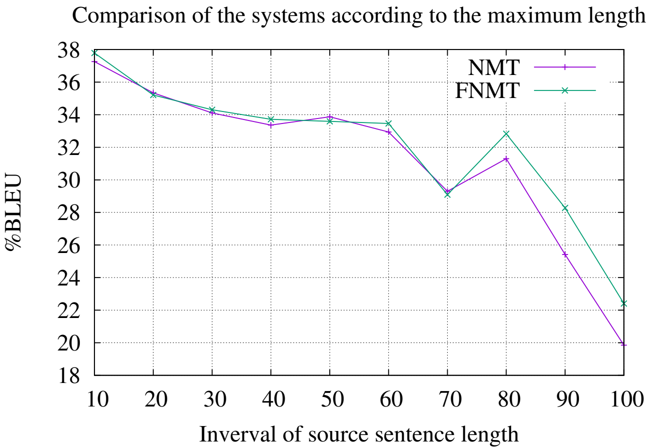


Factors dependency:



		%BLEU		
Model	**Depend.**	**word**	**lemma**	**factors**
NMT	-	34.69	35.10	40.79
FNMT	-	34.56	37.44	42.44
FNMT	prev. lem.	34.34	37.39	42.33
FNMT	curr. lem.	34.62	37.30	43.36
FNMT	prev. fact.	34.72	37.56	43.09

Results using different sentence length



- FNMT helps when translating sentences longer than 80 words
- Might be due to less sparsity on the lemma and factors space

Qualitative analysis. FNMT vs. NMT

1	Src	... set of adaptive choices that our lineage made ...	(len=90)
	Ref	... de choix adaptés établis par notre lignée ...	
	NMT	... de choix UNK que notre UNK a fait ...	
	FNMT	... de choix **adaptatifs** que notre **lignée** a fait ...	
2	Src	... enzymes that repair them and put them together	(len=23)
	Ref	... enzymes qui les réparent et les assemblent .	
	NMT	... enzymes qui les UNK et les UNK .	
	FNMT	... enzymes qui les **réparent** et les **mettent ensemble** .	
3	Src	... santa marta in north colombia	(len=26)
	Ref	... santa marta au nord de la colombie .	
	NMT	... santa UNK dans le nord de la colombie .	
	FNMT	... santa **marta** dans le nord de la colombie .	

- FNMT generates less *UNK* (*adaptés* and *lignée* are in the NMT target voc.)
- %BLEU penalizes some correct translations that are not the same as reference
- *réparent* and *marta* are not included in NMT vocabulary

Qualitative analysis.
FNMT vs. FNMT with current lemma dependency

W	Src	no one knows what the hell we do								
W	Ref	personne	ne	sait	ce	que	nous	faisons		.
W	FNMT	personne	ne	sait	ce	qu'	être	l'	enfer	.
L		personne	ne	savoir	ce	qu'	être	l'	enfer	.
F		pro-s	advn	v-P-3-s	prep	prorel	cln-3-s	det	nc-m-s	poncts
W	FNMT	personne	ne	sait	ce	que	nous	faisons		.
L	dep.	personne	ne	savoir	ce	que	nous	faire		.
F		nc-f-s	advn	v-P-3-s	det	prorel	**cln-1-p**	v-P-1-p		poncts

- Dependency improves factors prediction

Qualitative analysis. FNMT vs. BPE

Subwords (BPE) versus factors:

Src	we in medicine , I think , are baffled									
Ref	Je pense que en médecine nous sommes dépassés									
BPE_w	nous	,	en	médecine	,	je	pense	,	sommes	bafés
BPE	nous	,	en	médecine	,	je	pense	,	sommes	b+af+és
FNMT_w	nous	,	en	médecine	,	je	pense	,	sont	déconcertés
FNMT_l	lui	,	en	médecine	,	je	penser	,	être	déconcerter
FNMT_f	pro-1-p-l	pct-l	prep-l	nc-f-s-l	pct-l	cln-1-s-l	v-P-1-s-l	pct-l	v-P-3-p-l	vppart-K-m-p-l

- BPE translates *baffled* to *bafés* that does not exist in French
- FNMT translates *baffled* to *déconcertés*

Conclusions

- Factored NMT based on linguistic *a priori* knowledge
- Cons.
  - require up-to-date linguistic resources (difficult in low resources languages)
  - slight increase of model complexity (architecture with 2 outputs)
- Pros
  - handles very large vocabulary (6 times bigger in our experiments)
  - generates new words not included in the *shortlist*
  - performs better than other state of the art systems
  - reduces the generation of #UNK tokens
  - produces only correct words compared to subwords

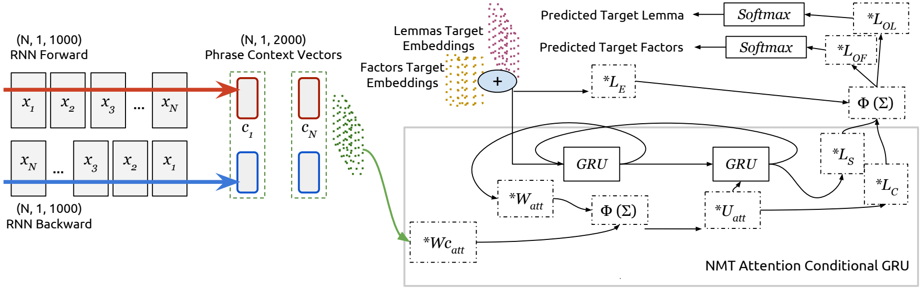
Future work

- Include factors in input side
  - Decrease the number of UNK in the source sequence
  - Motivated by [Sennrich et al,2016]
- Extend for N factors
  - Actually, the factors vocabulary is determined by the training set
  - Increase the generalisation power of the system to unseen word forms
- Apply to highly inflected languages
  - e.g. Arabic language

Thanks for your attention!

Le Mans, France

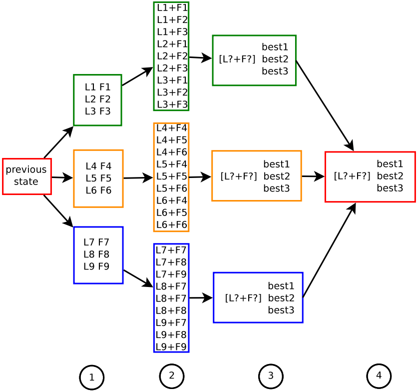
FNMT model



(N, 1, 1000)
RNN Forward
x_1 x_2 x_3 ... x_N
x_N ... x_3 x_2 x_1
(N, 1, 1000)
RNN Backward
(N, 1, 2000)
Phrase Context Vectors
c_1
c_N
Lemmas Target Embeddings
Factors Target Embeddings
+
Predicted Target Lemma
Predicted Target Factors
Softmax
Softmax
$*L_{OL}$
$*L_{OF}$
$*L_E$
$\Phi\ (\Sigma)$
GRU
GRU
$*L_S$
$*L_C$
$*W_{att}$
$\Phi\ (\Sigma)$
$*U_{att}$
$*Wc_{att}$
NMT Attention Conditional GRU

Handling beam search with 2 outputs

Timestep generation:



- Sum both costs to get 1-best translation
- 1 lemma - 1 factors: Limit length of the factors to the lemmas length
- Cross product of the nbest of the 2 outputs of each produced word
- Limit the options to the beam size

Feedback equations

$$
\begin{aligned}
\mathrm{GRU}_1\left(y_{j-1}, \mathbf{s}_{j-1}\right) &= (1-\mathbf{z}_j) \odot \underline{\mathbf{s}}_j + \mathbf{z}_j \odot \mathbf{s}_{j-1}, \\
\underline{\mathbf{s}}_j &= \tanh\left(\mathbf{W}\mathbf{fb}(y_{j-1}) + \mathbf{r}_j \odot \left(\mathbf{U}\mathbf{s}_{j-1}\right)\right), \\
\mathbf{r}_j &= \sigma\left(\mathbf{W}_r\mathbf{fb}(y_{j-1}) + \mathbf{U}_r\mathbf{s}_{j-1}\right), \\
\mathbf{z}_j &= \sigma\left(\mathbf{W}_z\mathbf{fb}(y_{j-1}) + \mathbf{U}_z\mathbf{s}_{j-1}\right),
\end{aligned}
$$

$$
\begin{array}{ll}
\text{Lemma :} & \mathbf{fb}(y_{t-1}) = y^L_{t-1} \\
\text{Factors:} & \mathbf{fb}(y_{t-1}) = y^F_{t-1} \\
\text{Sum:} & \mathbf{fb}(y_{t-1}) = y^L_{t-1} + y^F_{t-1} \\
\text{Linear sum:} & \mathbf{fb}(y_{t-1}) = (y^L_{t-1} + y^F_{t-1}) \cdot W_{fb} \\
\text{Tanh sum:} & \mathbf{fb}(y_{t-1}) = \tanh\left((y^L_{t-1} + y^F_{t-1}) \cdot W_{fb}\right) \\
\text{Linear concat:} & \mathbf{fb}(y_{t-1}) = [y^L_{t-1}; y^F_{t-1}] \cdot W_{fb} \\
\text{Tanh concat:} & \mathbf{fb}(y_{t-1}) = \tanh\left([y^L_{t-1}; y^F_{t-1}] \cdot W_{fb}\right)
\end{array}
$$

- y^L_{t-1} : lemma embedding at previous timestep
- y^F_{t-1} : factors embedding at previous timestep

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473.
- Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. CoRR, abs/1603.06147.
- Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2014. On using very large target vocabulary for neural machine translation. CoRR, abs/1412.2007.
- Philipp Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hai-Son. Le, Ilya Oparin, Abdel. Messaoudi, Alexandre Allauzen, Jean-Luc Gauvain, and François Yvon. 2011. Large vocabulary SOUL neural network language models. In INTERSPEECH.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. CoRR, abs/1508.07909.
- Barry Sennrich, Rico; Haddow. 2016. Linguistic input features improve neural machine translation. aeprint arXiv:1606.02892, June.