Low-Resource Neural Machine Translation

Alexandra Birch, Kenneth Heafield, Proyag Pal

University of Edinburgh

4/12/2017

Subword Neural MT [**Sennrich**, Haddow, Birch, ACL 2016a]

goal

subword segmentation that:

- uses a closed vocabulary of subword units
- can represent open vocabulary (including unknown words)
- minimizes the sequence length (given the vocabulary size)

solution

- greedy compression algorithm: byte pair encoding (BPE) [Gage, 1994]
- we adapt BPE to word segmentation
- hyperparameter: vocabulary size

vocabulary size	text
300	t+ h+ e i+ n+ d+ o+ o+ r t+ e+ m+ p+ e+ r+ a+ t+ u+ r+ e i+ s v+ e+ r+ y p+ l+ e+ a+ s+ a+ n+ t
1300	the in+ do+ or t+ em+ per+ at+ ure is very p+ le+ as+ ant
10300	the in+ door temper+ ature is very pleasant
50300	the indoor temperature is very pleasant

Deeper models

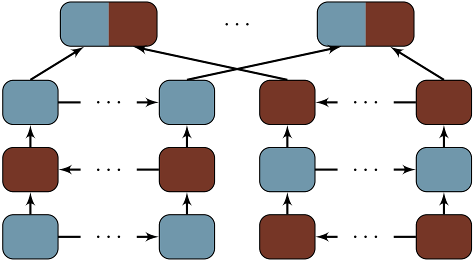


Figure: Alternating stacked encoder [Zhou et al., 2016].

Deep Transition Networks

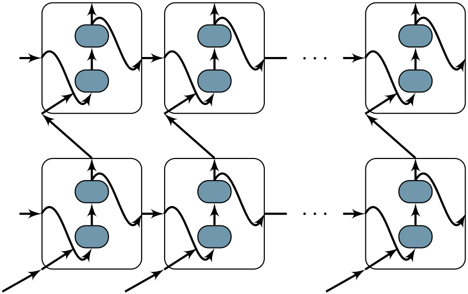


Figure: Illustration of BiDeep RNN architecture: stack of layers of recurrent cells; each cell is composed of multiple GRU transitions.

Semi-Supervised Training for NMT [**Sennrich**, Haddow, Birch, ACL 2016b]

why?

monolingual data

- is much less sparse than parallel data
- may be used for domain adaptation

why is this hard?

- standard in SMT: monolingual LM as feature in linear model
- linear combination of NMT and LM barely effective [Gülçehre et al., 2015]

our solution

end-to-end training of NMT model with parallel and monolingual data

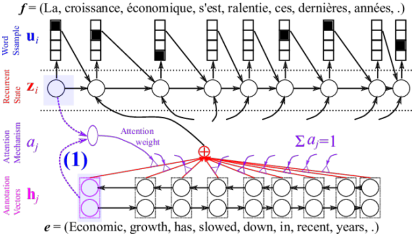
Monolingual Data in NMT

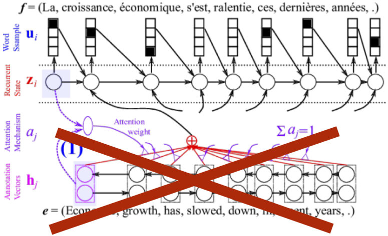
NMT is a conditional language model

$$p(u_i) = f(z_i, u_{i-1}, c_i)$$

Problem

for monolingual training instances, source context c_i is missing

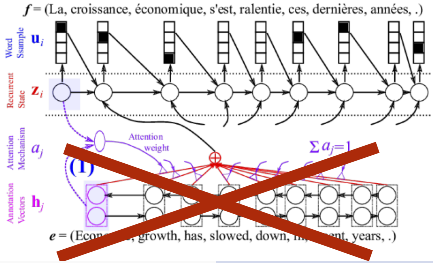




Monolingual Training Instances

solutions: missing data imputation for c_i

- missing data indicator: $\overrightarrow{0}$
  $\rightarrow$ works, but danger of catastrophic forgetting
- impute c_i with translation model
  $\rightarrow$ we do this indirectly by back-translating the target sentence
- make c_i a copy of the target
  $\rightarrow$ works especially for names and copied terms



Transfer learning

Harness resources from other language pairs

- Transfer learning takes knowledge gained while solving one task and applies it to a different but related task.
- Learning from parallel data in other languages: apply domain adaptation techniques to fine-tune multilingual MT models

Examples

- Zoph et al. (2016) train a high-resource language pair, then fine-tune on a low-resource language pair
  $\rightarrow$ French-English parent helps Hausa Turkish and Uzbeck to English child languages
- Nguyen and Chiang (2017) initially train on low-resource but related language pair
  $\rightarrow$ Uzbek-English parent helps and Turkish and Uyghur child languages

Bibliography I

Gage, P. (1994).
A New Algorithm for Data Compression.
C Users J., 12(2):23–38.

Gülçehre, c., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., and Bengio, Y. (2015).
On Using Monolingual Corpora in Neural Machine Translation.
CoRR, abs/1503.03535.

Zhou, J., Cao, Y., Wang, X., Li, P., and Xu, W. (2016).
Deep Recurrent Models with Fast-Forward Connections for Neural Machine Translation.
Transactions of the Association of Computational Linguistics – Volume 4, Issue 1, pages 371–383.