Memory-augmented Neural Machine Translation

Paper Review

Bojan Božić

December 1, 2017

Introduction

- NMT has highly promising performance for large training data.
- Common principle: encoding meaning of input into concept space and performing translation based on encoding $\rightarrow$ deeper understanding and learning of translation rules, better translation than SMT.
- Problem: tendency towards overfitting to frequent observations and overlooking special cases.
- Cause: Translational function is shared, so high- and low-frequency pairs impact each other by adapting shared parameters. Smoothness of translation function makes infrequent pairs seem like noise.

Neural Machine Translation

Is an approach to machine translation that uses a large neural network. It departs from phrase-based statistical approaches that use separately engineered subcomponents. E.g. Google uses Google Neural Machine Translation (GNMT) in preference to its previous statistical methods.

Problem: Low-frequency pairs

src.	人类共有二十三对染色体。
ref.	Humans have 23 pairs of chromosomes.
NMT	There are 23-year history of human history.

Table 1: An example of Chinese-to-English meaning drift with NMT.

Errors


带好孩子
Take the child
小心落水
Fall into water carefully


中國東方航空 上海航空公司
CHINA EASTERN SHANGHAI AIRLINES
请在一米线外等候
Please wait outside rice-flour noodle
www.engrish.com


文明观赏
请勿伤蛇
www.engrish.com
Do not hurt the snake viewing civilization


公共旅游卫生间
Public Toilet Tourism
Alamy

Statistic Machine Translation

- Based on statistics of words and phrases (i.e. symbolic method with discrete model).
- Discrete model = probability of infrequent pairs cannot be smoothed out.
- Lack of shared parameters = frequent pairs have much less impact on infrequent pairs.
- SMT memorises as many observed patterns as possible by using a phrase table.
- Ideal: Neural model with complementary statistical support.

Attention-based NMT

- Attention-based RNN model with encoder-decoder frame is used.
- MLP similarity function: $\alpha_{ij} = \frac{e_{ij}}{\sum e_{ik}}; e_{ij} = a(s_{i-1}, h_j)$
- Semantic content: $c_i = \sum \alpha_{ij} h_j$
- Update with recurrent function: $s_i = f_d(y_{i-1}, s_{i-1}, c_i)$
- Next word: $p(y_i) = \sigma(y_i^T W z_i)$
- Intermediate variable: $z_i = g(y_{i-1}, s_{i-1}, c_i)$

M-NMT Architecture

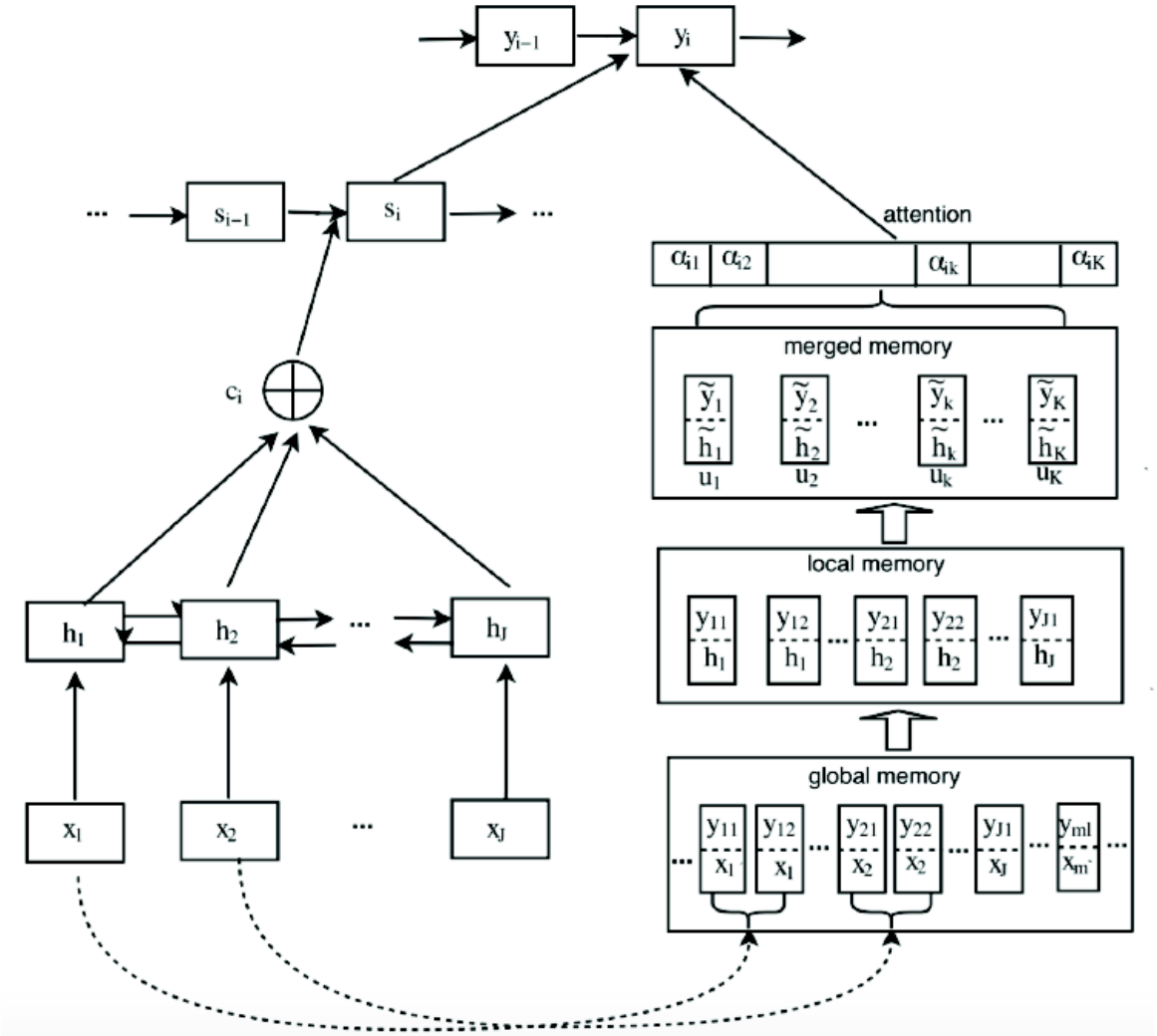

Figure 1: The structure of the M-NMT architecture.

Memory Elements

- Attention-based RNN model is good with frequent words, and memory elements provide knowledge of infrequent words.
- Memory element: $u_{jl} = \begin{bmatrix} y_{jl} \\ x_j \end{bmatrix}$
- Local memory: $u_{jl} = \begin{bmatrix} y_{jl} \\ h_j \end{bmatrix}$
- Compression: $u_k = \begin{bmatrix} \tilde{y}_k \\ \tilde{h}_k \end{bmatrix} = \begin{bmatrix} \tilde{y}_k \\ \sum_j p(x_j|\tilde{y}_k) h_j \end{bmatrix}; \forall \tilde{y}_k \in \{y_{jl}\}$

Memory Attention

- Attention factor: $\alpha_{ik}^{m} = \frac{e_{ik}^{m}}{\sum_{k=1}^{K} e_{ik}^{m}}$
- Relevance function: $e_{ik}^{m} = (v^{m})^{T} tanh(W_{s}^{m} s_{i-1} + W_{u}^{m} u_{k} + W_{y}^{m} y_{i-1})$
- Consolidated posterior: $\tilde{p}(y_i) = \beta \alpha_{ik}^{m} + (1 - \beta) p(y_i)$
- Objective function:

$$L(\theta) = \sum_{n} \sum_{i} log(\alpha_{ik_i^n}^{m})$$

Memory Integration

- Translation dictionary produced by SMT system is used.
- Align training sentence pairs using GIZA++ and apply intersection refinement rules to get a single 1:1 alignment for each sentence pair and extract translation dictionary.
- Key information is conditional probability that source and target word are translated to each other. This is used twice:
  1. conditional $p(y_{jl}|x_j)$ used to select target words y_{jl}
  2. conditional $p(x_j|\tilde{y}_k)$ used to merge element with target words $\tilde{y}_k$

OOV Treatment

- Manually defined dictionary specifies how to translate OOV words.
- Use this to construct local memory at run-time.
- When OOV is encountered the vector of a similar word is borrowed.
- Alternative choices should prevent collisions of the similar word.
- Problem: Vocabulary of neural model is fixed $\rightarrow$ no probabilities.
- Solution: Rewrite similar word by OOV word and redirect prediction.

Experiments

- Two datasets: small IWSLT (44k sentences from tourism domain) and large NIST (1M sentence pairs from LDC corpora)
- Memory construction: GIZA++ toolkit
- Baselines: conventional SMT and attention-based RNN NMT

M-NMT Configurations

System	Attending	Attended
$M-NMT(s, u^{y})$	s_{i-1}	$u_k(y)$
$M-NMT(s, u^{xy})$	s_{i-1}	$u_k(x), u_k(y)$
$M-NMT(sy, u^{y})$	s_{i-1}, y_{i-1}	$u_k(y)$
$M-NMT(sy, u^{xy})$	s_{i-1}, y_{i-1}	$u_k(x), u_k(y)$

Table 2: M-NMT systems with different configurations.

BLEU Scores

System	IWSLT05	NIST03
Moses	52.5	30.6
NMT	43.9	31.3
NMT-L	45.9	31.7
$M - NMT(s, u^y)$	49.8	32.3
$M - NMT(s, u^{xy})$	50.7	32.5
$M - NMT(sy, u^y)$	51.4	32.8
$M - NMT(sy, u^{xy})$	**52.9**	**34.0**

Table 3: BLEU scores with different translation systems on the two Chinese-English translation datasets.

OOV Recall Rates

	T-INV		T-OOV	
System	Recall	BLUE	Recall	BLEU
NMT	0.06	15.1	0	13.7
M-NMT	0.05	16.0	0	14.6
NMT-PL	0.09	15.4	0.08	14.3
M-NMT+OOV	**0.28**	**17.0**	**0.40**	**15.9**

Table 4: The OOV recall rates and BLEU scores on sentences with OOV words. 'T-INV' refers to the case where the target words of the OOV input are in-vocabulary, and 'T-OOV' means the case where the target words are also OOV.

Word Frequency

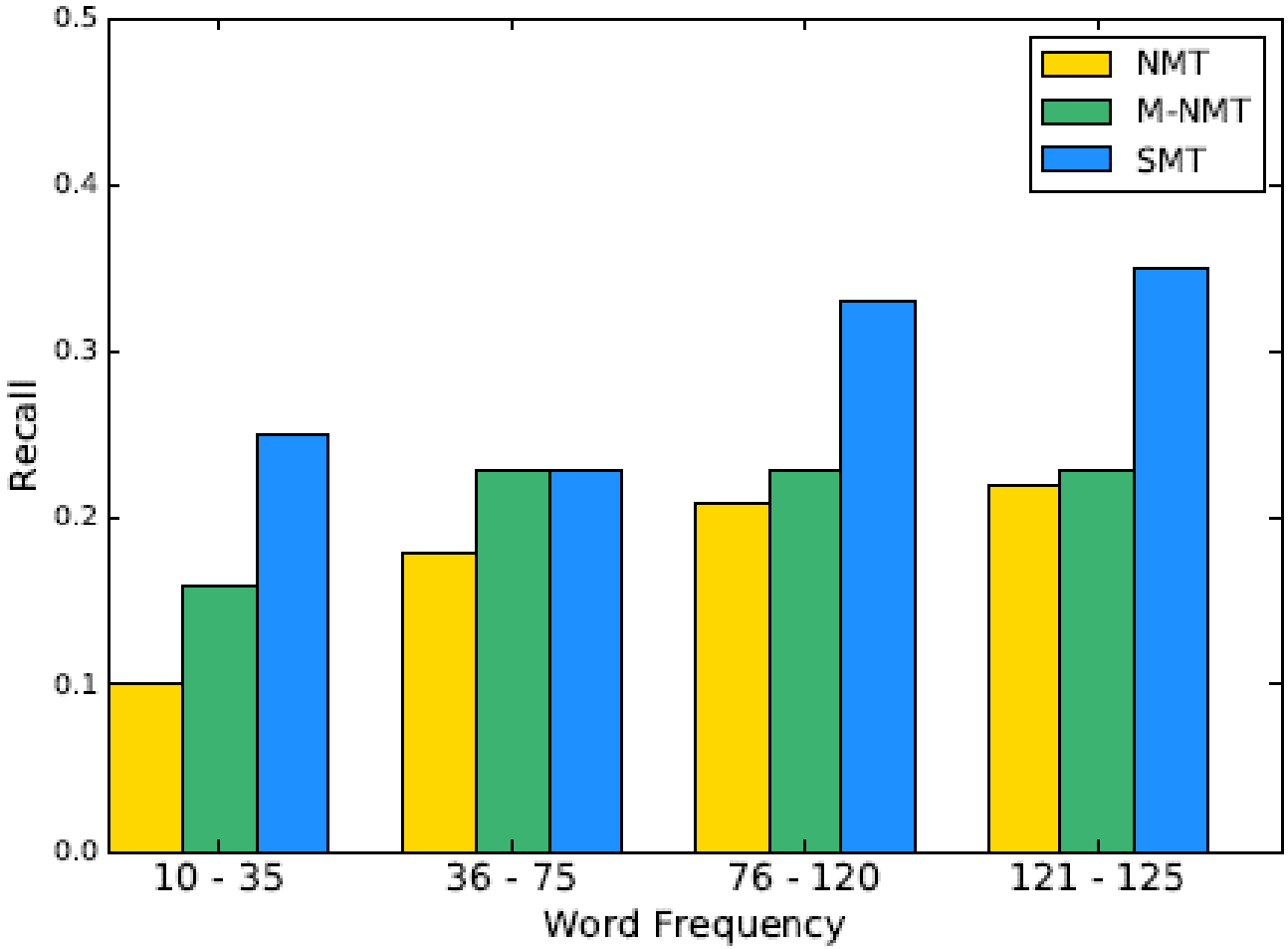

Figure 2: The recall rates of words in different frequency bins.

Translation Results

src.	人类共有二十三对染色体。
ref.	Humans have 23 pairs of chromosomes.
Moses	A total of 23 human chromosome.
NMT	There are 23-year history of human history.
M-NMT	There have a total of 23 species of chromosomes.

Table 5: The translations from different systems for the Chinese-to-English 'meaning drift' example.

Conclusion

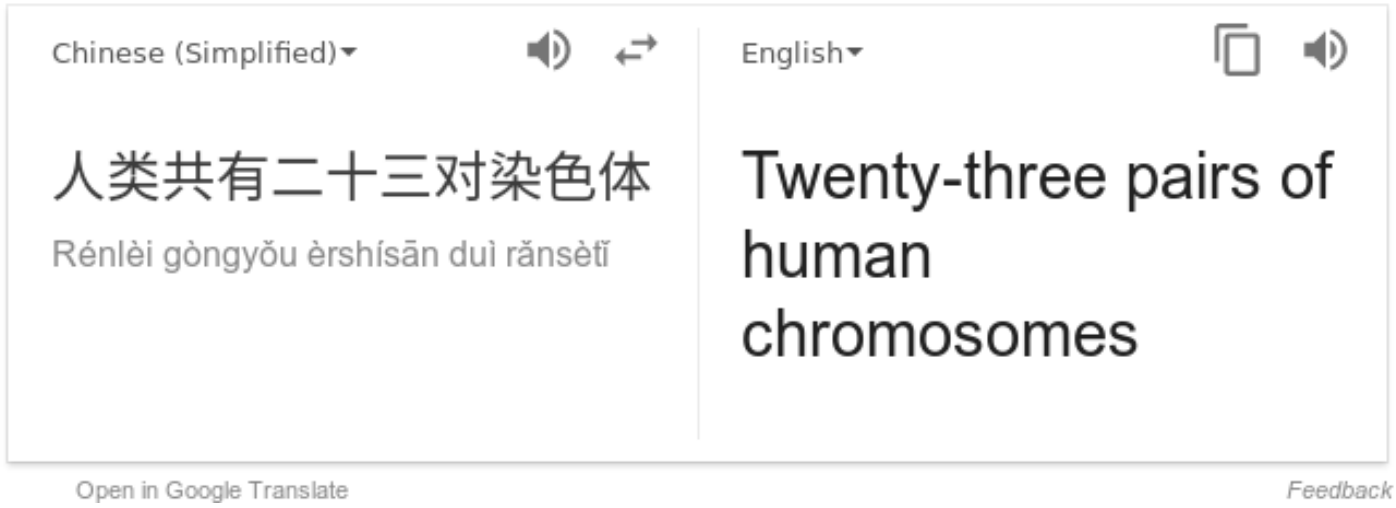

Figure 3: Even Google Translate does a decent job.

Conclusion

- Interesting approach (according to related work claims)
- Well defined experiments and metrics (although not always appropriate)
- Quite pessimistic goals and setting
- Lack of innovation and novelty
- But overall a good paper and well worth reading

Questions?