Paper Reading
Evaluating NMT

Longyue Wang

About evaluating NMT

Three papers from different layers/perspectives:

- **Human Evaluation**: does NMT outperforms professional human translator or not?

↓

- **Statistical Analysis**: analyse the syntactic properties in NMT outputs?

↓

- **Automatic Estimation**: the recent progress in Quality Estimation (QE，无reference自动评价)

Human Evaluation

Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

Samuel Laubli, Rico Sennrich, Martin Volk

本文**重新**探讨NMT翻译质量究竟有没有接近专业译员的翻译水平。假设：先前结论，即NMT与人工翻译质量等价，是在确实document-level context的情况下进行的，不准确。本文评价（adequacy和fluency）方法：

- 评估颗粒度：等级制（0-5）、打分制（0-100），**区分制**（NMT vs Human）
- Rater：crowd-sourcing SMT，**expert** NMT
- 单元：single sentence，**document-level**

Experiment & Results

- 123 Chinese-English articles from WMT 2017 test set
- Plot reversal in Adequacy
- Turning Fluency down
- It is time to shift towards document-level evaluation

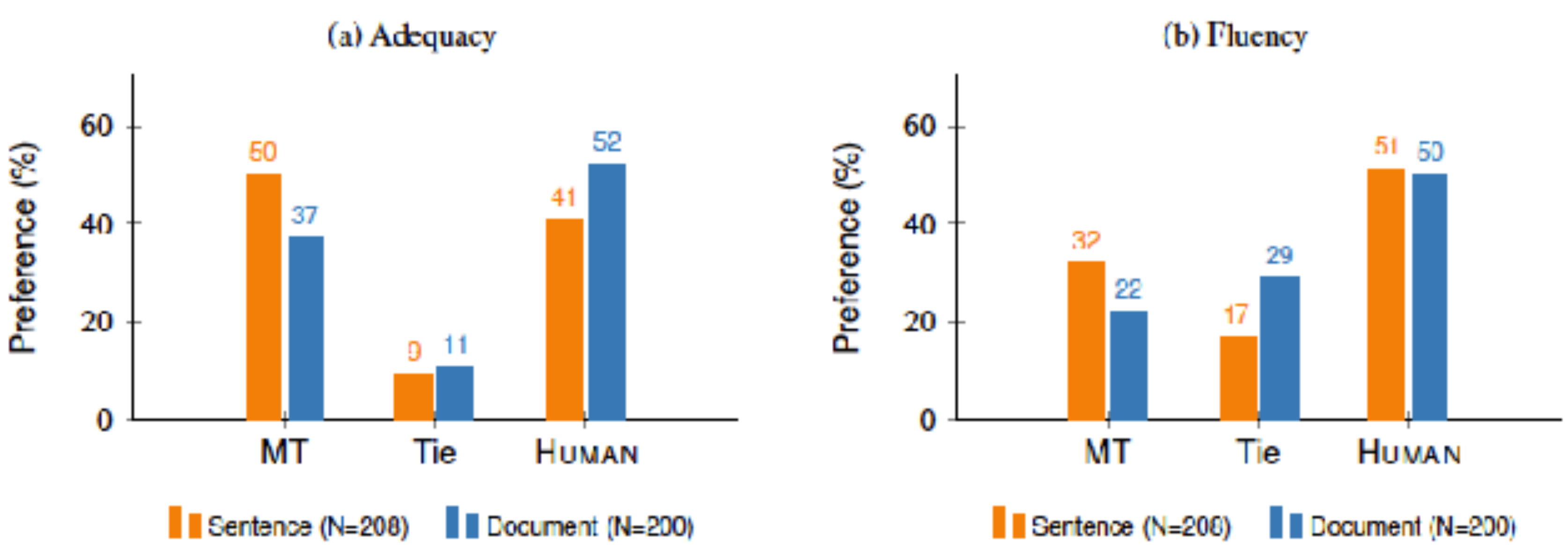


Figure 1: Raters prefer human translation more strongly in entire documents. When evaluating isolated sentences in terms of adequacy, there is no statistically significant difference between HUMAN and MT; in all other settings, raters show a statistically significant preference for HUMAN.

启发：

- 慎用“outperform”的句子
- document-level的人工和自动评价、NMT
- 总结document-level 的10篇NMT工作，梳理思路。可能会从速度优化角度去考虑新模型。

Syntactic Properties

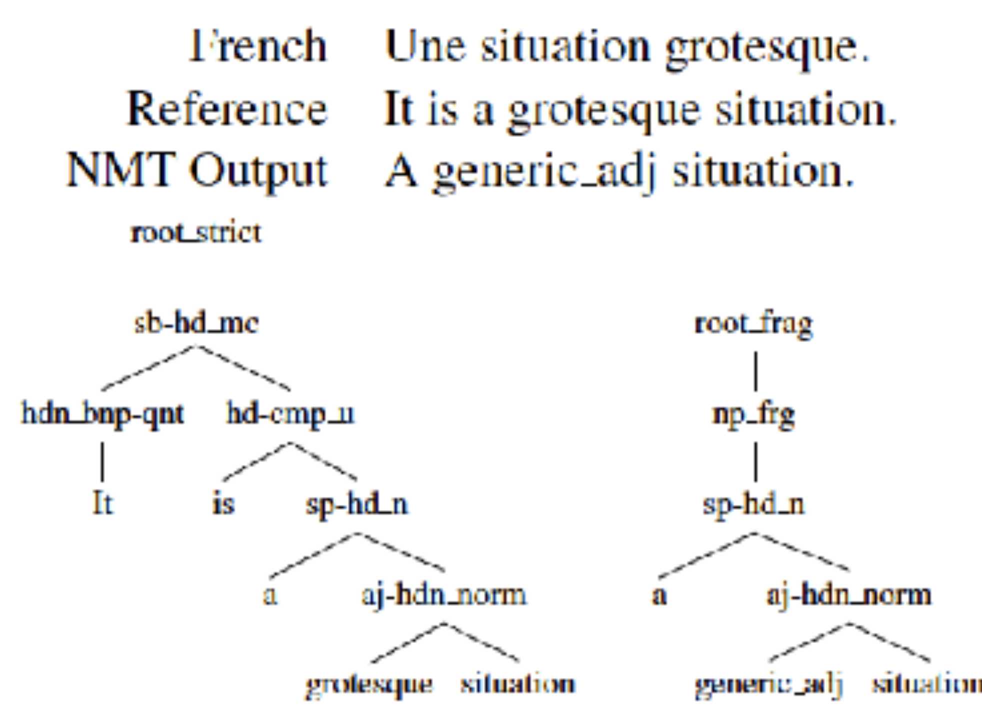
Evaluating Syntactic Properties of Seq2seq Output with a Broad Coverage HPSG: A Case Study on Machine Translation

Johnny Tian-Zheng Wei

seq2seq模型无法很好的理解语法属性。

本文引入了一个较新颖的角度来定量和定性地分析NMT翻译结果，即“符合语法规则的(grammartically)”的程度。发现模型在rarer syntactic rules上学习能力不足。

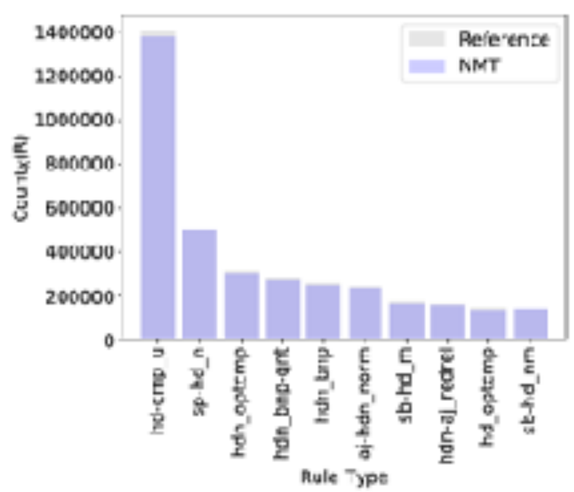
- 充分讨论HPSG-based English Resource Grammar可以用来分析译文的grammartical与否的工具：
  - 85% wiki can be parsed by ERG
  - fine-grained labels of linguistic constructions
  - unlike statistical parsers, these grammars are hand-built

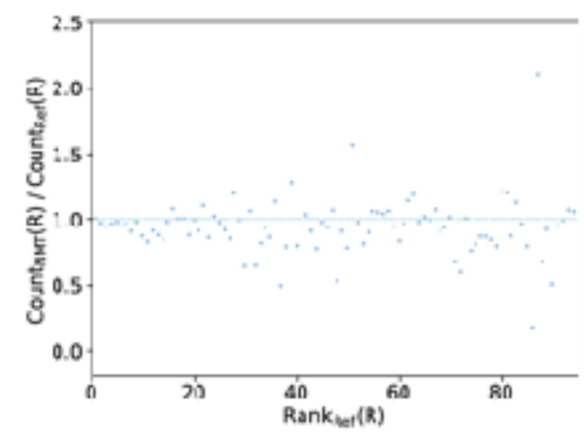


- 利用LKB/PET ERG parser对翻译结果和reference解析成树形结构，对比多个指标上在Pearson Correlation的一致性。
  - 1.4M EN-FR ERG-parseble sentences and 200K EN-FR for analysis
  - 93% outputs can be parsed, NMT is a little bit better than unigram-model

Feature	Equation	r
LP NMT	$\log P_m(S_o)$	0.313
LP Unigr. (src-fr)	$\log P_u(S_i)$	0.289
LP Unigr. (ref-en)	$\log P_u(S_r)$	0.273
LP Unigr. (out-en)	$\log P_u(S_o)$	0.304

- 定性分析不能parsed的output，及统计规则片段。
    - 37% 仍然是符合语法规则的，有一定gap。
    - 对于规则学习有歧视性：使用少的学的差。





    - NMT有源语侵蚀现象，即将源语端的语法规则直接用到了目标语，造成很多语法错误。

French	je le répète , vous avez raison .
Reference	i repeat ; you are quite right .
NMT Output	i repeat , you are right .

French	quel paradoxe !
Reference	what a paradox this is !
NMT Output	what a paradox !

启发：

- 可以用GRG parsing去量化语法规则的学习情况。
- 语言磨蚀(语蚀，Language Attrition)是语言学习过程中语言能力减弱或损失现象。
- 在机器翻译中的现象是，源语中的patent被错误地transfer到了目标语中。

Quality Estimation

Contextual Encoding for Translation Quality Estimation

Junjie Hu, Wei-Cheng Chang, Yuexin Wu, Graham Neubig

Word-level QE任务，是在无参考答案下预测每个词翻译的好坏（序列标注任务），本文是WMT18第一名系统。

前期工作很少考虑local context与target word之间的交互。本文提出利用CNN+RNN的混合模型，可以更好的关注short-term 和 long-term 的关系。此外，为了达到最好效果，本文还引入了POS、alignment、人工feature信息：

1，ok与bad的标签数据不平衡导致模型预测bias

2，人工feature最后拼接到FF中可以进一步提升，neural 和 人工feature的融合

3，CNN 可以很好的capture local information

- CNN可以更好的学习周围词汇的pattent
- RNN encoding 可以refine学习到的知识，具体地FNN与Bi-GRU交替多层
- 在输出之前再加入人工feature（maxon）

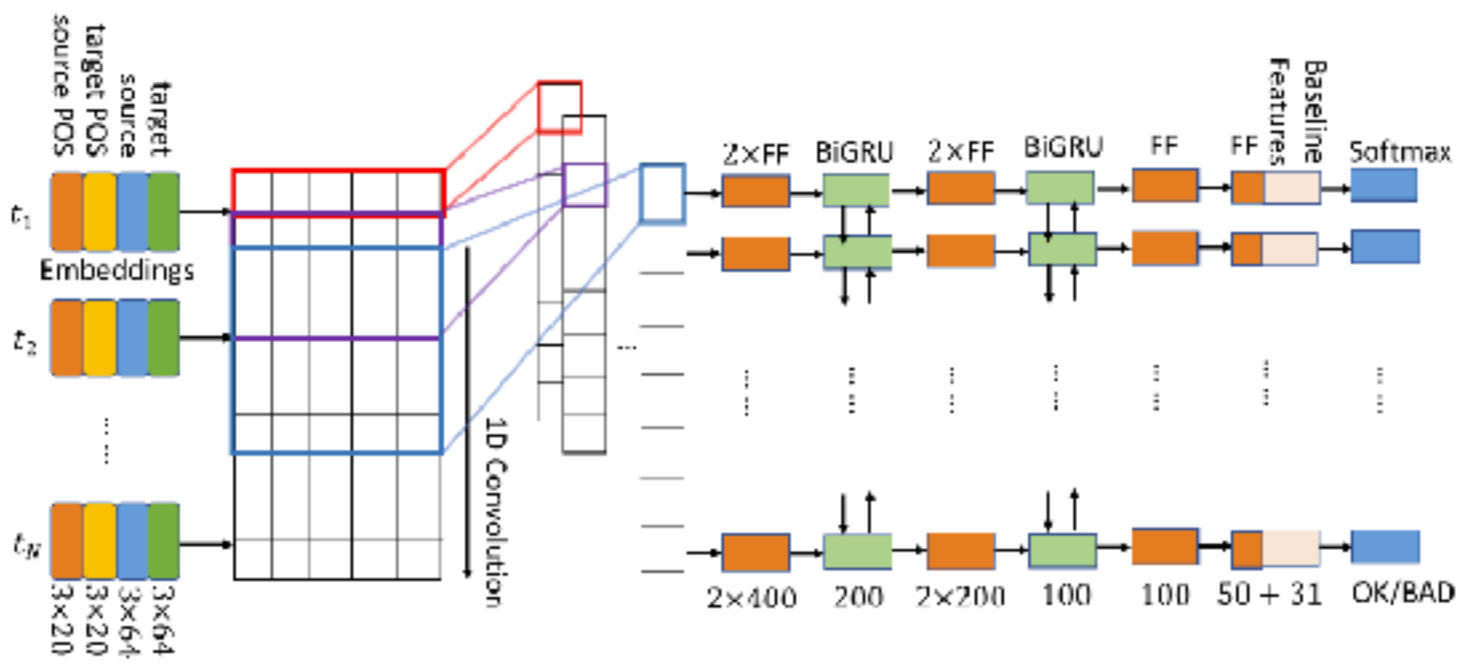
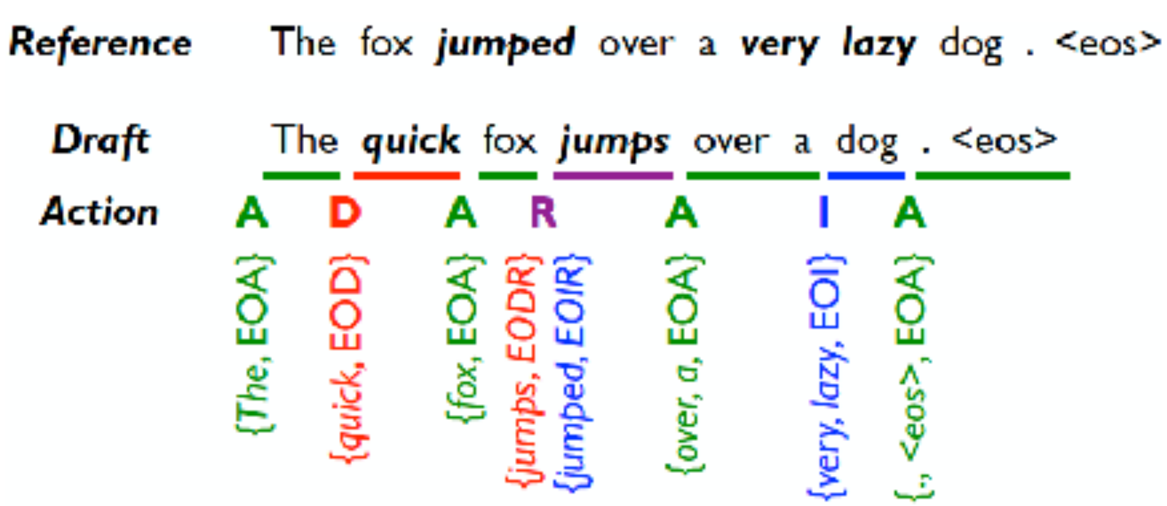


Figure 1: The architecture of our model, with the convolutional encoder on the left, and stacked RNN on the right.

Language Pairs	F1-BAD	F1-OK	F1-Multi	Rank
En-De (SMT)	0.5075	0.8394	0.4260	3
En-De (NMT)	0.3565	0.8827	0.3147	2
De-En	0.4906	0.8640	0.4239	2
En-Lv (SMT)	0.4211	0.8592	0.3618	1
En-Lv (NMT)	0.5192	0.8268	0.4293	1
En-Cz	0.5882	0.8061	0.4741	1

启发：

- 思考Learning to Revise工作



- Multi-task learning / reinforcement learning