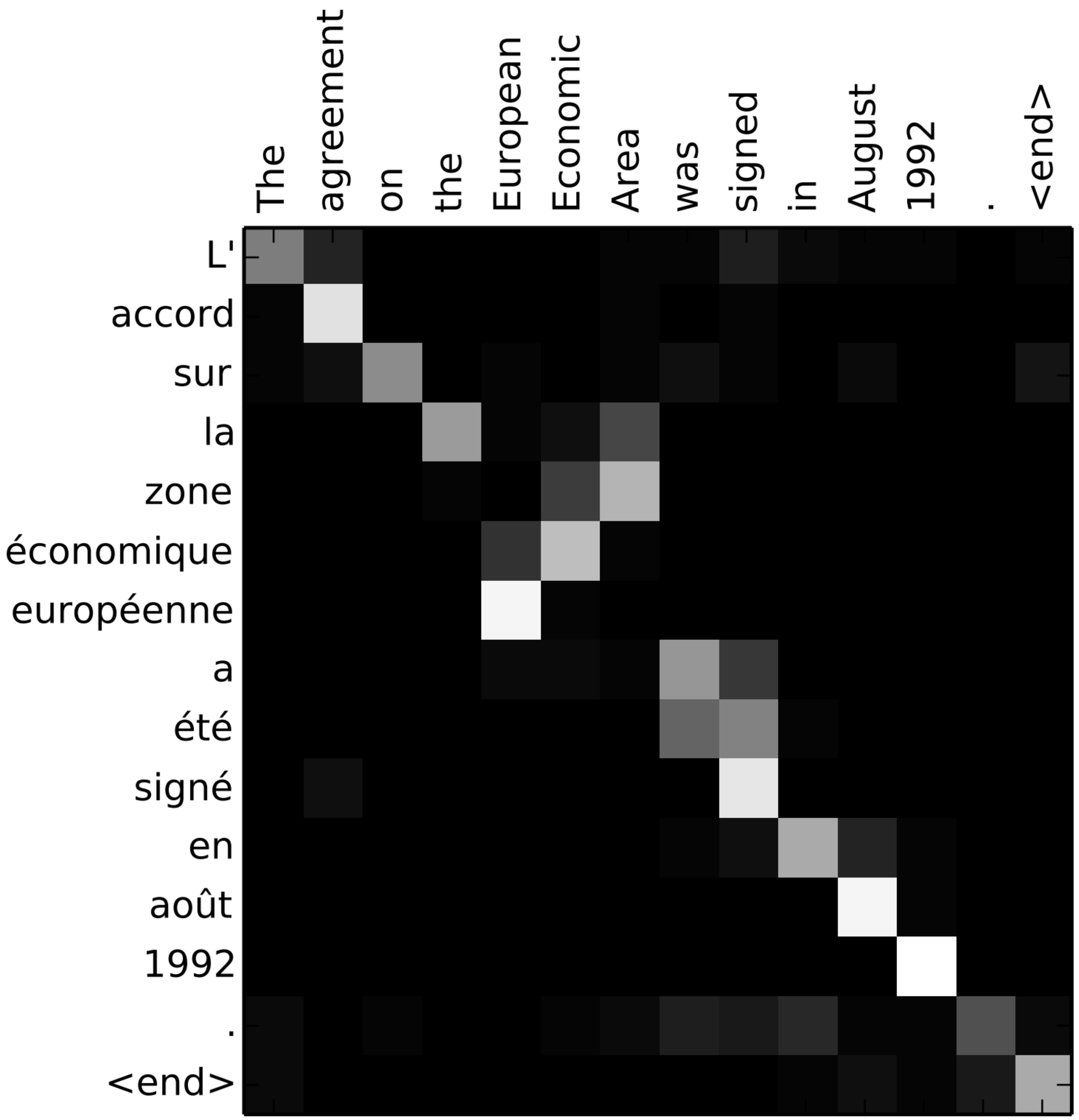
Attention

Attention

- Encoder hidden states capture contextual source word identity
- Decoder hidden states are now mostly responsible for selecting what to attend to
- Doesn’t take a complex hidden state to walk monotonically through a sentence and spit out word-by-word translations



Neural MT

Results: WMT English-French

- 12M sentence pairs

Classic phrase-based system: ~**33** BLEU, uses additional target-language data

Rerank with LSTMs: **36.5** BLEU (long line of work here; Devlin+ 2014)

Sutskever+ (2014) seq2seq single: **30.6** BLEU (input reversed)

Sutskever+ (2014) seq2seq ensemble: **34.8** BLEU

Luong+ (2015) seq2seq ensemble with attention and rare word handling: **37.5** BLEU

- But English-French is a really easy language pair and there's *tons* of data for it! Does this approach work for anything harder?

Results: WMT English-German

- 4.5M sentence pairs

Classic phrase-based system: **20.7** BLEU

Luong+ (2014) seq2seq: **14** BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: **23.0** BLEU

- Not nearly as good in absolute BLEU, but BLEU scores aren't really comparable across languages
- French, Spanish = easiest
  German, Czech = harder
  Japanese, Russian = hard (grammatically different, lots of morphology…)

MT Examples

src	In einem Interview sagte Bloom jedoch , dass er und Kerr sich noch immer lieben .
ref	However , in an interview , Bloom has said that he and *Kerr* still love each other .
best	In an interview , however , Bloom said that he and *Kerr* still love .
base	However , in an interview , Bloom said that he and **Tina** were still <unk> .

- best = with attention, base = no attention
- NMT systems can hallucinate words, especially when not using attention — phrase-based doesn't do this

Luong et al. (2015)

MT Examples

src	Wegen der von Berlin und der Europäischen Zentralbank verhängten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhalten an der gemeinsamen Währung genötigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The *austerity imposed by Berlin and the European Central Bank , coupled with the straitjacket* imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict *austerity measures imposed by Berlin and the European Central Bank in connection with the straitjacket* in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

- best = with attention, base = no attention

Luong et al. (2015)

MT Examples

Source	such changes in reaction conditions include , but are not limited to , an increase in temperature or change in ph .
Reference	所(such) 述(said) 反 应(reaction) 条 件(condition) 的(of) 改 变(change) 包 括(include) 但(but) 不(not) 限 于(limit) 温度(temperature) 的(of) 增加(increase) 或(or) pH 值(value) 的(of) 改变(change) 。
PBMT	中(in) 的(of) 这 种(such) 变 化(change) 的(of) 反 应(reaction) 条 件(condition) 包 括(include) , 但(but) 不(not) 限 于(limit) , 增加(increase) 的(of) 温度(temperature) 或(or) pH 变化(change) 。
NMT	这种(such) 反应(reaction) 条件(condition) 的(of) 变化(change) 包括(include) 但(but) 不(not) 限于(limit) pH 或(or) pH 的(of) 变化(change) 。

- NMT can repeat itself if it gets confused (pH or pH)
- Phrase-based MT often gets chunks right, may have more subtle ungrammaticalities

Zhang et al. (2017)

Handling Rare Words

en: The *ecotax* portico in *Pont-de-Buis* , . . . [truncated] . . . , was taken down on Thursday morning

fr: Le *portique* *écotaxe* de *Pont-de-Buis* , . . . [truncated] . . . , a été *démonté* jeudi matin

nn: Le `unk` de `unk` à `unk` , . . . [truncated] . . . , a été pris le jeudi matin

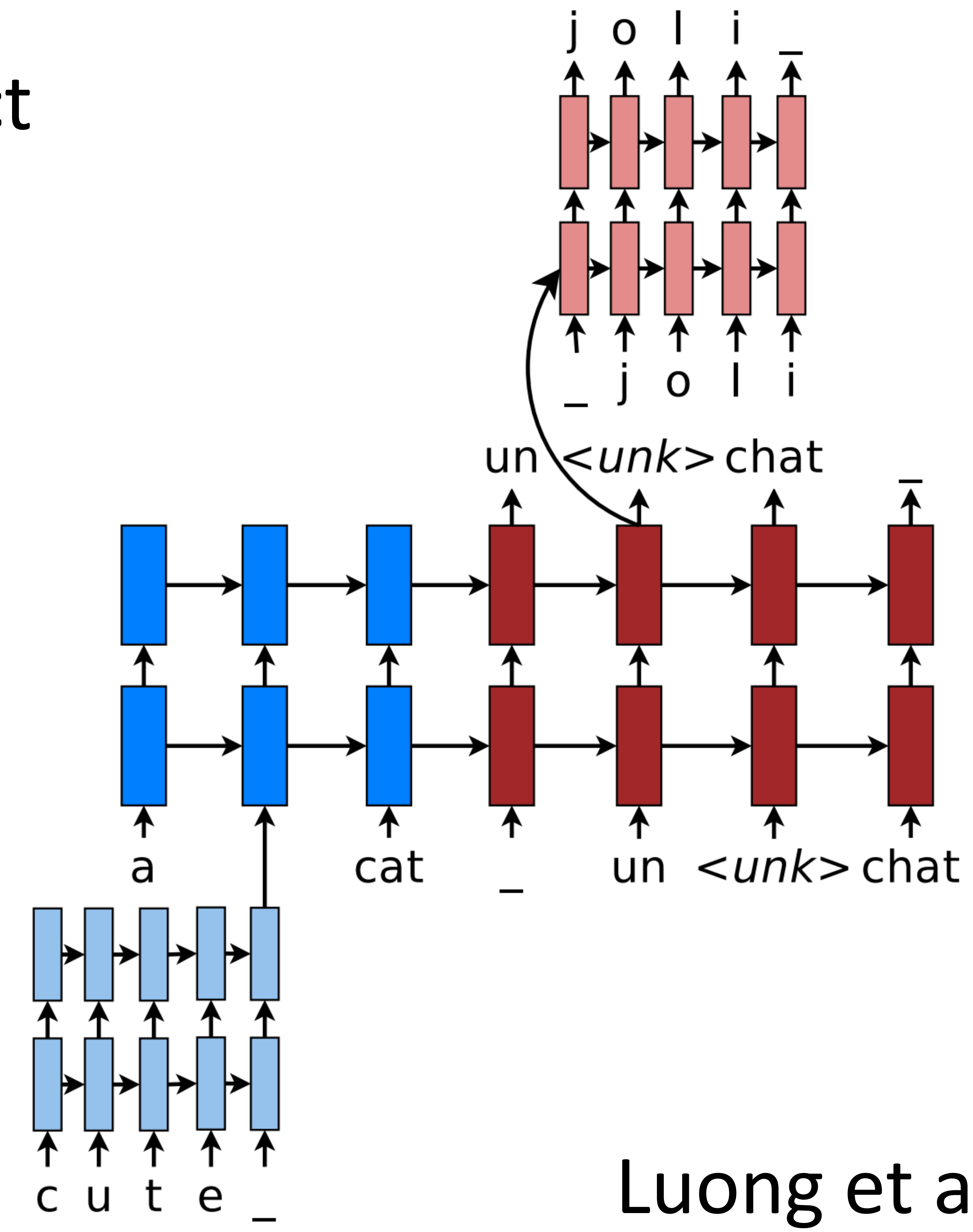
- Need to transliterate or copy OOV words

Jean et al. (2015), Luong et al. (2015)

Character-level Approaches

- Hybrid word-character models: predict unk and then “switch into” character generation mode

- Hard to handle, does not parallelize well



Luong et al. (2016)

Word Piece Models

- Use Huffman encoding on a corpus, keep most common *k* (~10,000) character sequences for source and target

Input: _the **_eco tax** _port ico _in _Po nt - de - Bu is …

Output: _le _port ique **_éco taxe** _de _Pont - de - Bui s

- Captures common words and parts of rare words
- Subword structure may make it easier to translate
- Model balances translating and transliterating without explicit switching

Wu et al. (2016)

Rare Words: Byte Pair Encoding

- Simpler procedure, based only on the dictionary
- Input: a dictionary of words represented as characters

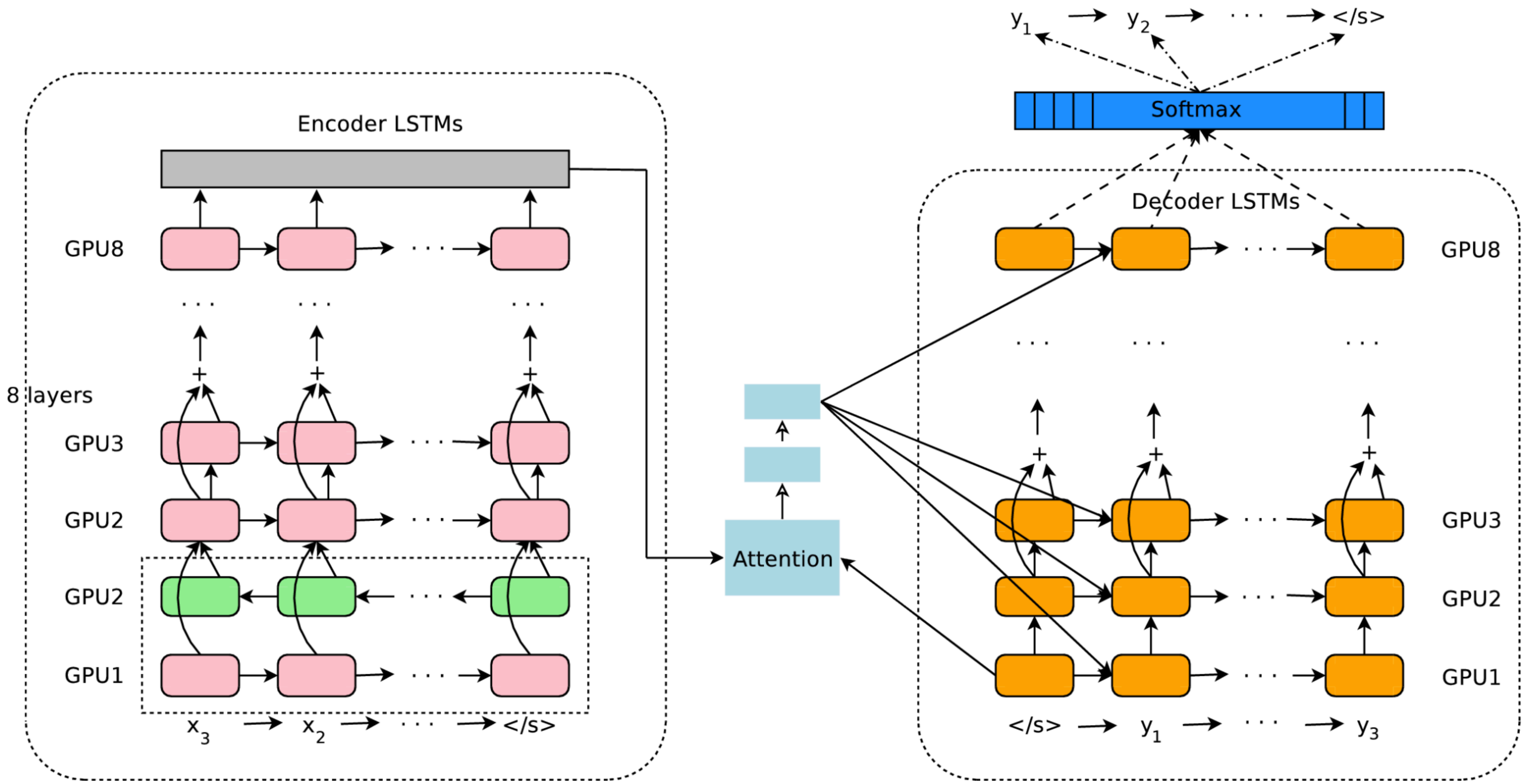
```
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
```

- Count bigram character cooccurrences
- Merge the most frequent pair of adjacent characters

- Final size = initial vocab + num merges. Often do 10k - 30k merges
- Most SOTA NMT systems use this on both source + target

Sennrich et al. (2016)

Google's NMT System



- 8-layer LSTM encoder-decoder with attention, word piece vocabulary of 8k-32k

Wu et al. (2016)

Google’s NMT System

English-French:

Google’s phrase-based system: 37.0 BLEU

Luong+ (2015) seq2seq ensemble with rare word handling: 37.5 BLEU

Google’s 32k word pieces: 38.95 BLEU

English-German:

Google’s phrase-based system: 20.7 BLEU

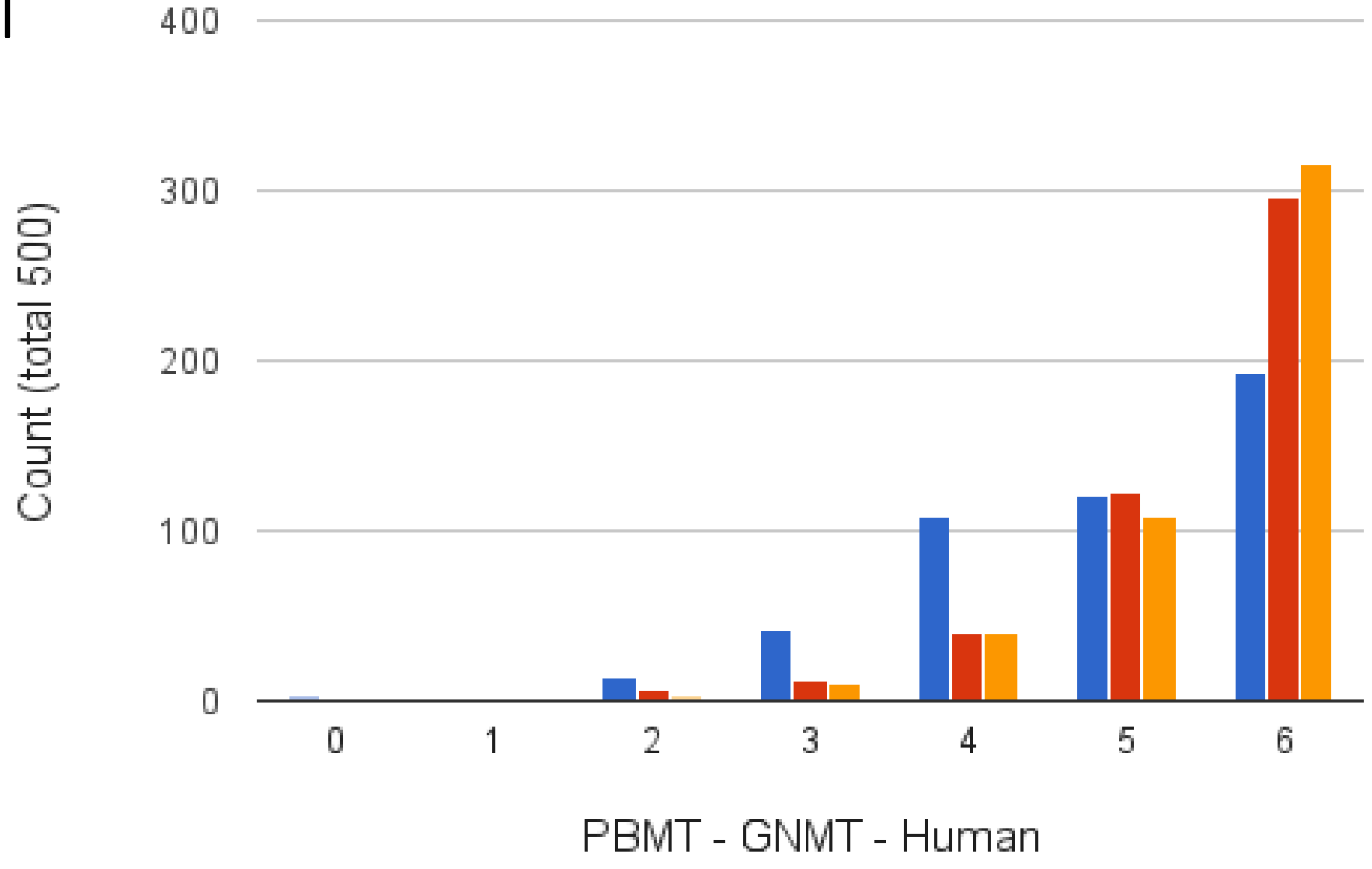
Luong+ (2015) seq2seq ensemble with rare word handling: 23.0 BLEU

Google’s 32k word pieces: 24.2 BLEU

Wu et al. (2016)

Human Evaluation (En-Es)

- Similar to human-level performance *on English-Spanish*



Wu et al. (2016)

Google's NMT System

Source	She was spotted three days later by a dog walker trapped in the quarry	
PBMT	Elle a été repéré trois jours plus tard par un promeneur de chien piégé dans la carrière	6.0
GNMT	Elle a été repérée trois jours plus tard par un traîneau à chiens piégé dans la carrière.	2.0
Human	Elle a été repérée trois jours plus tard par une personne qui promenait son chien coincée dans la carrière	5.0

Gender is correct in GNMT but not in PBMT

"sled"

"walker"

Wu et al. (2016)

Backtranslation

- Classical MT methods used a bilingual corpus of sentences B = (S, T) and a large monolingual corpus T' to train a language model. Can neural MT do the same?

- Approach 1: force the system to generate T' as targets from null inputs

  s_1, t_1
  s_2, t_2
  …
  [null], t'_1
  [null], t'_2
  …

- Approach 2: generate synthetic sources with a T->S machine translation system (backtranslation)

  s_1, t_1
  s_2, t_2
  …
  MT(t'_1), t'_1
  MT(t'_2), t'_2
  …

Sennrich et al. (2015)