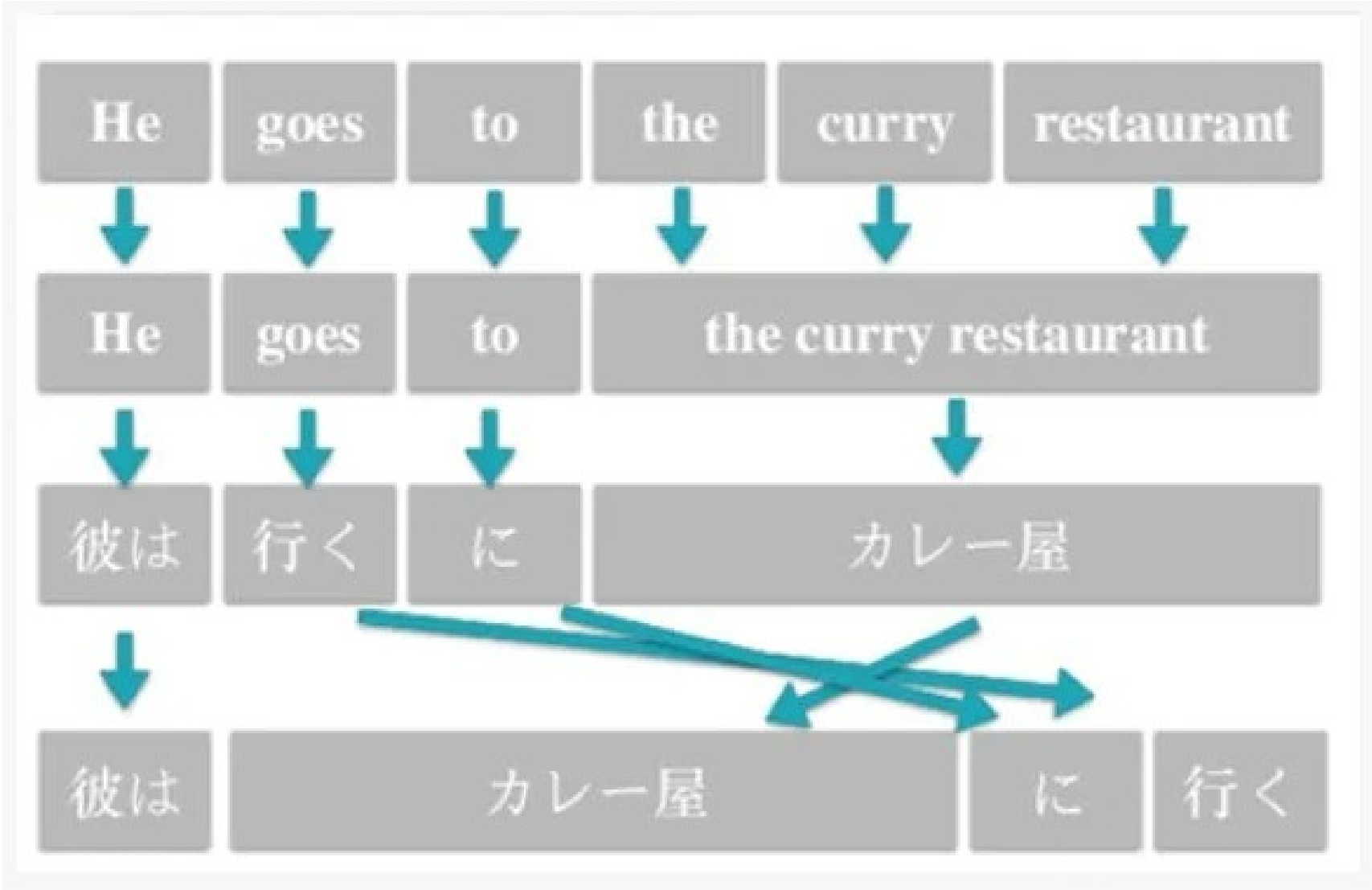
SEQUENCE TO SEQUENCE LEARNING WITH NEURAL NETWORKS
&
BLEU: A METHOD FOR AUTOMATIC EVALUATION OF MACHINE TRANSLATION

Nilesh Kumar Srivastava
2022.04.22

Contents

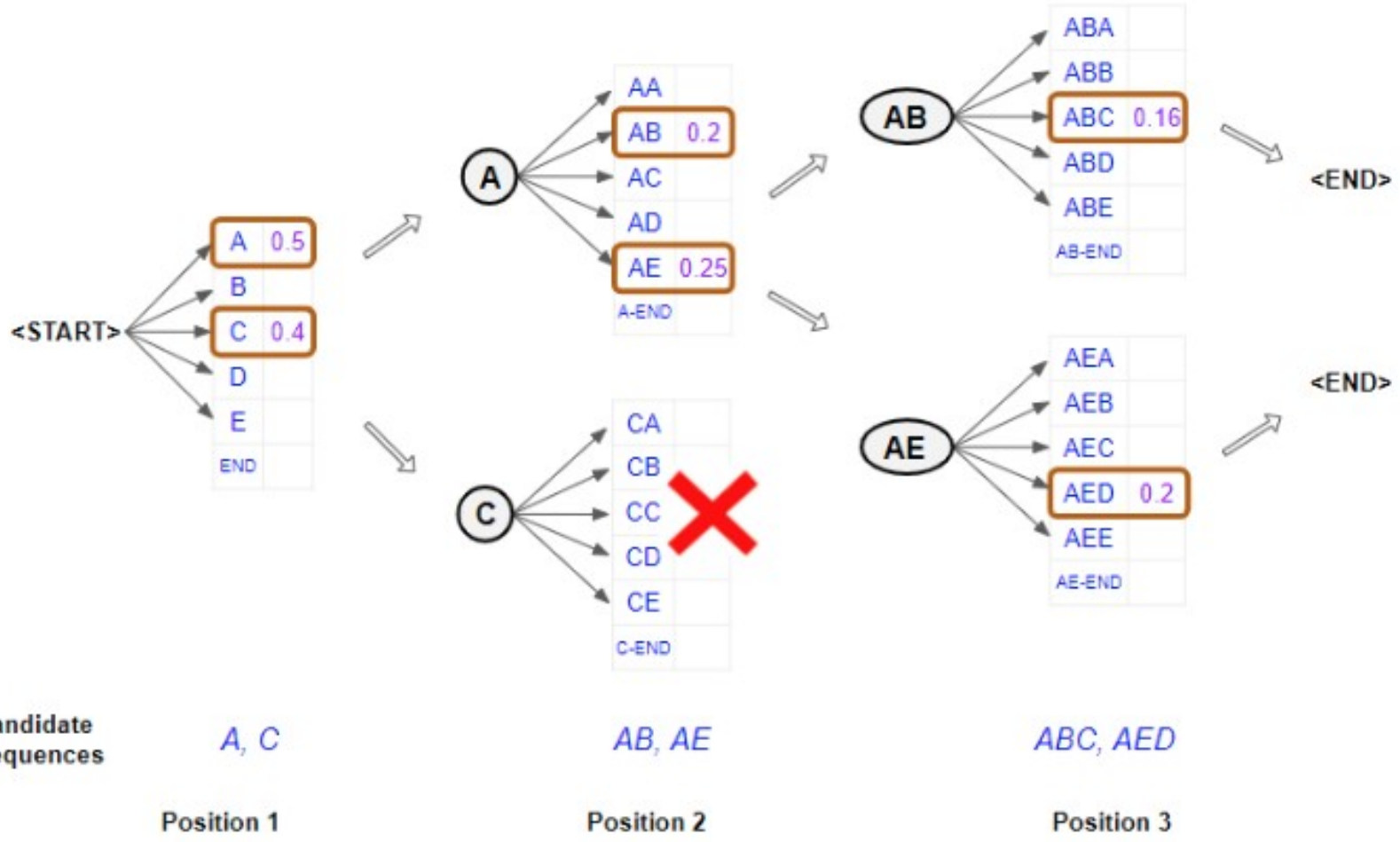
Overview : SMT

- Phrase
- Translate
- Rearrange



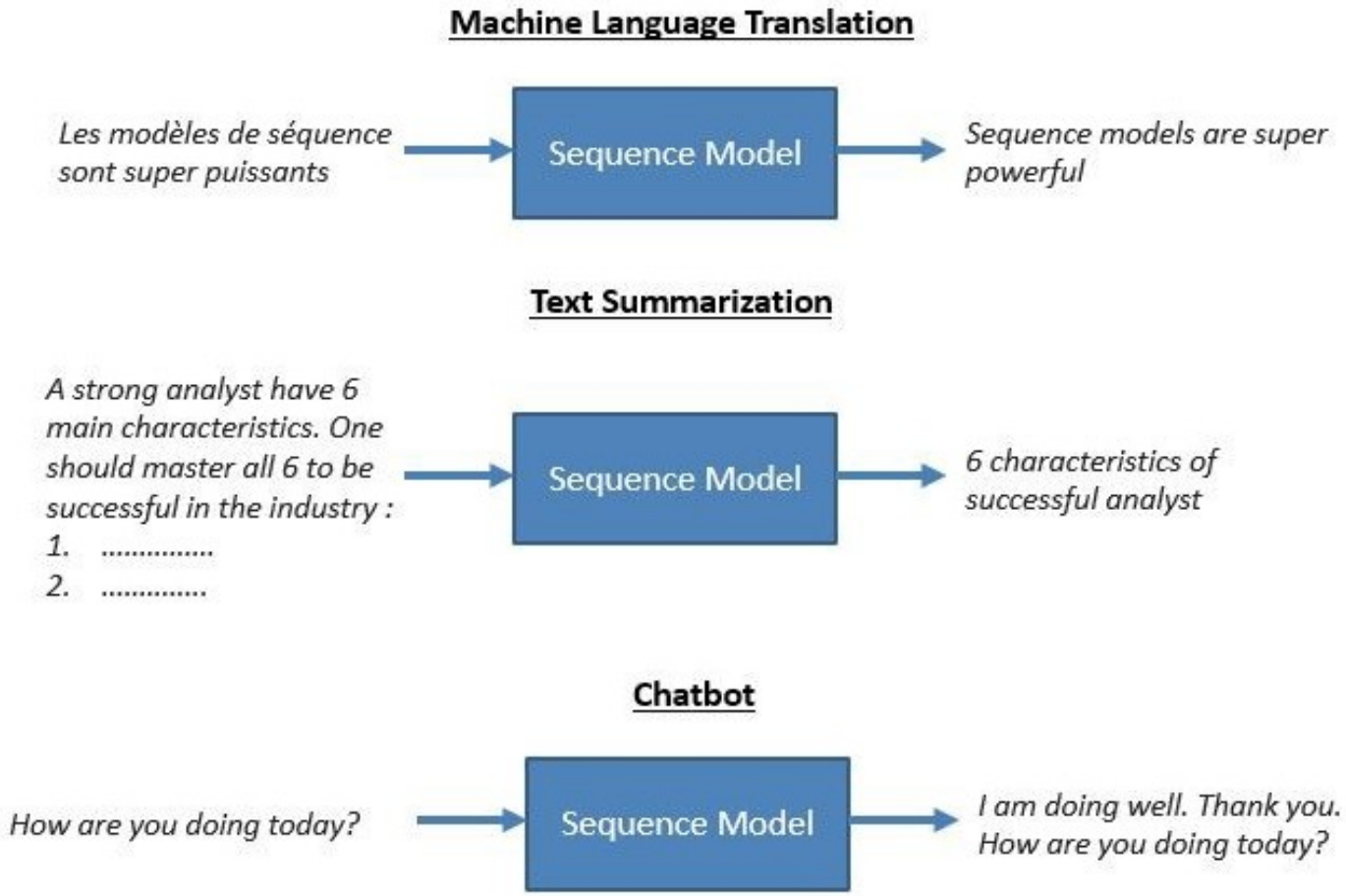
Source: https://kantanmtblog.com/2019/04/02/a-short-introduction-to-the-statistical-machine-translation-model/

Overview: Beam Search



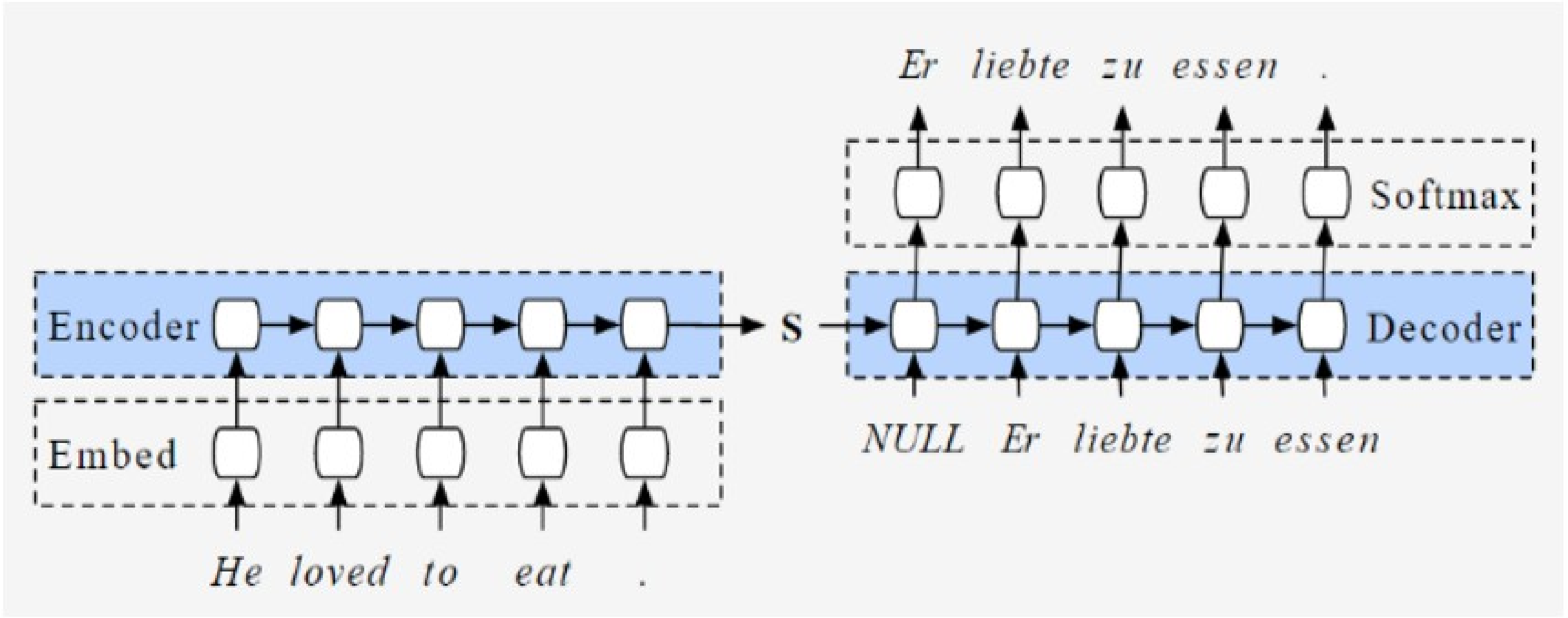
Source: https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24

Overview: Sequence to Sequence Models - Applications



Source: https://www.analyticsvidhya.com/blog/2020/08/a-simple-introduction-to-sequence-to-sequence-models/

Overview : Sequence to Sequence Models - Architechture



Source: https://www.analyticsvidhya.com/blog/2020/08/a-simple-introduction-to-sequence-to-sequence-models/

The Model

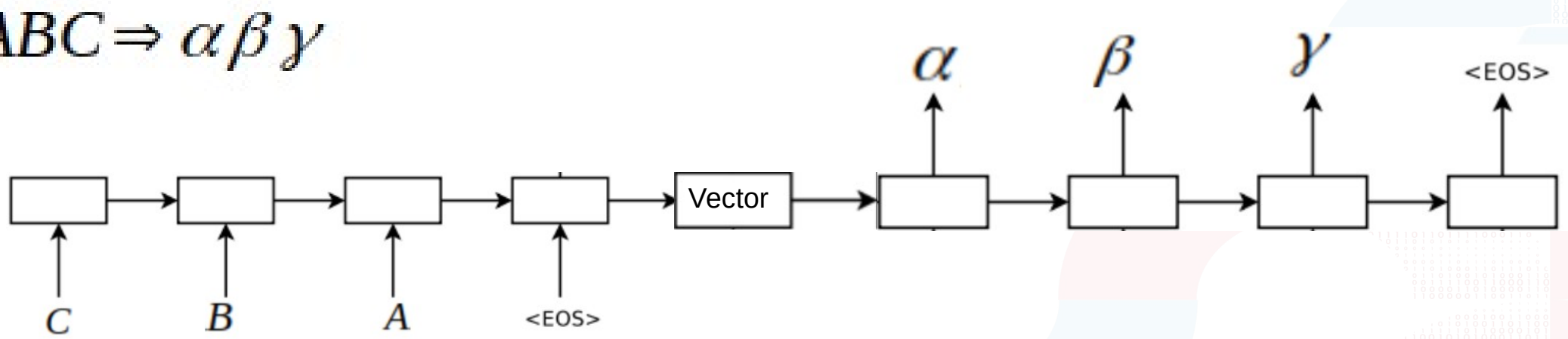
- LSTM Cells

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1})$$

  - one for the input sequence
  - another for the output sequence

- Deep LSTM (4 layers)

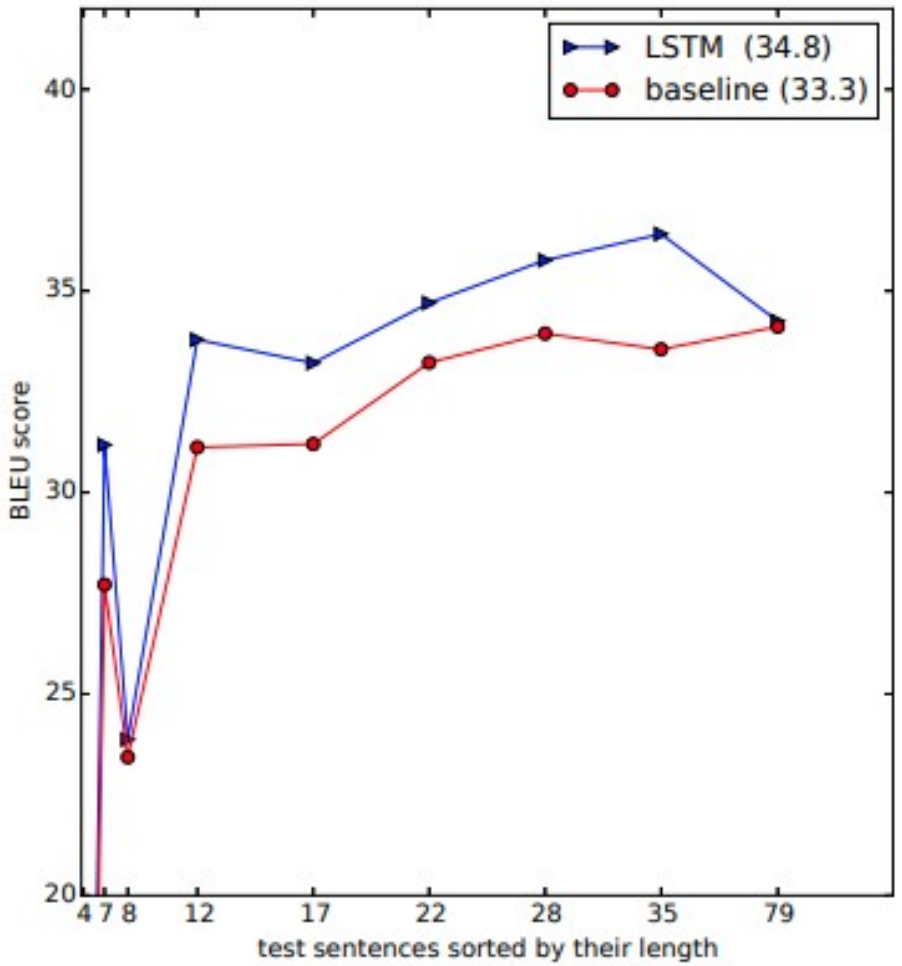
- Reversed the order of Input

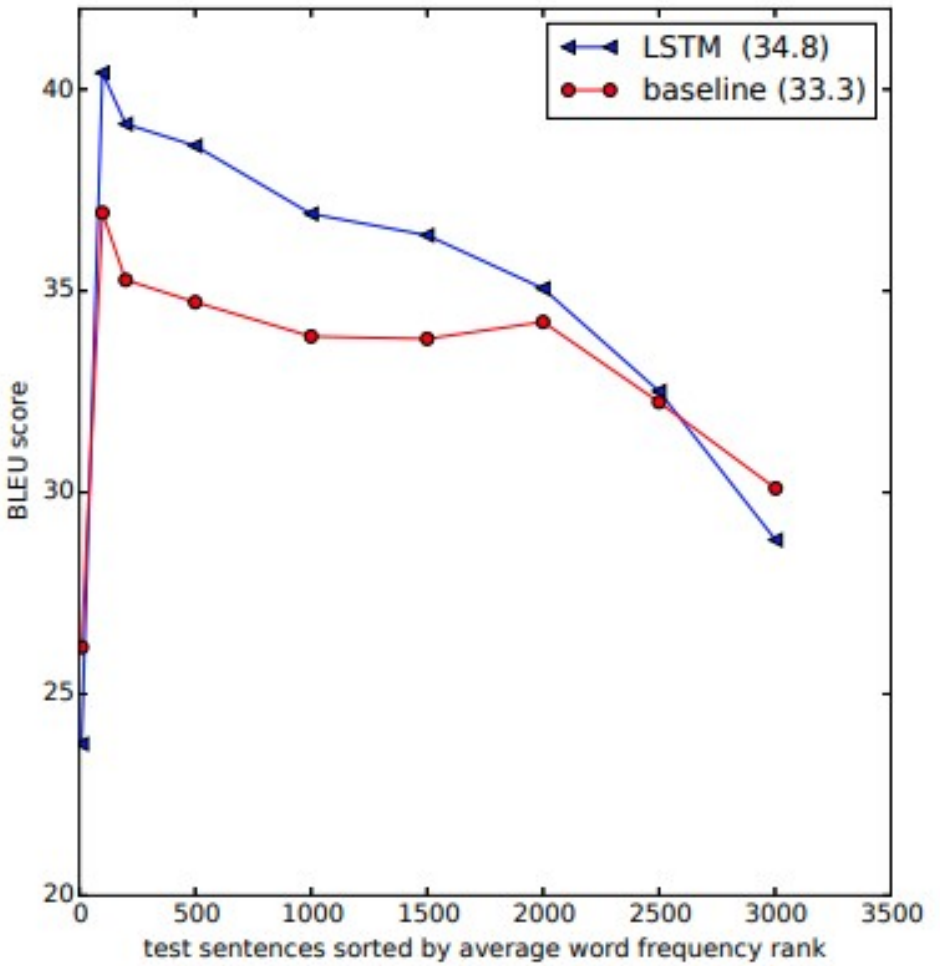


Model Analysis

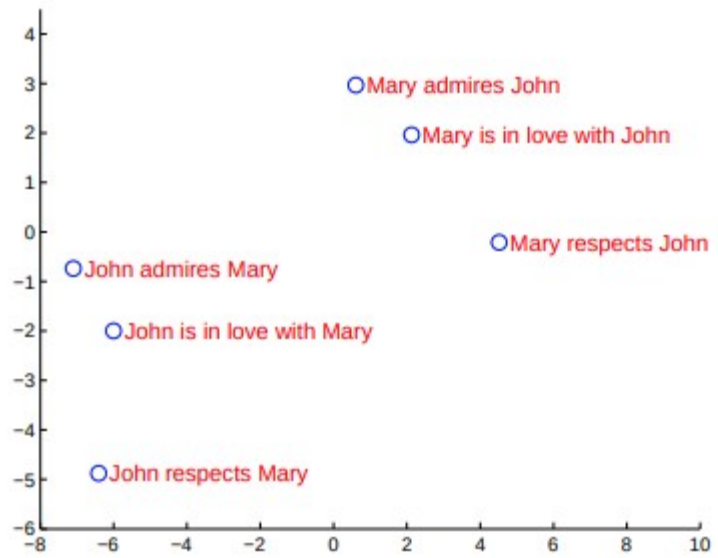
Type	Sentence
Our model	Ulrich UNK , membre du conseil d' administration du constructeur automobile Audi , affirme qu' il s' agit d' une pratique courante depuis des années pour que les téléphones portables puissent être collectés avant les réunions du conseil d' administration afin qu' ils ne soient pas utilisés comme appareils d' écoute à distance .
Truth	Ulrich Hackenberg , membre du conseil d' administration du constructeur automobile Audi , déclare que la collecte des téléphones portables avant les réunions du conseil , afin qu' ils ne puissent pas être utilisés comme appareils d' écoute à distance , est une pratique courante depuis des années .
Our model	" Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu' ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu' ils pourraient interférer avec les tours de téléphone cellulaire lorsqu' ils sont dans l' air " , dit UNK .
Truth	" Les téléphones portables sont véritablement un problème , non seulement parce qu' ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d' après la FCC , qu' ils pourraient perturber les antennes-relais de téléphonie mobile s' ils sont utilisés à bord " , a déclaré Rosenker .
Our model	Avec la crémation , il y a un " sentiment de violence contre le corps d' un être cher " , qui sera " réduit à une pile de cendres " en très peu de temps au lieu d' un processus de décomposition " qui accompagnera les étapes du deuil " .
Truth	Il y a , avec la crémation , " une violence faite au corps aimé " , qui va être " réduit à un tas de cendres " en très peu de temps , et non après un processus de décomposition , qui " accompagnerait les phases du deuil " .

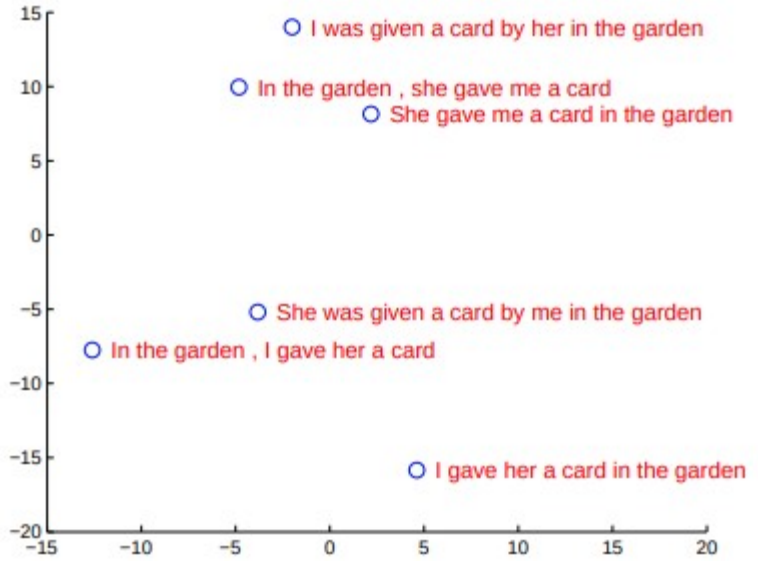
Model Analysis





Model Analysis





Model Results

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
Best WMT'14 result [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

BLEU - Background

Human Evaluation aspects:

- Adequacy
- Fidelity
- Fluency

Challenges:

- Time consuming
- Expensive

Requirements:

- A numerical “translation closeness” metric
- A corpus of good quality human reference translations

BLEU – Baseline Metric

Translation from a Chinese source:

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Human Translation:

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Modified Precision

Precision = unigram match in translation / total words in the candidate translation

Candidate : the the the the the the the
Reference 1 : The cat is on the mat.
Reference 2 : There is a cat on the mat.

Precision = 7/7 = 1

Modified Unigram precision = Σ(min(count,max_ref_count)) / total word count = 2/7

This approach address the two main aspects of translation:

- Adequacy (same words)
- Fluency (longer n-gram matches)

Precision simulations

Figure 1: Distinguishing Human from Machine

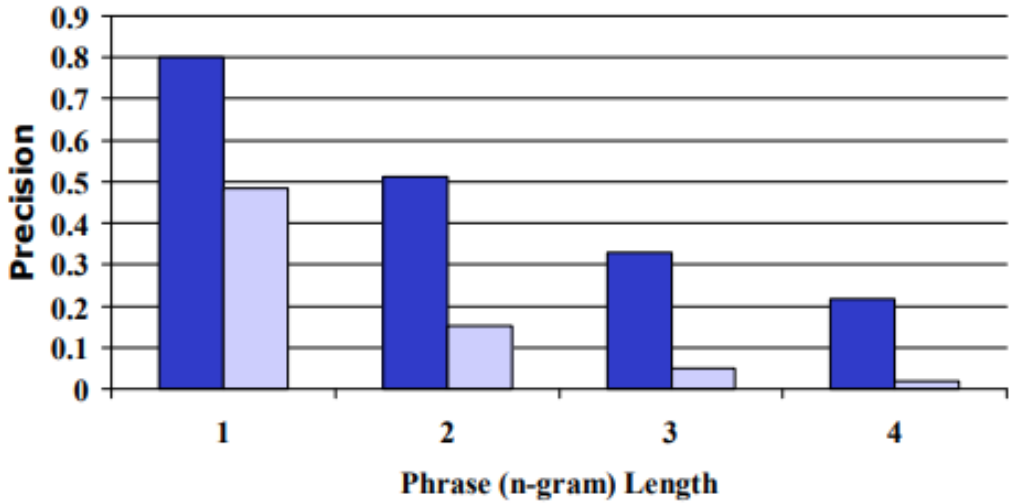
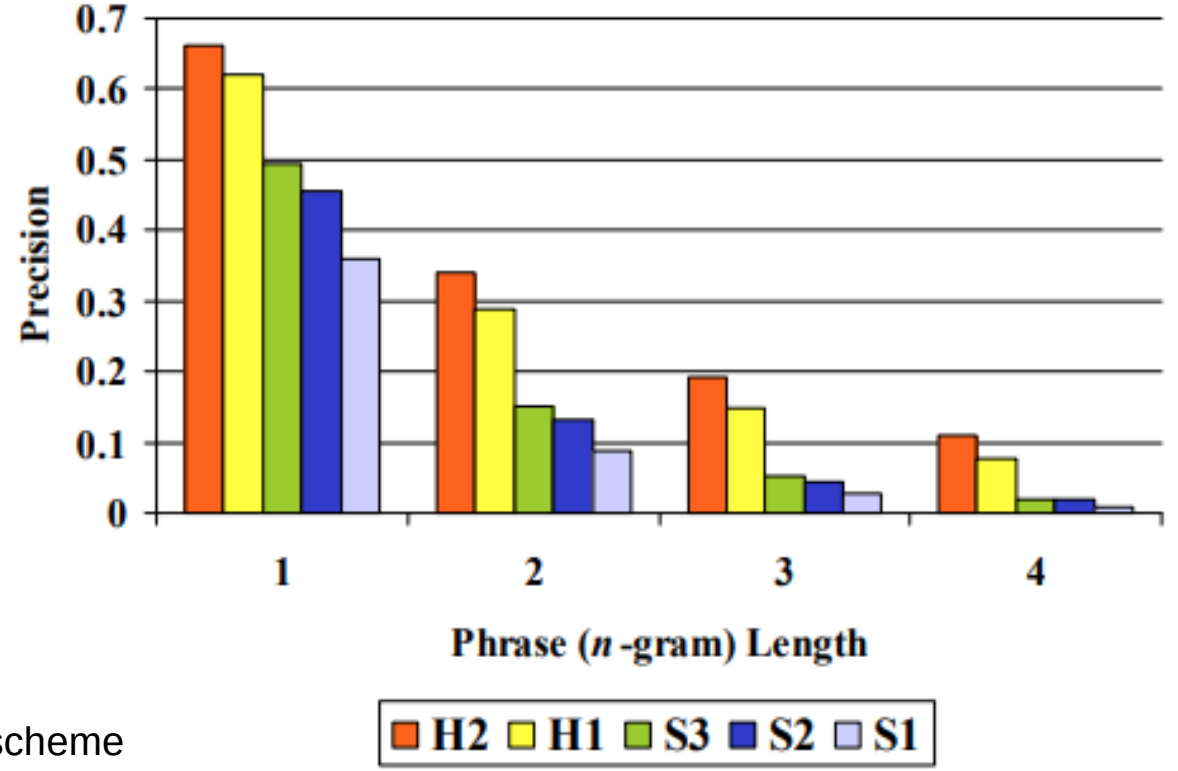


Figure 2: Machine and Human Translations



Problem: Exponential Decay due to Weighted linear averaging scheme

Solution: weighted average of the logarithm of modified precision

Problem with Sentence Length

Candidate: of the

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Unigram precision = 2/2

Bigram precision = 1/1

Problem with Recall

Candidate 1: I always invariably perpetually do.

Candidate 2: I always do.

Reference 1: I always do.

Reference 2: I invariably do.

Reference 3: I perpetually do.

Synonymous Words

Brevity Penalty & BLEU

- To overcome the problem with recall
- Computed on the corpus level

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c = total length of the translation of candidate corpus

r = total length of the reference corpus

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

P_n = geometric average of the modified n-gram precision

W_n = positive weights

N = n-grams upto length N

$$\log \mathrm{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n.$$

In Baseline:

N = 4 and $W_n = 1/N$

-

BLEU Evaluation

- Range: 0-1
- More reference translation = higher BLEU score
- BLEU scores of 5 Systems against two references:

Table 1: BLEU on 500 sentences

S1	S2	S3	H1	H2
0.0527	0.0829	0.0930	0.1934	0.2571

- Is the difference in BLEU metric reliable?
- What is the variance of the BLEU score?
- If we were to pick another random set of 500 sentences, would we still judge S3 to be better than S2?

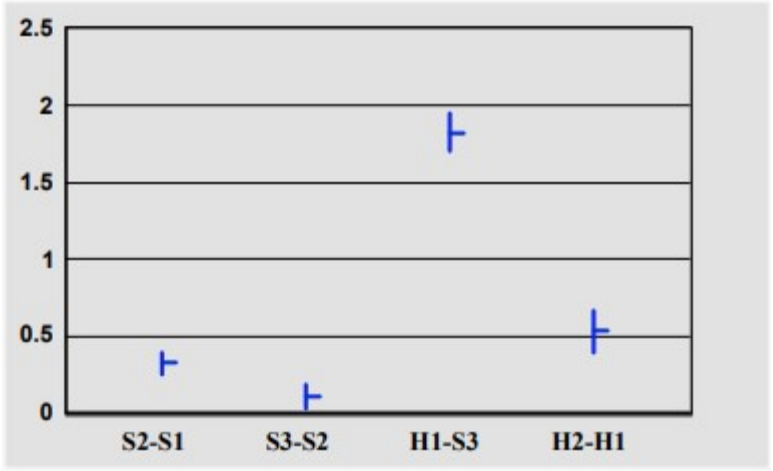
Next set: 20 blocks of 25 sentences

Table 2: Paired t-statistics on 20 blocks

	S1	S2	S3	H1	H2
Mean	0.051	0.081	0.090	0.192	0.256
StdDev	0.017	0.025	0.020	0.030	0.039
t	—	6	3.4	24	11

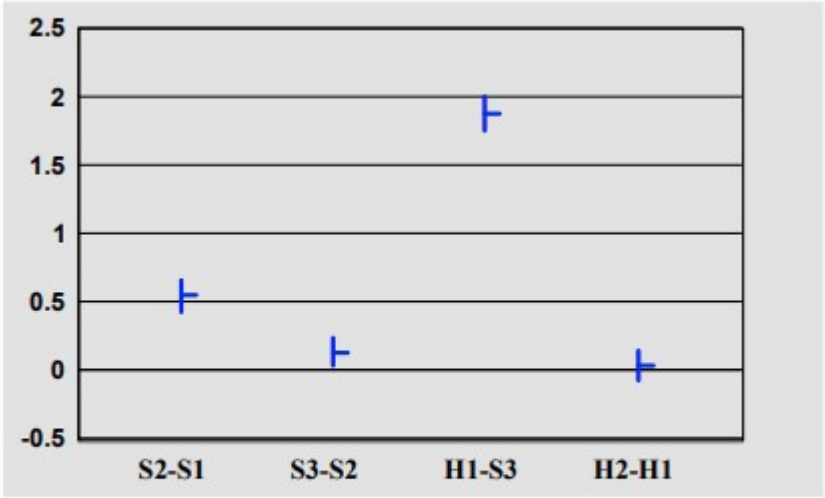
Human Evaluation

Figure 3: Monolingual Judgments - pairwise differential comparison



+95%	0.400	0.194	1.945	0.670
-95%	0.252	0.034	1.705	0.400
monolingual	0.326	0.114	1.825	0.535

Figure 4: Bilingual Judgments - pairwise differential comparison



+95%	0.667	0.238	2.007	0.145
-95%	0.435	0.042	1.759	-0.069
bilingual	0.551	0.140	1.883	0.038

Evaluation – Human vs BLEU

Figure 5: BLEU predicts Monolingual Judgments

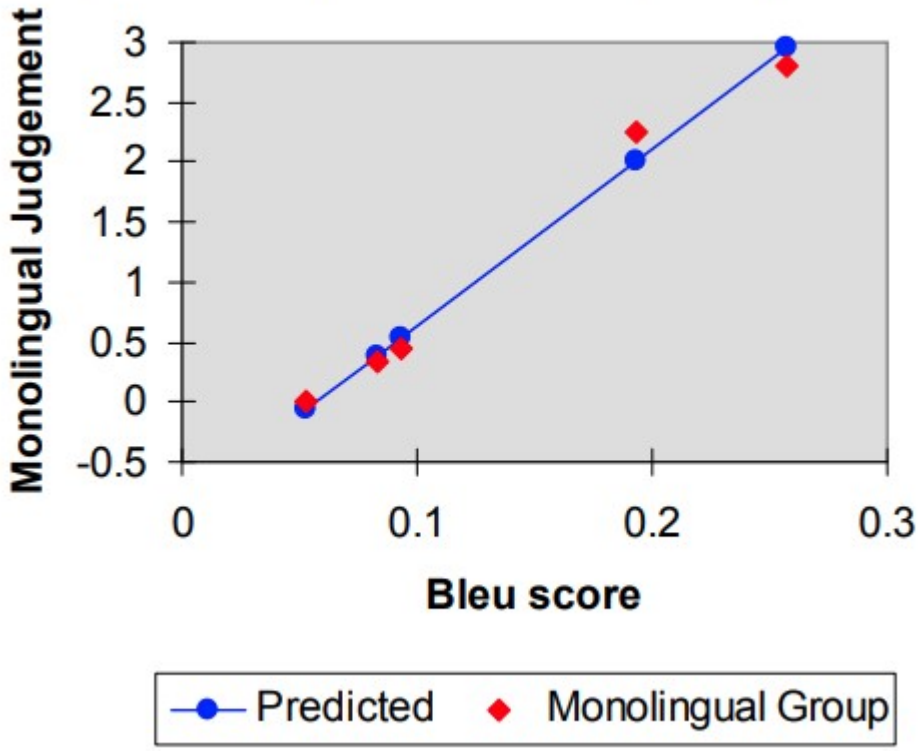
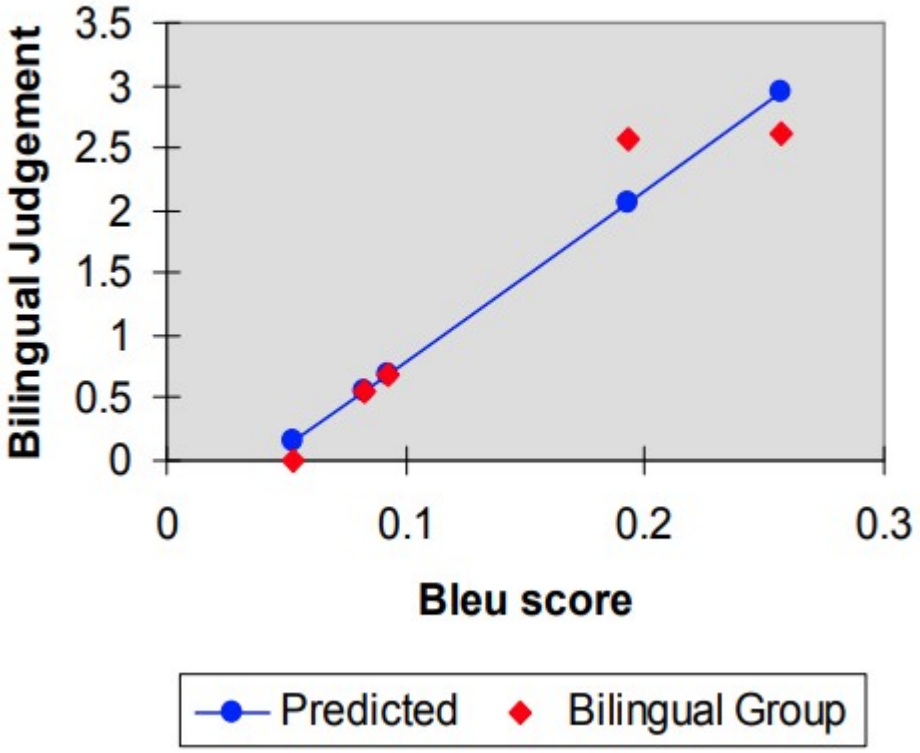


Figure 6: BLEU predicts Bilingual Judgments



- Figures shows linear regression of the group scores as a function of the BLEU score
- Correlation coefficient for monolingual group = 0.99
- Correlation coefficient for bilingual group = 0.96

Conclusion

- Deep LSTM with Limited Vocabulary (input vocabulary of 160,000 and an output vocabulary of 80,000) can outperform the SMT-based system whose vocabulary is unlimited.
- Reversing the input sequence yields better results.
- Model is able to translate very long sentences correctly.

- BLEU could be adopted as a metric for evaluating summarization or similar NLG tasks