Indian Institute of Technology Kharagpur
Department of Computer Science & Engineering

CS60075
Natural Language Processing
Autumn 2020

Module 7:

Machine Translation 5

Neural Machine Translation

28 October 2020

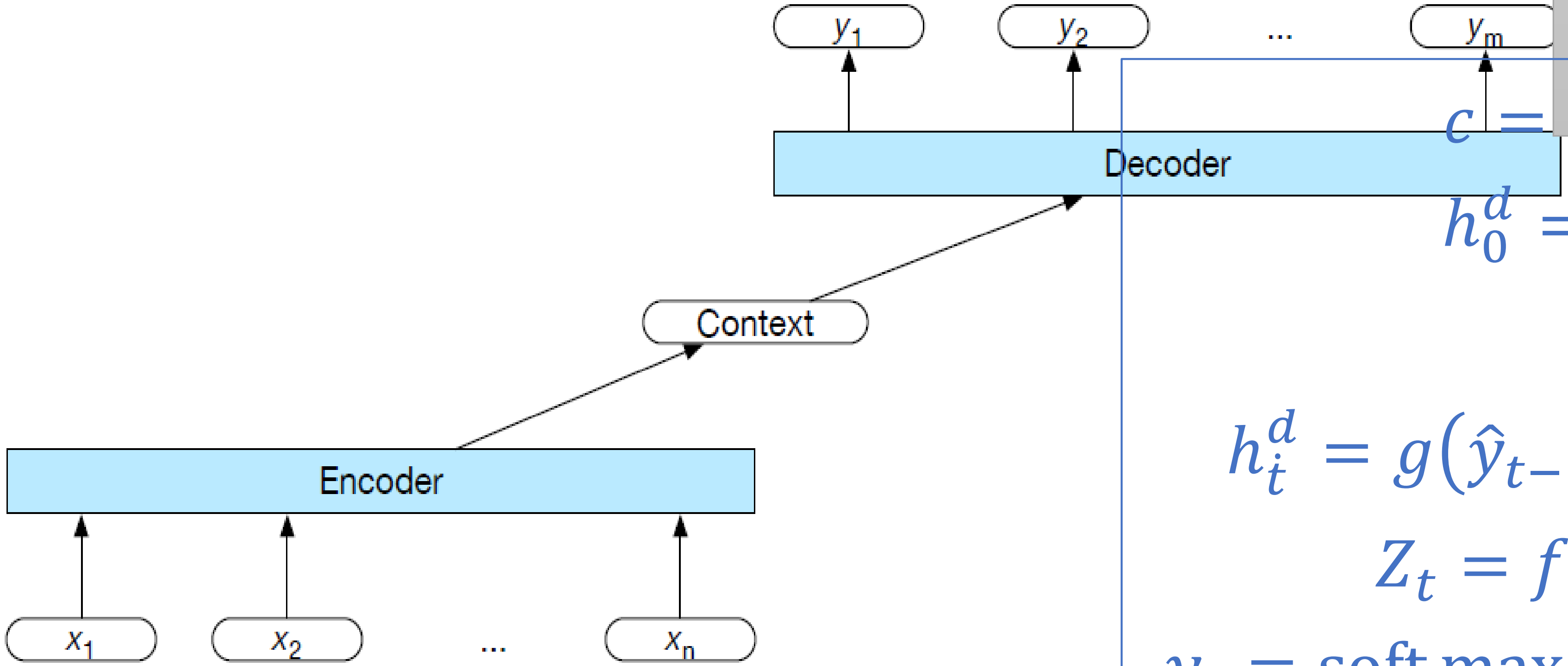
Conditional Language Modeling for Machine Translation

$$P(Y|X) = \prod_{j=1}^{J} P(y_j \mid X, y_1, \ldots, y_{j-1})$$

Encoder-decoder networks

$$c = h_n^e$$
$$h_0^d = c$$

$$h_t^d = g(\hat{y}_{t-1}, h_{t-1}^d)$$
$$Z_t = f(h_t^d)$$
$$y_t = \text{soft max } (z_t)$$

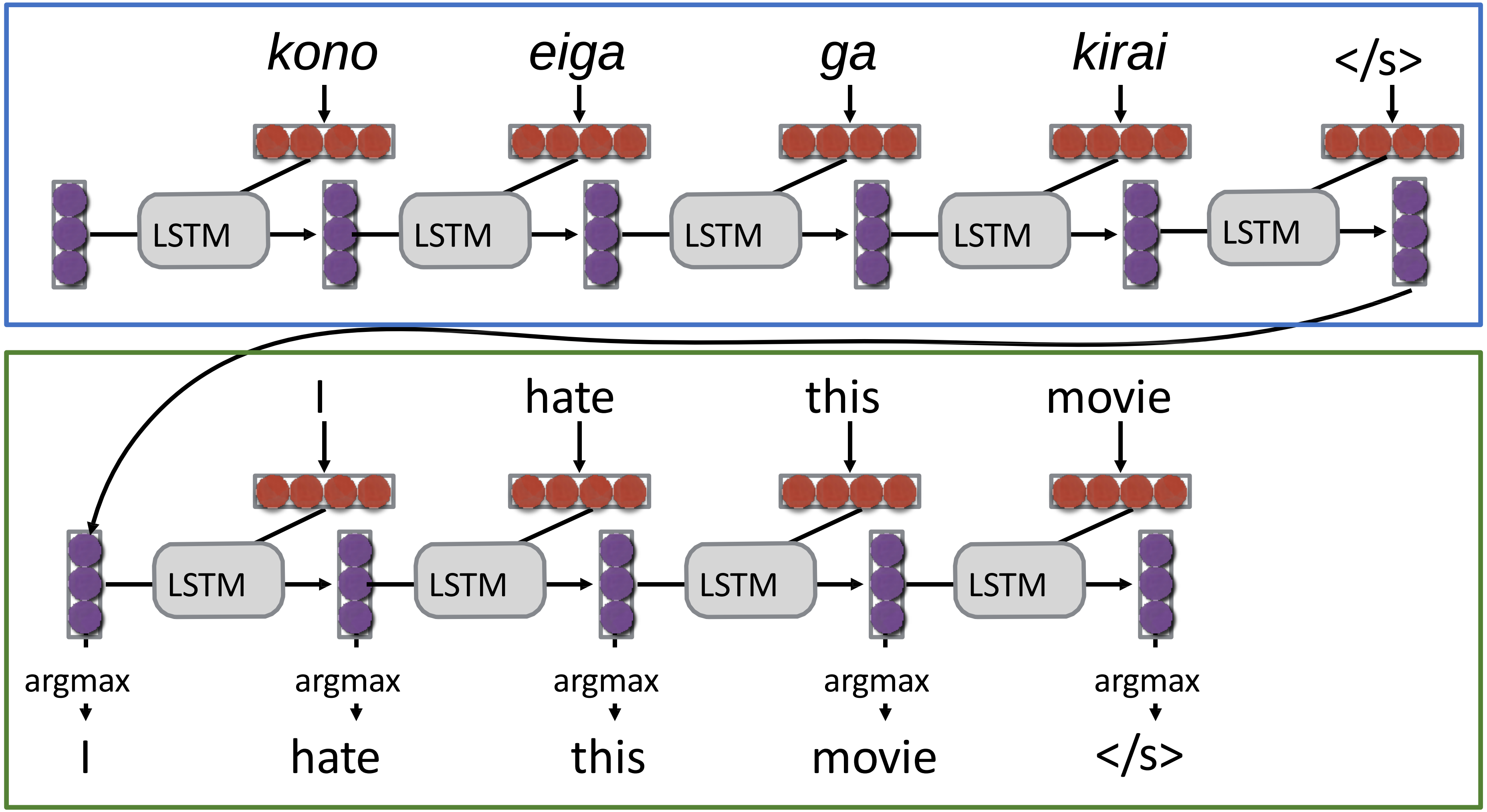


$$c = h_n^e$$
$$h_0^d = c$$

$$h_t^d = g(\hat{y}_{t-1}, h_{t-1}^d, c)$$
$$Z_t = f(h_t^d)$$
$$y_t = \text{soft max } (\hat{y}_{t-1}, z_t, c)$$

Conditional LM for MT

Encoder

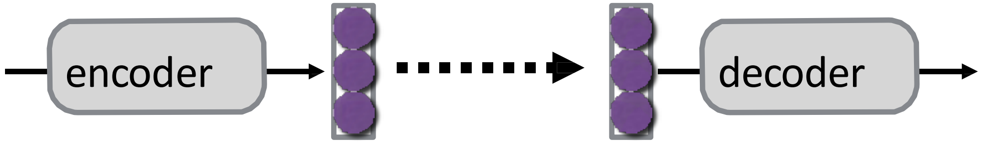


Decoder

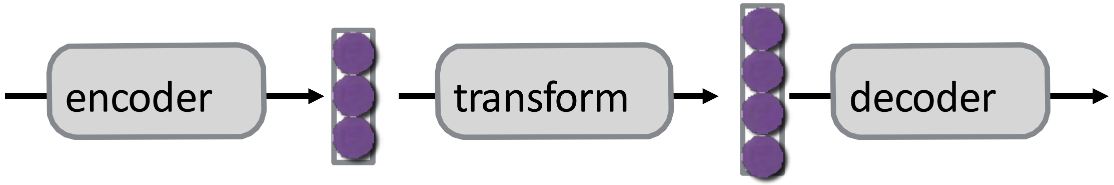
Sutskever et al. 2014

How to pass the hidden state

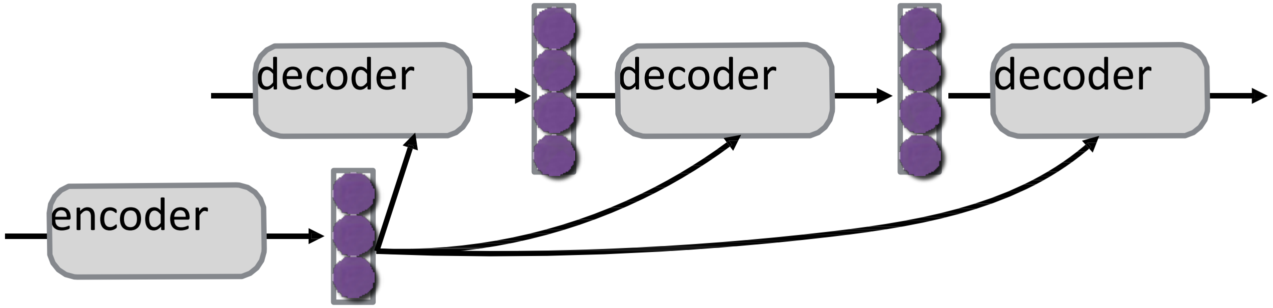
1. Initialize decoder w/ encoder (Sutskever et al. 2014)



2. Transform (can be different dimensions)



3. Input at every time step (Kalchbrenner & Blunsom 2013)



Training Conditional LMs

- Get parallel corpus of inputs and outputs

Maximize likelihood

Standard corpora for MT:

- WMT Conference on Machine Translation runs an evaluation every year with large-scale (e.g. 10M sentence) datasets
- Smaller datasets, e.g. 200k sentence TED talks from IWSLT, can be more conducive to experimentation

The Generation Problem

- We have a model of P(Y|X), how do we use it to generate a sentence?

- Two methods:
  1. Sampling: Try to generate a random sentence according to the probability distribution.
  2. Argmax: Try to generate the sentence with the highest probability.

Ancestral Sampling

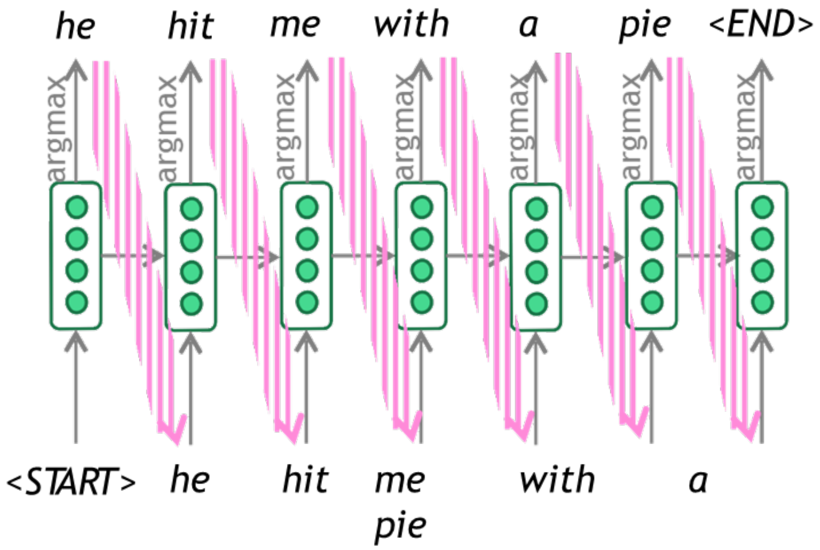
- Randomly generate words one-by-one.

```
while $y_{j-1}$ != “</s>”:
  $y_j \sim P(y_j \mid X, y_1, \ldots, y_{j-1})$
```

Greedy Search

- One by one, pick the single highest-probability word

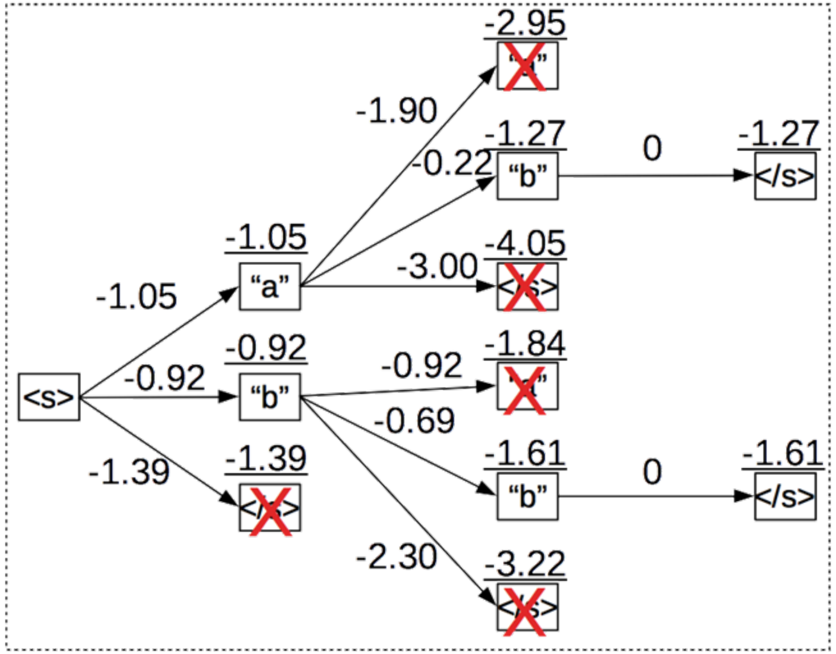
```
while $y_{j-1}$ != “</s>”:
  $y_j = \text{argmax}\ P(y_j \mid X, y_1, \ldots, y_{j-1})$
```



1. Will often generate the “easy” words first
2. Will prefer multiple common words to one rare word

Beam Search

- Instead of picking one high-probability word, maintain several paths



Evaluation

Evaluating MT Quality

- Why Evaluate?
  1. Want to rank systems
  2. Want to evaluate incremental changes
  3. What to make scientific claims

- How not to do it?

Evaluating MT Quality

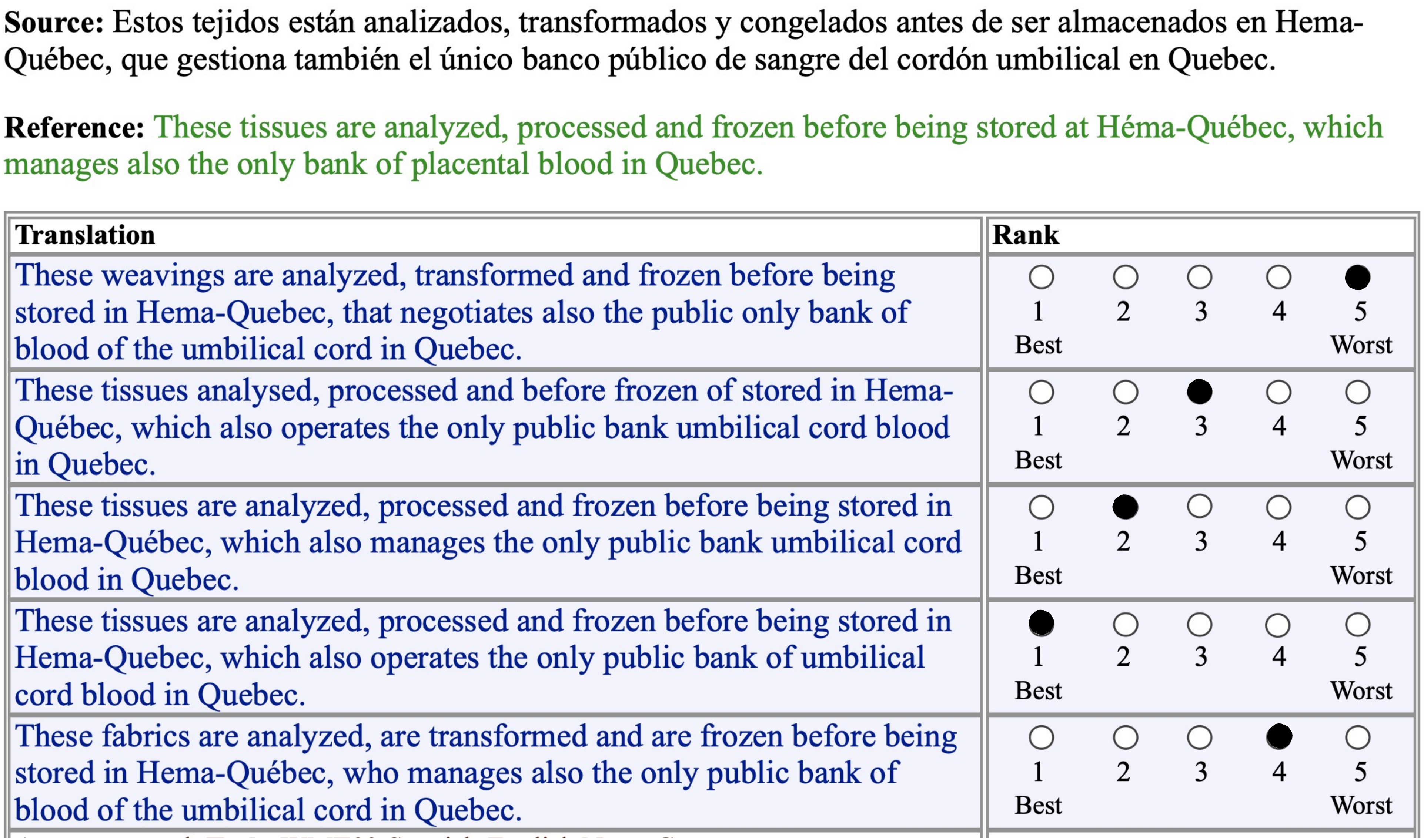
- Why Evaluate?
  1. Want to rank systems
  2. Want to evaluate incremental changes
  3. What to make scientific claims

- How not to do it?
  - Back-translation

Human Evaluation of MT vs Automatic Evaluation

- Human Evaluation is
    - Ultimately what we're interested in, but
    - Very time consuming
    - Not re-usable
- Automatic evaluation is
    - Cheap and reusable, but
    - Not necessarily reliable

Manual Evaluation



Source: Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

Reference: These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

Translation	Rank
These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ○ 4 ● 5 Worst
These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.	○ 1 Best ○ 2 ● 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.	○ 1 Best ● 2 ○ 3 ○ 4 ○ 5 Worst
These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.	● 1 Best ○ 2 ○ 3 ○ 4 ○ 5 Worst
These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.	○ 1 Best ○ 2 ○ 3 ● 4 ○ 5 Worst

Goals for Automatic Evaluation

- No cost evaluation for incremental changes
- Ability to rank systems
- Ability to identify which sentences we're doing poorly on, and categorize errors
- Correlation with human judgments
- Interpretability of the score

Methodology

- Comparison against reference translations
- Intuition: closer we get to human translations, the better we're doin
- Could use WER like in speech recognition?

How to evaluate?

1. Compare against lots of test sentences

2. Use multiple reference translations for each test sentence

3. Look for phrase / n-gram matches, allow movement

BLEU

Bi**L**ingual **E**valuation **U**nderstudy

- Uses multiple reference translations
- Look for n-grams that occur anywhere in the sentence

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

N-gram precision

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

- BLEU modifies this precision to eliminate repetitions that occur across sentences.

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him **to Miami**, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him **to Miami**, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him **to Miami** in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him in Florida. **to Miami**

- Multiple references

“to Miami” can only be counted as correct once

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.

American, Florida, Miami, Orejuela, **appeared**, as, being, calm, carry, escorted, he, him, in, led, **plane**, quite, seemed, take, that, the, **to**, **to**, to, **was** , was, **which**, while, **will**, would, ,, **.**

1-gram precision = 15/18

Hyp	**appeared calm** when **he was** taken **to the American plane** , **which will to Miami , Florida .**

American plane, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, , Florida

2-gram precision = 10/17

Hyp	**appeared calm** when **he was** taken **to the American plane** , **which will to Miami , Florida .**

N-gram precision

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.

1-gram precision = 15/18 = .83

2-gram precision = 10/17 = .59

3-gram precision = 5/16 = .31

4-gram precision = 3/15 = .20

- Geometric average

(0.83 * 0.59 * 0.31 * 0.2)^(1/4) = 0.417

or equivalently

exp(ln .83 + ln .59 + ln .31 + ln .2/4) = 0.417

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Ref 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Ref 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Ref 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Hyp	to the American plane

Is this better?

Hyp	to the American plane

1-gram precision = 4/4 = 1.0

2-gram precision = 3/3 = 1.0

3-gram precision = 2/2 = 1.0

4-gram precision = 1/1 = 1.0

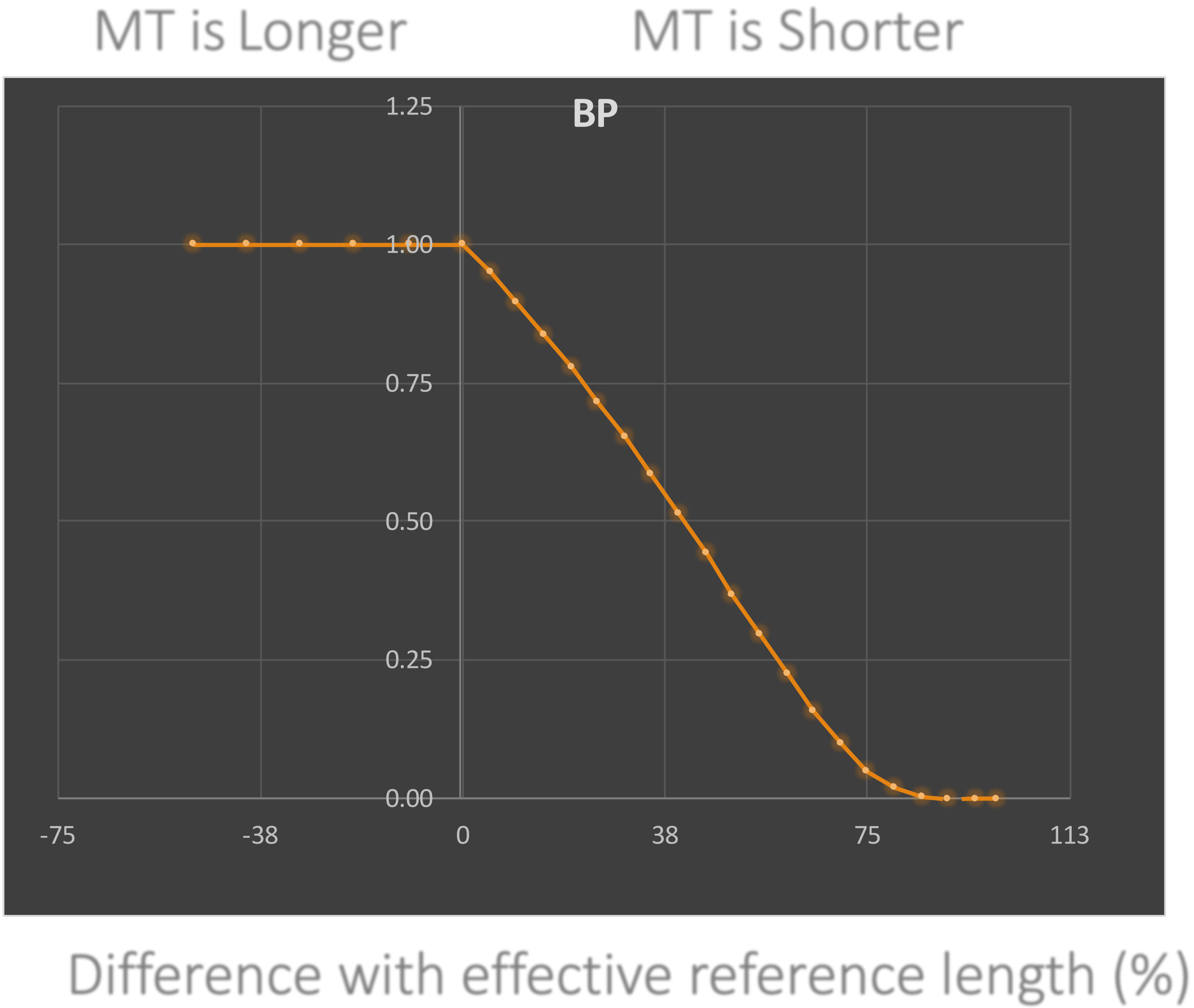
$$\exp(\ln 1 + \ln 1 + \ln 1 + \ln 1) = 1$$

Brevity Penalty

- c is the length of the corpus of hypothesis translations
- r is the effective reference corpus length
- The effective reference corpus length is the sum of the single reference translation from each set that is closest to the hypothesis translation.

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

Brevity Penalty



Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. $r = 20$
Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida. $c = 18$

$$BP = \exp(1-(20/18)) = 0.89$$

Ref 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida. $r = 20$
Hyp	to the American plane $c = 4$

$$BP = \exp(1-(20/4)) = 0.02$$

BLEU

- Geometric average of the n- gram precisions
- Optionally weight them with w
- Multiplied by the brevity penalty

$$\text{Bleu} = \text{BP} \ * \exp(\sum_{n=1}^{N} w_n \log p_n)$$

BLEU

Hyp	appeared calm when he was taken to the American plane, which will to Miami, Florida.

$$\exp(1-(20/18)) * \exp((\ln .83 + \ln .59 + \ln .31 + \ln .2)/4) = 0.374$$

Hyp	to the American plane

$$\exp(1-(20/4)) * \exp((\ln 1 + \ln 1 + \ln 1 + \ln 1)/4) = 0.018$$

Problems with BLEU

- **Synonyms and paraphrases** are only handled if they are in the set of multiple reference translations
- The scores for **words are equally weighted** so missing out on content-bearing material brings no additional penalty.
- The brevity penalty is a stop-gap measure to compensate for the fairly serious problem of not being able to calculate **recall**.