Machine Translation

Machine Translation

- **The task:** given a string in one language (source) and a target language, generate a translation for it in the target language

- Variants of the task:
  - Scope of the text: sentence/paragraph/document
  - Single output or space of possible outputs

Ambiguity Resolution

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - “John plays the guitar” → ג'ון **מנגן** בגיטרה
  - “John plays soccer” → ג'ון **משחק** בכדורגל

- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - “The spirit is willing but the flesh is weak.” ⇒
    “The liquor is good but the meat is spoiled”
  - “Out of sight, out of mind.” ⇒ “Invisible idiot”

Translation Quality

- Achieving literary quality translation is very difficult
- Existing MT systems can generate rough translations that frequently at least convey the **gist** of a document
- High quality translations possible when specialized to narrow domains
- Some MT systems used in ***computer-aided translation*** in which a bilingual human post-edits the output to produce more readable accurate translations
- Frequently used to aid ***localization*** of software interfaces and documentation to adapt them to other languages

Linguistic Differences between Languages

- Syntactic variation between ***SVO*** (e.g. English), ***SOV*** (e.g. Hindi), and ***VSO*** (e.g. Arabic) languages
  - SVO languages tend to use prepositions
  - SOV languages tend to use postpositions

- ***Pro-drop*** languages regularly omit subjects that must be inferred when translated to other languages
  - I don't know. Did you like it?

    知らない。気に入った？
    Shiranai. Ki ni itta?
    know-NEGATIVE. like-PAST?
  - הלכתי הביתה I went home

Reminder: “Word” Order Differences

- Compare English with Japanese glosses:

“IBM bought Lotus”

“IBM Lotus bought”

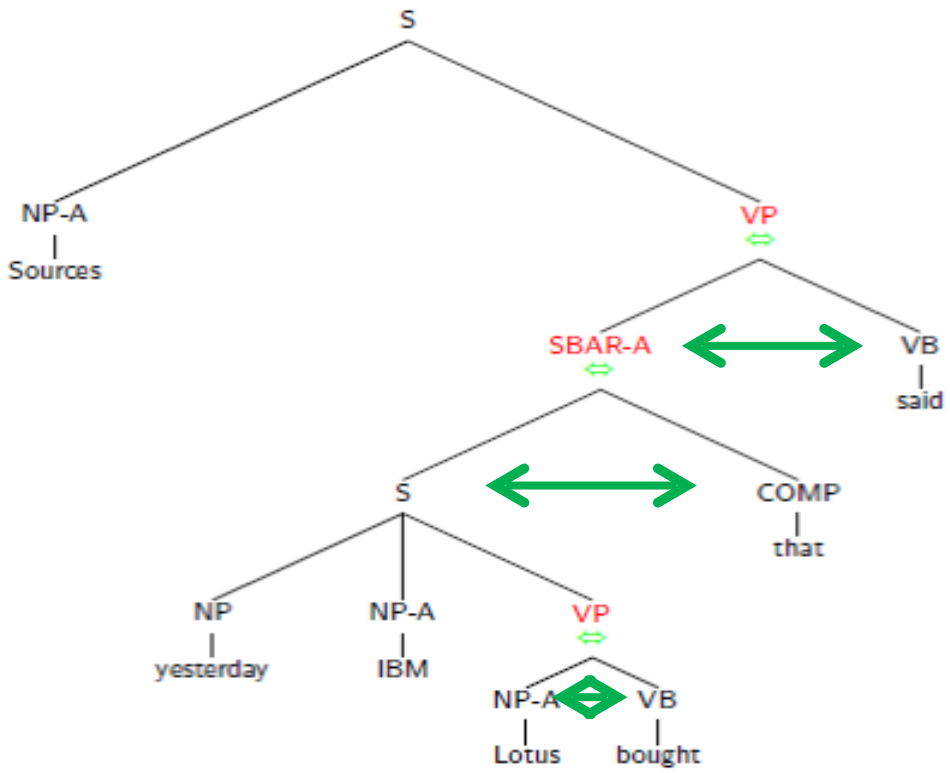
“Sources said that IBM bought Lotus yesterday”

“Sources yesterday IBM Lotus bought that said”

- It is difficult to phrase the possible permutations in terms of words, easier in terms of phrases

Word Order Differences

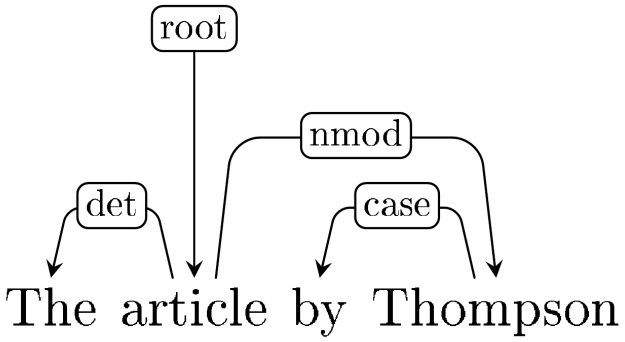
- This can be seen by comparing English and Japanese Constituency Structures:

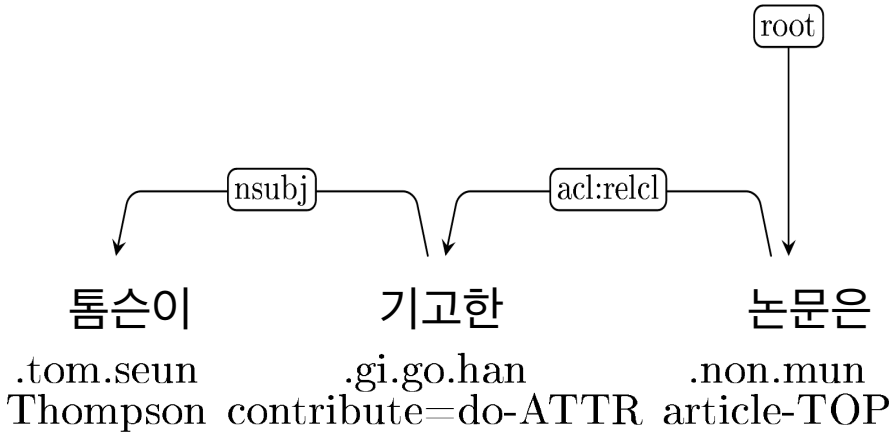


Linguistic Divergences

- Changes between some languages can be much more substantial

- Linguistic divergences (which are not word order differences):
  - “I like him” ⟺ “הוא מוצא חן בעיניי"
  - “to be cold” ⟺ “avoir froid” (lit. “to have cold”)

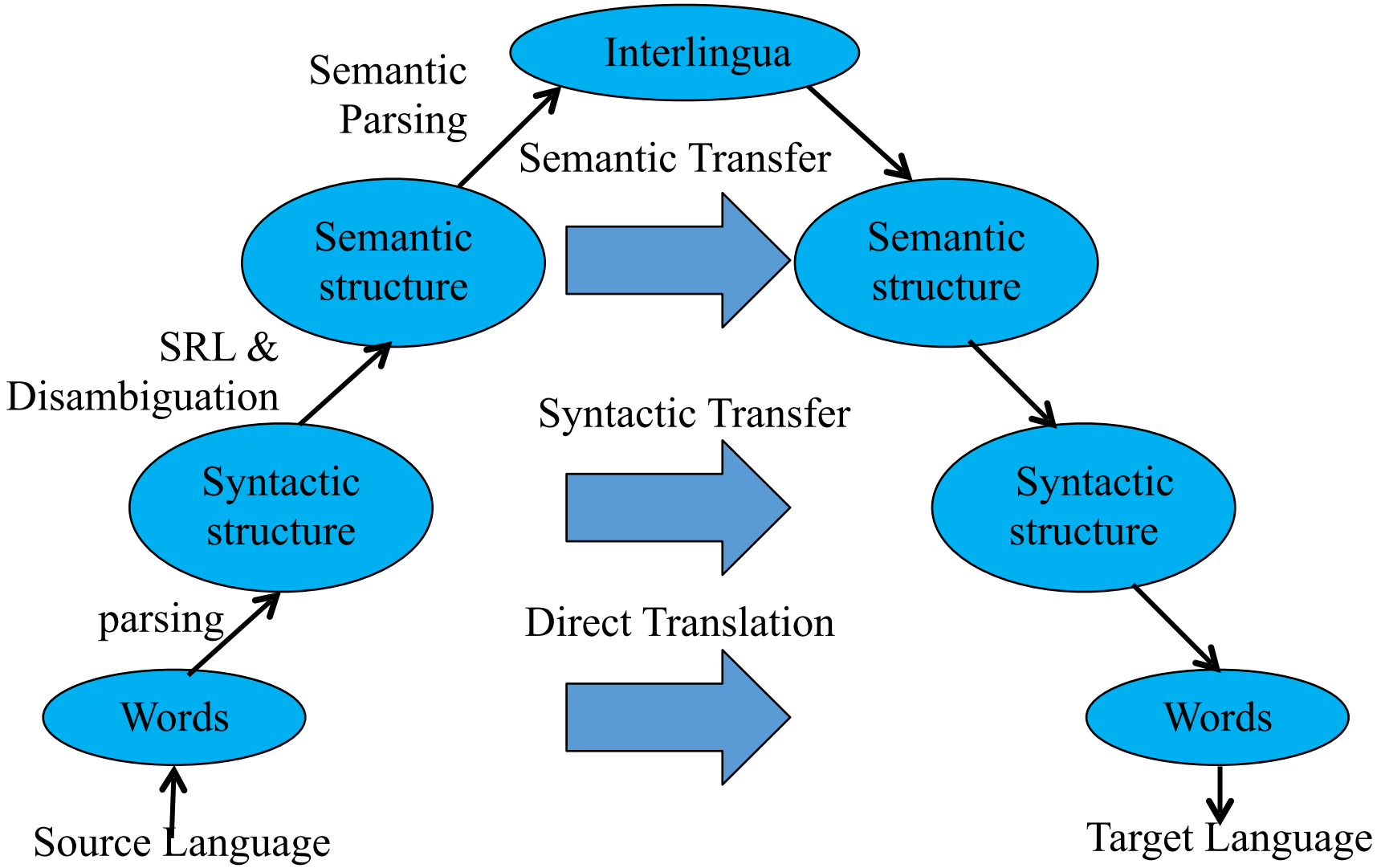




Lexical Gaps

- Some words in one language do not have a corresponding term in the other
  - *Fleuve* (river that flows into ocean) and *Rivière* (river that does not flow into ocean) in French
  - *Schadenfreude* (feeling good about another's pain) in German has no correspondent in English
    - So is *Zechprellerei* (sometimes called 'dine and dash')

Vauquois Triangle



Direct Transfer: Classic Approach

1. Morphological analysis as pre-processing:
   - “Mary didn’t slap the green witch” → “Mary DO:PAST not slap the green witch”
   - "מרי לא סטרה למכשפה הירוקה" →
     “מרי לא סטר:עבר,נקבה ל מכשפה ה ירוק:נקבה"
2. Lexical Transfer
   - Mary DO:PAST not slap the green witch →
     Maria no dar:PAST una bofetada a la verde bruja
3. Lexical Reordering
   - Maria no dar:PAST una bofetada a la bruja verde
4. Morphological generation
   - Maria no dió una bofetada a la bruja verde

Syntactic Transfer: Classic Approach

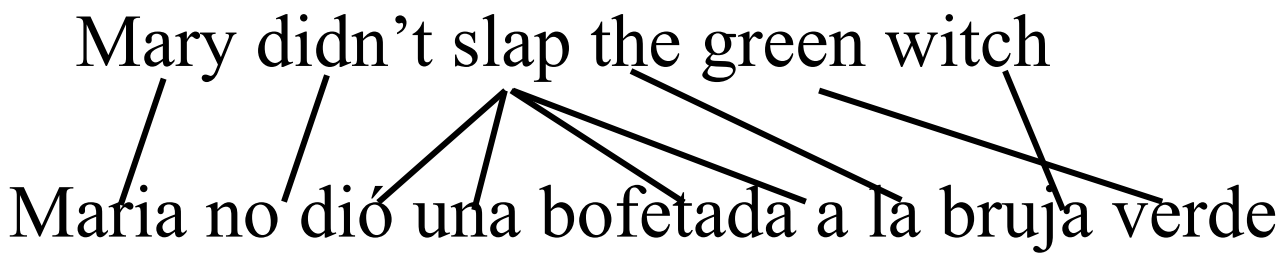
- Simple lexical reordering does not adequately handle more dramatic reordering such as that required to translate from an SVO to an SOV language.
- Need syntactic transfer rules that map parse tree for one language into one for another:
  - English to Spanish:
    - NP → Adj Nom ⇒ NP → Nom ADJ
  - English to Japanese:
    - VP → V NP ⇒ VP → NP V
    - PP → P NP ⇒ PP → NP P

Statistical Methods

- As with other areas of MT, manually crafted rules are low in coverage and don't scale well
- SMT acquires knowledge needed for translation from a ***parallel corpus*** or ***bitext*** that contains the same set of documents in two languages
  - With some models more supervision is required, such as a syntactic parser or a bi-lingual dictionary
- The *Canadian Hansards* (parliamentary proceedings in French and English) and the *Europarl* (protocols of the European Parliament, in all official EU languages) are well-known parallel corpora
  - Parallel corpora are also published through the annual conference for MT (WMT)

Word Alignment

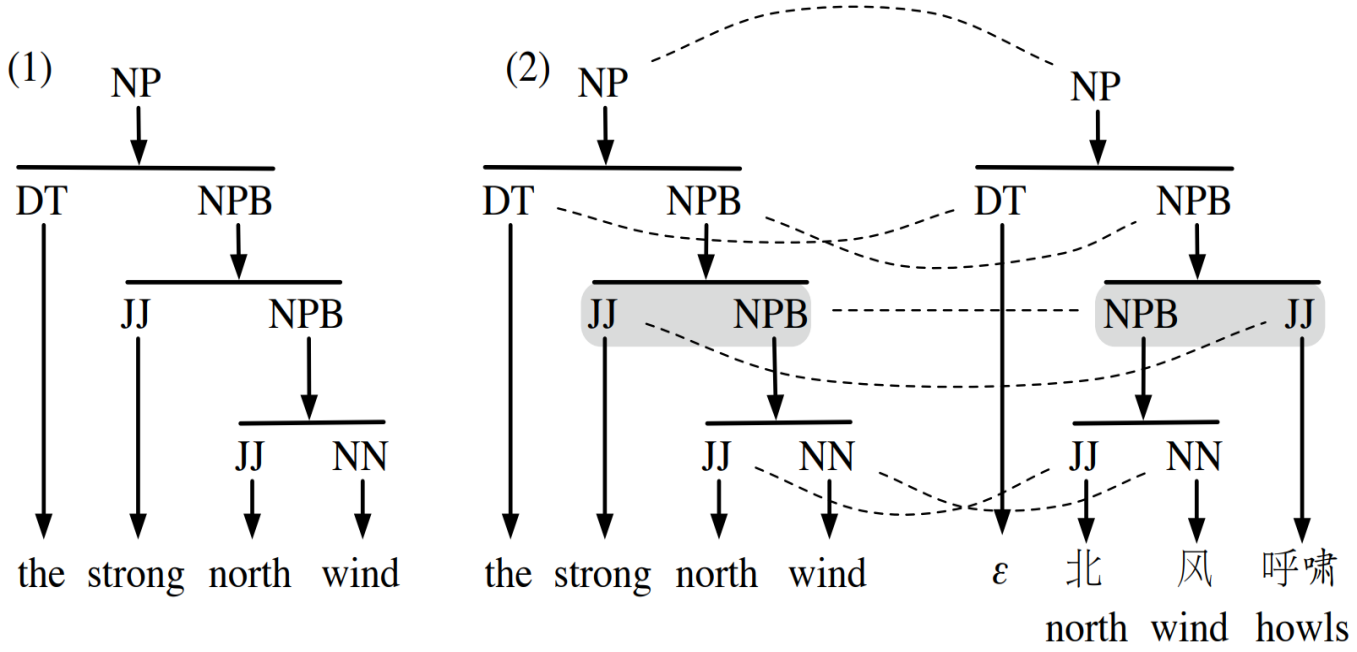
- One of the basic tools in MT
- Shows mapping between words in one language and the other



- Often done by finding co-occurrences across of words/phrases across parallel data

Syntax-Based Statistical Machine Translation

- Syntax-based approaches demonstrated improvement for translating between more distant language pairs, e.g., Chinese/ English
- They assume that the syntactic trees of the derivations in the two languages are similar, except that each right-hand-side of a production may be permuted relative to the production in the other language



Adapted from Lopez (2007) Survey on SMT

Syntax-Based Statistical Machine Translation

- Despite the promise of such models, their assumption is still too weak to deal with all cross-linguistic divergences where the syntactic tree is entirely different

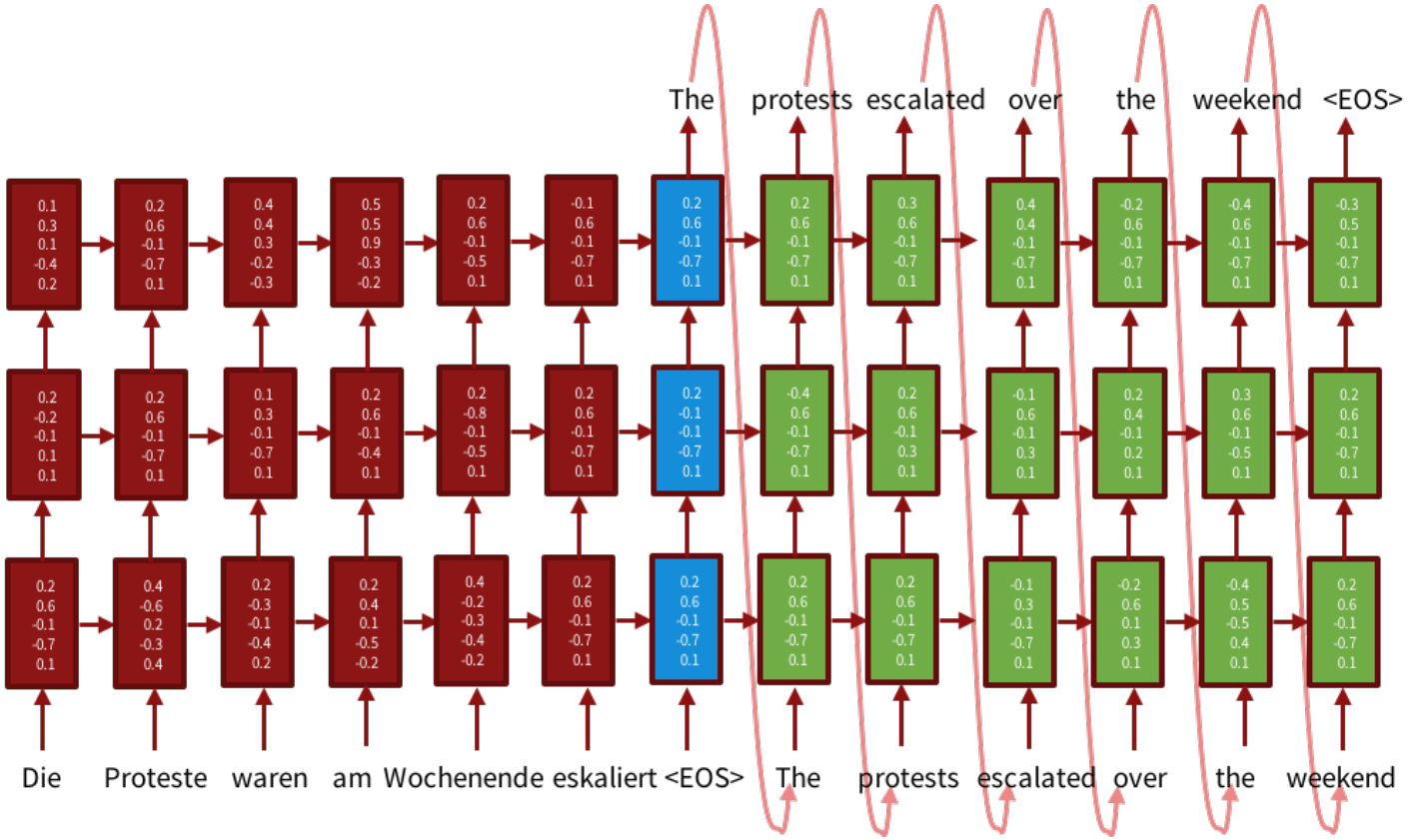
Jacky happened to meet the guy yesterday

ג'קי פגש במקרה את הבחור אתמול


ג'קי קרה לפגוש את הבחור אתמול

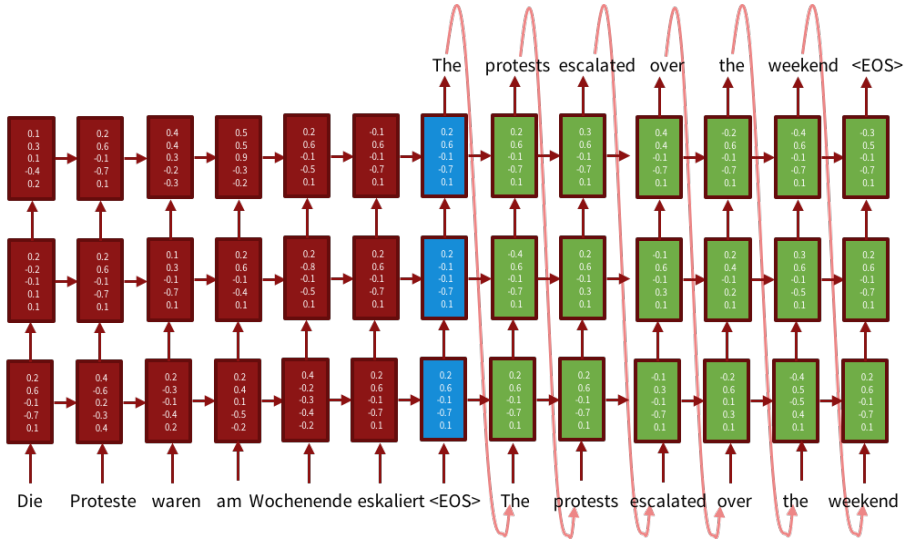
Neural Machine Translation

- The current state of the art outperforms phrase-based translation by a considerable margin in a variety of settings
- The standard architecture is called encoder-decoder



Neural Machine Translation

- Encoder-decoder systems have two stages:
  1. Encoding: the input sentence is encoded into a vector. This can be done by RNNs or other types of networks.
  2. Decoding: like a language model it assigns a probability to the next word. However, it also conditions on the encoded input.
- This is a neural LM, which is conditioned on the encoding of the input



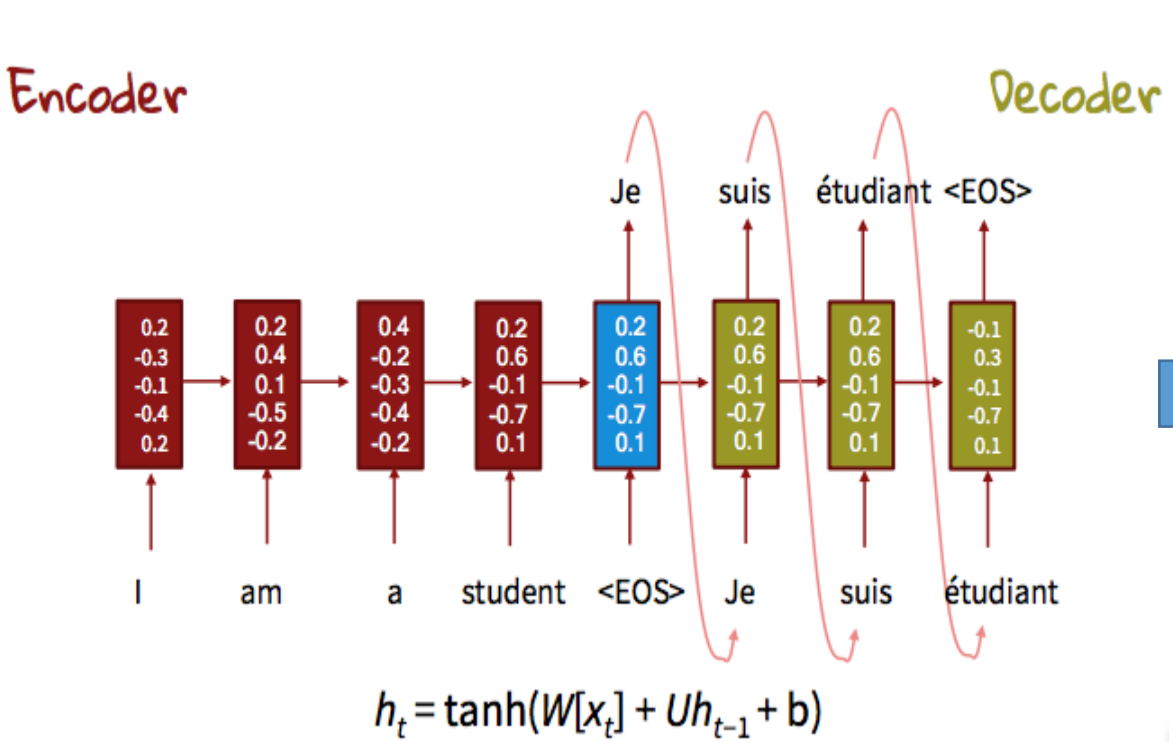
Neural Machine Translation

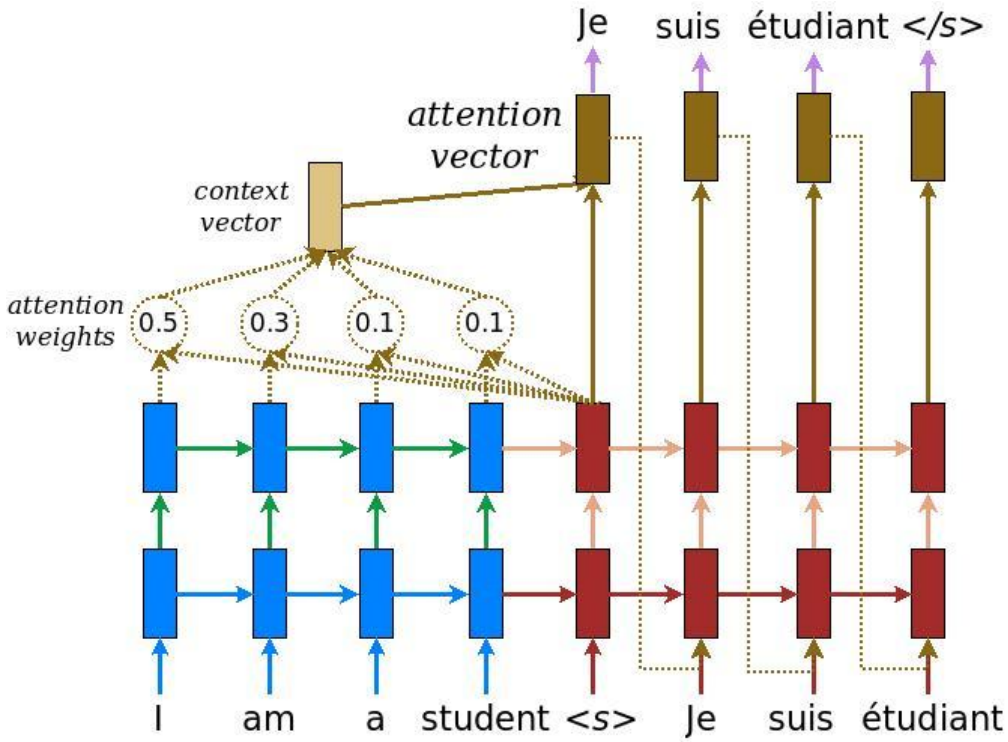
- Training: standard training is done by maximizing the average log probability of the next word, given the input and the prefix (*y* is the translation, *x* is the input sentence)

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} log[p_\theta(y_i | y_{i-1}, \ldots, y_1, \mathbf{x})]$$

- Decoding: similar to what we saw in transition-based parsing
  - Greedy decoding
  - Beam search
  - *Inter alia*

Adding Attention Mechanism

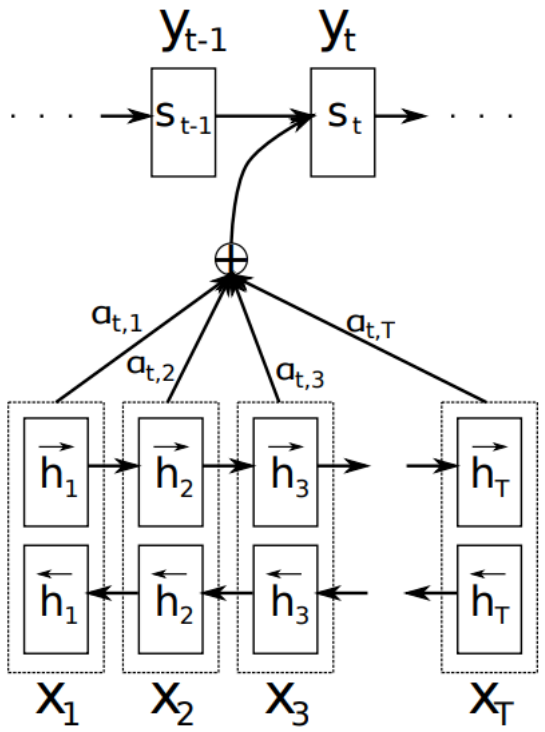




Neural Machine Translation by Jointly Learning to Align and Translate, Bahdanau, Cho and Bengio. ICLR 2015

Adding Attention Mechanism

- One of the problems in the simple encoder-decoder RNN is that the entire sentence is generated from summary vectors of the entire source sentence
- The attention mechanism allows translating only the relevant parts of the sentence
  - Like word alignment, but soft and contextual



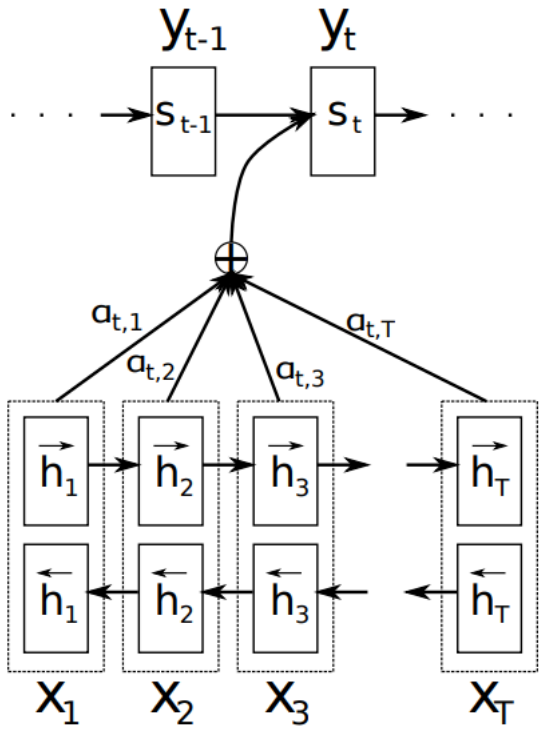
Adding Attention Mechanism

$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1},s_i,c_i)$$

$$s_i = f(s_{i-1},y_{i-1},c_i)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \qquad \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1},h_j)$$



Transformers

- Next class we will learn about recent work that bases the entire encoder-decoder architecture on the mechanism of attention

Machine Translation Evaluation

- Human subjective evaluation is the best but is time-consuming, expensive and inconsistent for long sentences.

- Automated evaluation comparing the output to **multiple human reference translations** is cheaper and generally correlates with human judgements.

The BLEU Score

- The most commonly measure is BLEU (Papineni et al., 2002)
- It determines number of *n*-grams of various sizes that the MT output shares with the reference translations
- Computes a modified precision measure of the *n*-grams in MT result
- Since precision-based measures favor short sentences, a penalty term is introduced for long translations

The BLEU Score

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 1 Unigram Precision: 5/6

BLEU Score

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to the green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 1 Bigram Precision: 1/5

BLEU Score

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to the green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 2 Unigram Precision: 7/10

BLEU Score

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to the green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 2 Bigram Precision: 4/9

Modified *N*-Gram Precision

- If we evaluate performance on a single candidate *c* against a set of references, then for each n-value (n=1,…,N):

$$p_n = \frac{\text{number of n-grams in } c \text{ that appear in at least one of the references}^*}{\text{number of n-grams in } c}$$

- We compute the geometric mean of *n*-gram precision over all *n*-grams up to size *N* (typically 4):

$$p = \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}}$$

* - if an n-gram appears multiple times in *c*, we do not count it more than the maximal number of times it appears in any of the references

Modified *N*-Gram Precision

- BLEU can be easily extended to evaluate many sets of candidates and references (a corpus)
  - We do not cover it here

- Returning to the example from before we get: (for N=2)

$$\text{Cand 1:} \quad p = \sqrt[2]{\frac{5}{6}\frac{1}{5}} = 0.408$$

$$\text{Cand 2:} \quad p = \sqrt[2]{\frac{7}{10}\frac{4}{9}} = 0.558$$

Brevity Penalty

- Not easy to compute recall to complement precision since there are multiple alternative gold-standard references and don't need to match all of them
- Instead, use a penalty for translations that are shorter than the reference translations
- Define effective reference length, *r*, for each sentence as the length of the reference sentence with the largest number of *n*-gram matches. Let *c* be the candidate sentence length:

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

BLEU Score

- Final BLEU Score: BLEU = $BP \times p$

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$$

$$BLEU = 0.846 \times 0.408 = 0.345$$

Cand 2: Mary did not give a smack to a green witch.

Best Ref: Mary did not smack the green witch.

$$c = 10, \quad r = 7, \quad BP = 1$$

$$BLEU = 1 \times 0.558 = 0.558$$

The BLEU Score

- BLEU has been shown to correlate with human evaluation when comparing outputs from different SMT systems
- However, it is does not correlate with human judgments when comparing SMT systems with human translations
- Other MT evaluation metrics have been proposed that claim to overcome some of the limitations of BLEU