Machine Translation

Kwankamol Nongpong, Ph.D.

CS4430 Selected Topic in Natural Language Processing and Social Interaction
CSX4210 / ITX4210 Natural Language Processing and Social Interaction
Vincent Mary School of Science and Technology
Assumption University

Machine Translation

- Translates from one language to another using computers.
- The common uses of machine translation are:
  - for information access
  - to aid human translators (machine generates the draft translation, post edited by human translator) i.e., computer-aided translation (CAT)
  - incremental translation, translate speech on-the-fly before the entire sentence is complete.
- Google Translate

Encoder-Decoder Network

- *aka* sequence to sequence network
- Standard algorithm of MT
- Can be implemented with RNNs or with transformers.
- Used for sequence modeling in which the output sequence is a complex function of the entire input sequence.
  - Not direct mappings from individual words
  - The words of the target language may not agree with the words of the source language (number and order of words).

Example

- English: He wrote a letter to a friend
- Japanese: tomodachi ni tegami-o kaita
friend to letter wrote
- Chinese: ?

Example: Chinese and English

- 大会/General Assembly 在/on 1982年/1982 12月/December 10日/10 通过 了/adopted 第37号/37th 决议/resolution， 核准了/approved 第二次/second 探索/exploration 及/and 和平peaceful 利用/using 外层空间/outer space 会 议/conference 的/of 各项/various 建议/suggestions。

- On 10 December 1982 , the General Assembly adopted resolution 37 in which it endorsed the recommendations of the Second United Nations Conference on the Exploration and Peaceful Uses of Outer Space .

Other Usage of Encoder-Decoder

- Summarization
- Dialogue
- Semantic Parsing

Language Divergence & Typology

- Word Order Typology
- Lexical Divergences
- Morphological Typology
- Referential Density

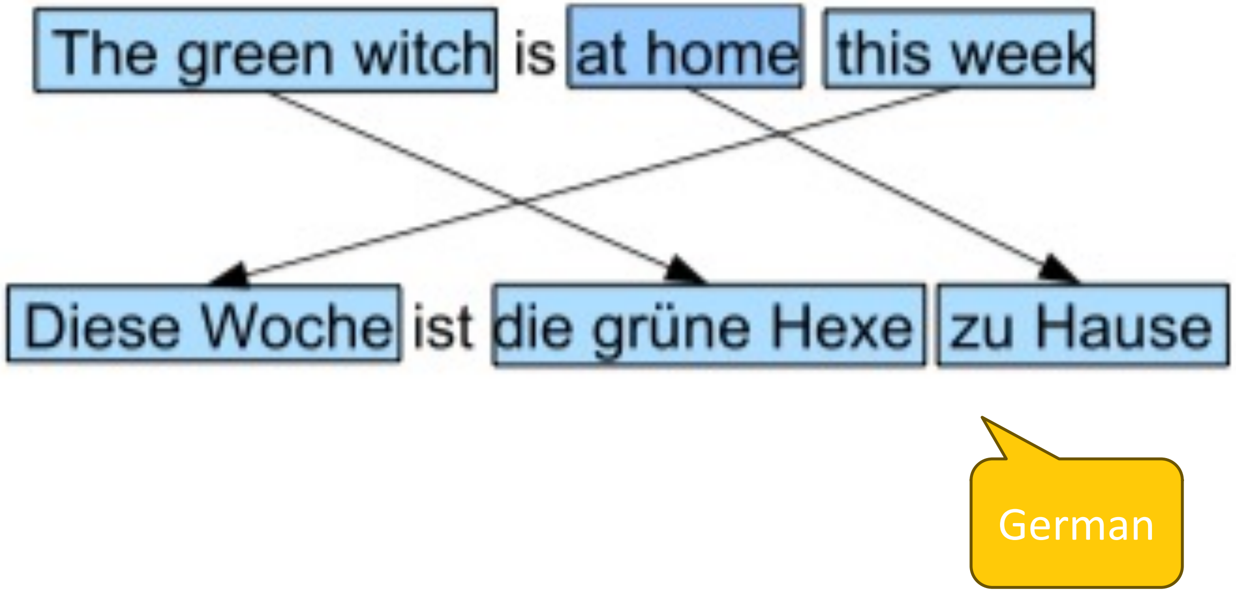
Word Order Typology

- Subject-Verb-Object (SVO)
  - German, French, English, Mandarin
- Subject-Object-Verb (SOV)
  - Hindi, Japanese
- Verb-Subject-Object (VSO)
  - Irish, Arabic

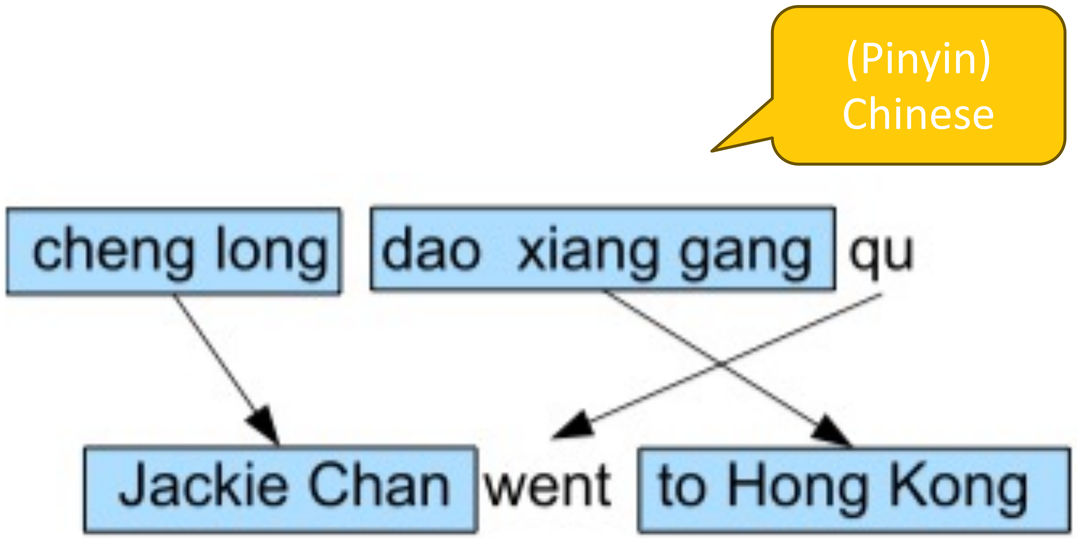
Example

• English:	He wrote a letter to a friend	
• Japanese:	tomodachi ni tegami-o kaita	
	friend to letter wrote	
• Arabic:	katabt risala li sadq	
	wrote letter to friend	

Example: Word Order Difference



Example: Word Order Difference



Lexical Divergences

- Translate individual words from one language to another.
- English uses brother for male sibling.
- Chinese has distinct words for older brother and younger brother.
  - 哥哥 and 弟弟
- Translation requires a kind of specialization, disambibuating the different uses of a word
  - Word Sense Disambiguation
- Moreover, there is also an issue of lexical gap.
  - no word or phrase, can express the exact meaning of a word in the other language.

Morphological Typology

- Languages are characterized by 2 dimensions:
  - Number of morphemes per words
  - Degree to which morphemes are segmentable

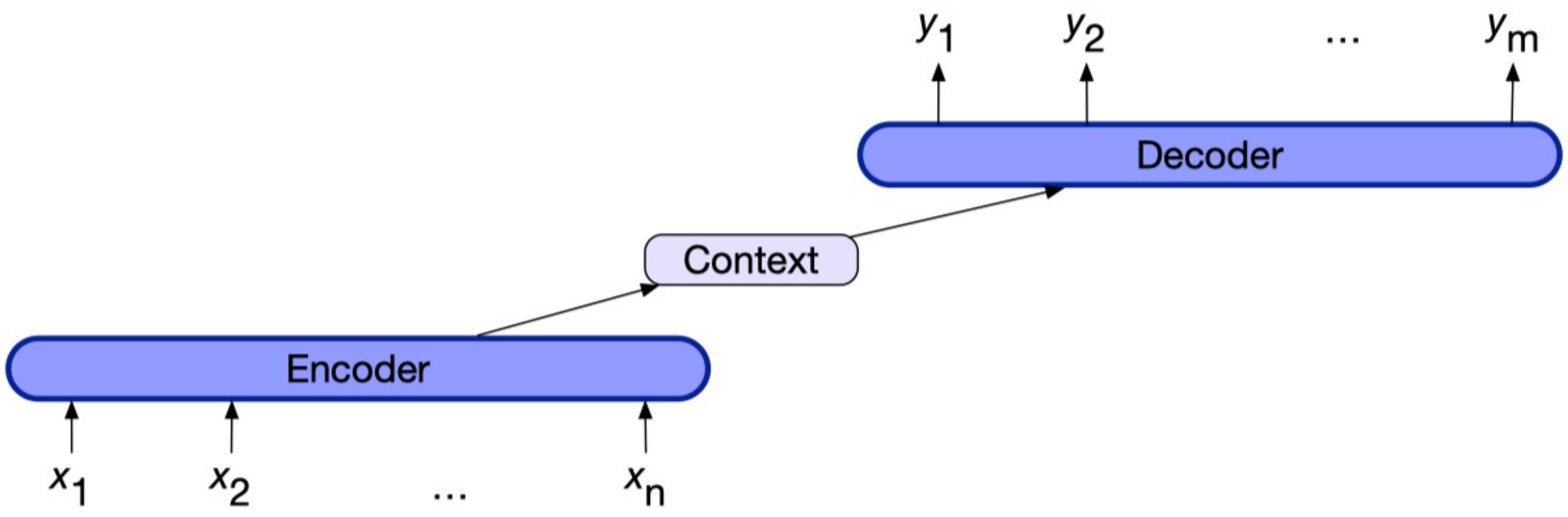
Referential Density

- Languages that can omit pronouns are called pro-drop languages.
- Languages that tend to use more pronouns are more referentially dense.
- Chinese and Japanese are referentially sparse (or cold) languages.
  - The hearer has to do more inference work to recover antecedents.
- Hot vs. Cold media
  - e.g. movie vs. comics
- Translating from languages with pro-drop to non-pro-drop languages can be difficult.
  - Recover who/what is being talked about and insert a proper pronoun.

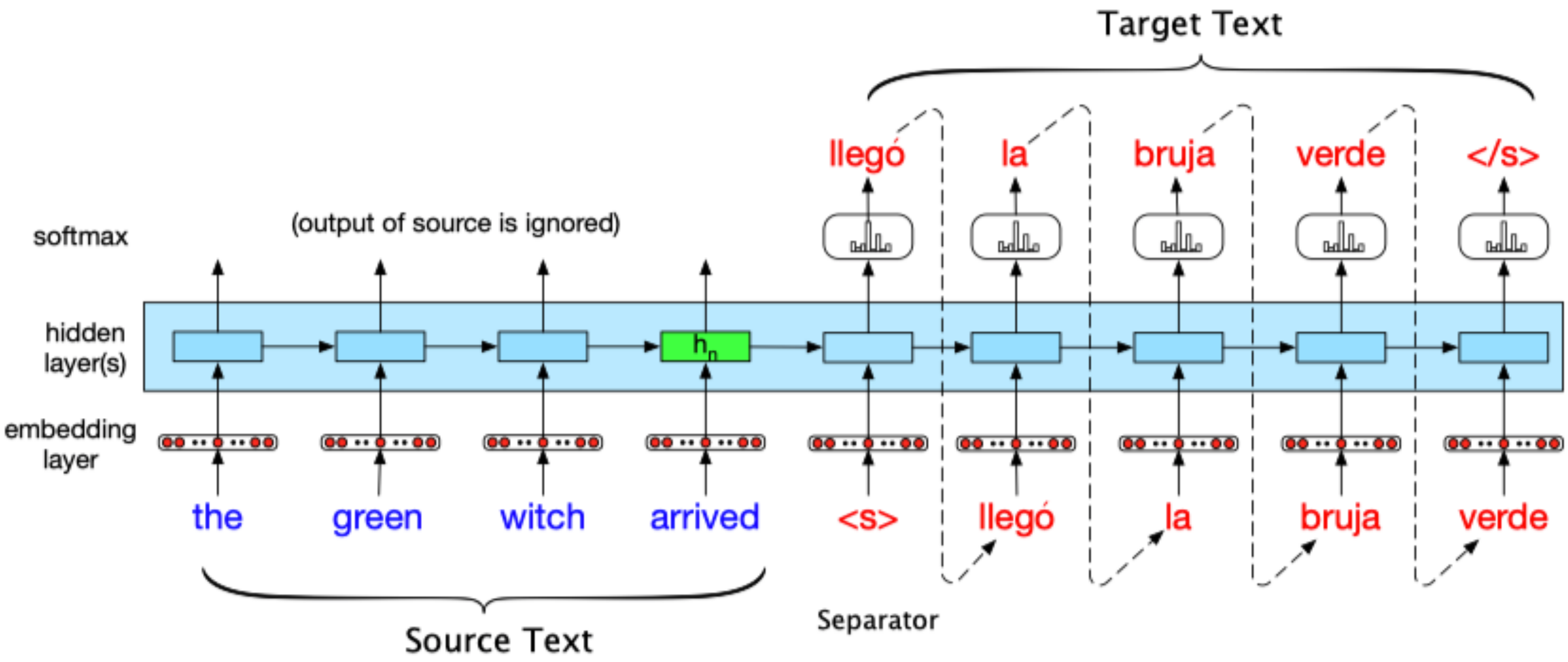
The Encoder-Decoder Model

Encoder-Decoder Model

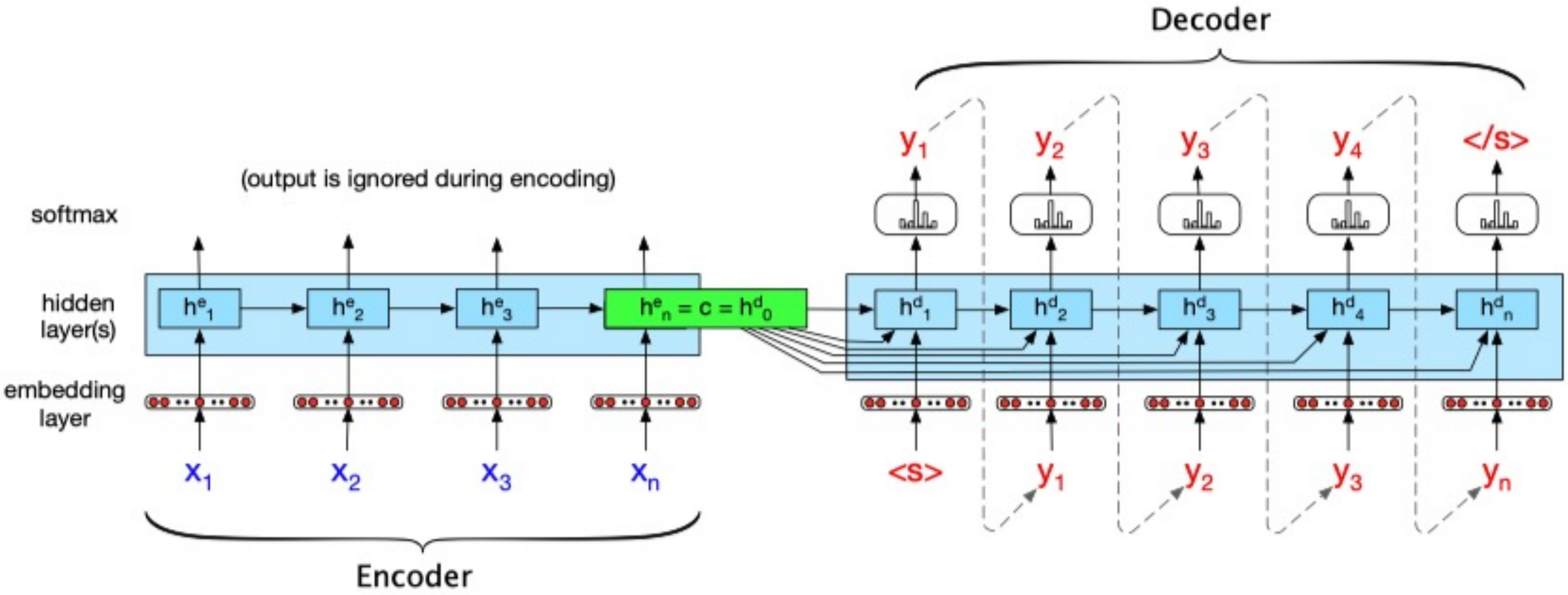
- Encoder network takes an input sequence and creates a contextualized representation *i.e.*, context of the input.
- Decoder generates a task specific output sequence.



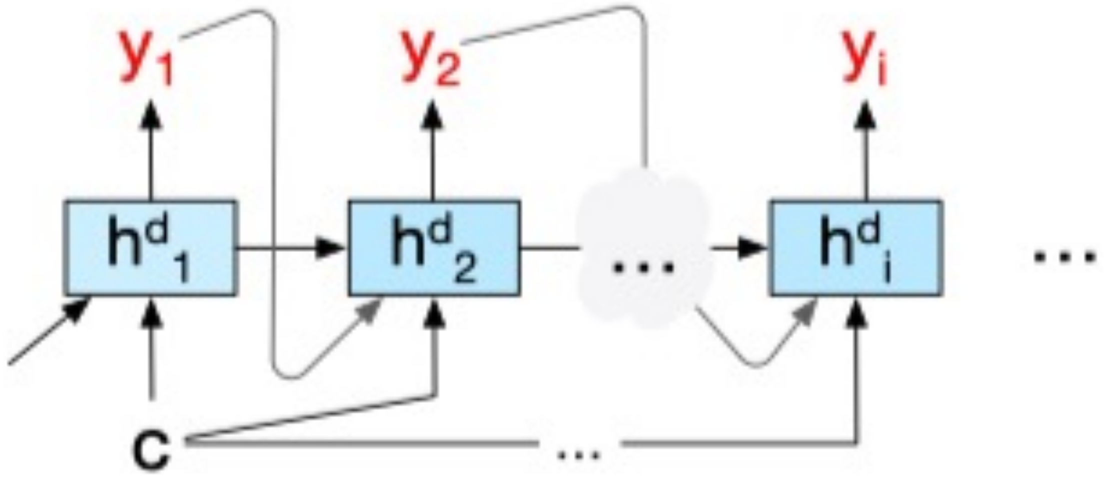
Encoder-Decoder with RNNs



Encoder-Decoder with RNNs (Formal)

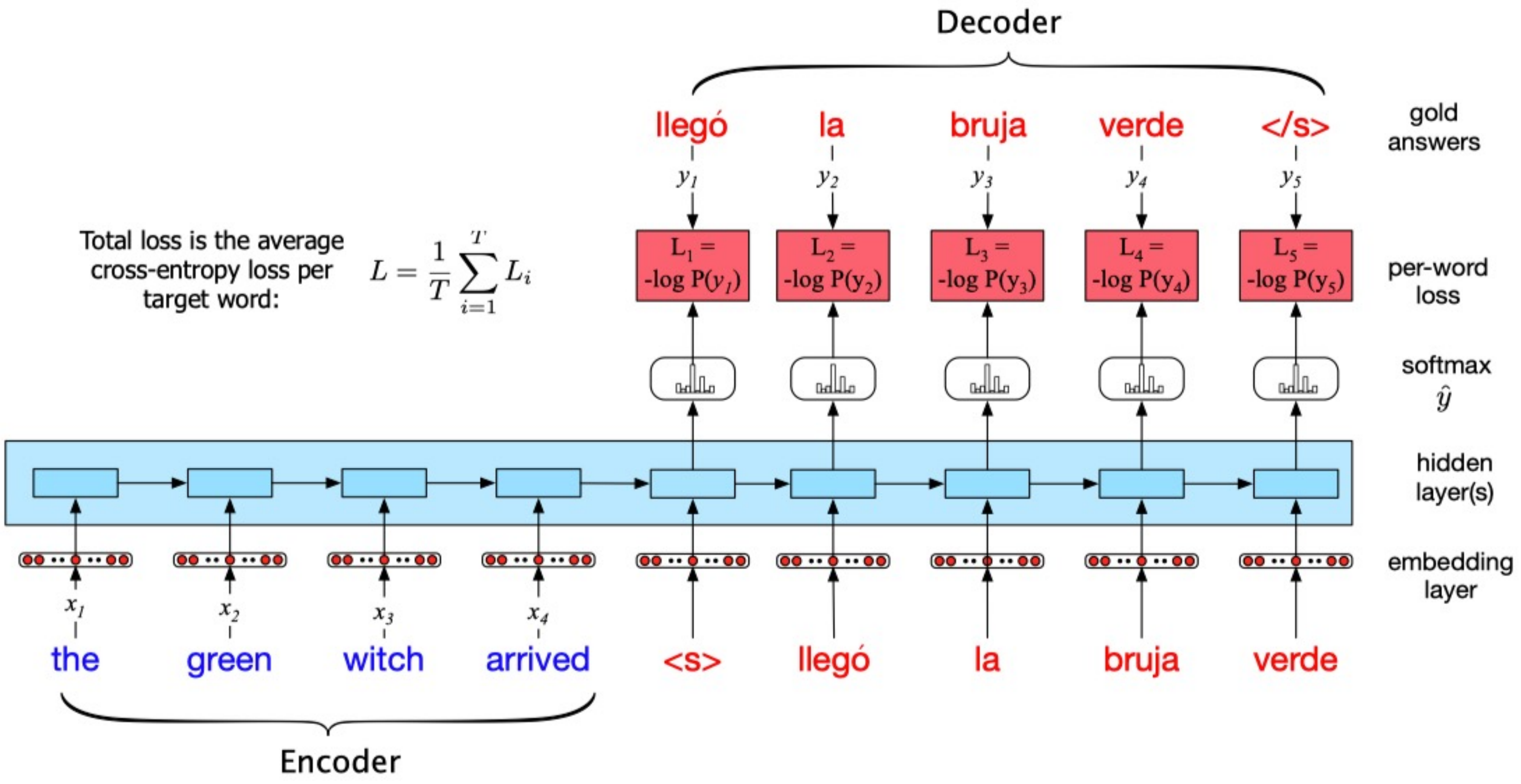


Hidden States and the Context

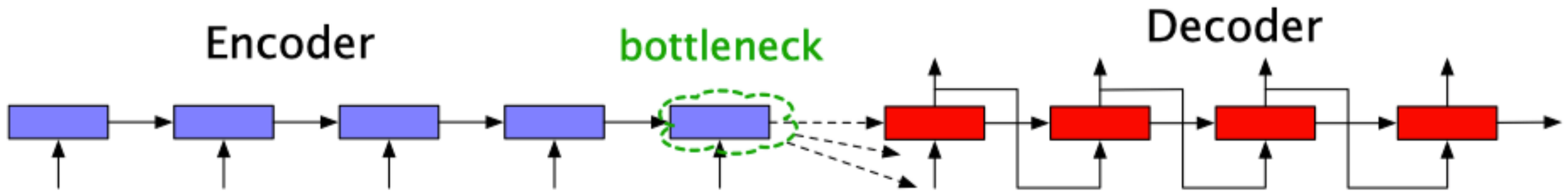


Training the Encoder-Decoder Model

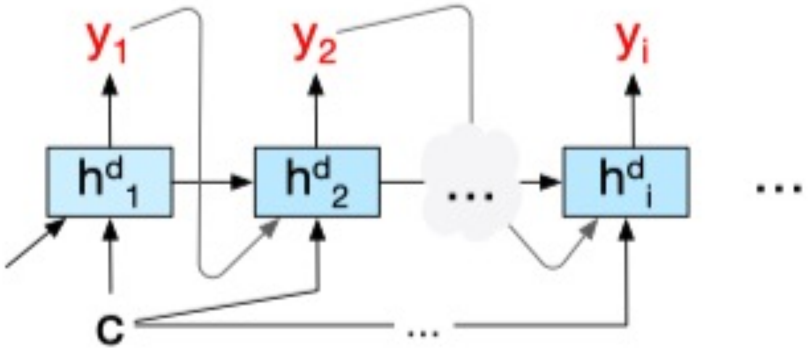
- Encoder-Decoder architectures are trained end-to-end.

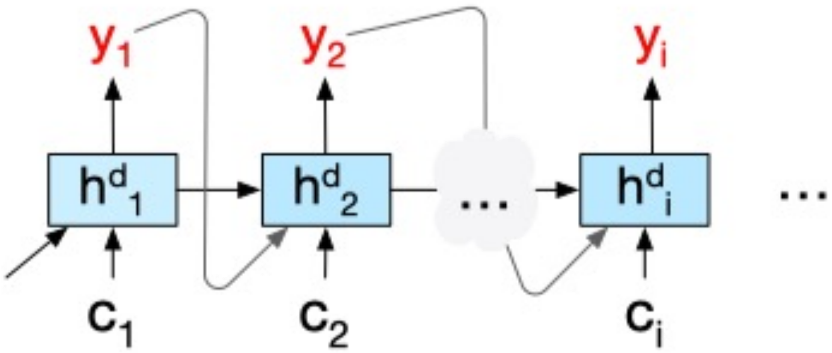


The Bottleneck



Attention Mechanism

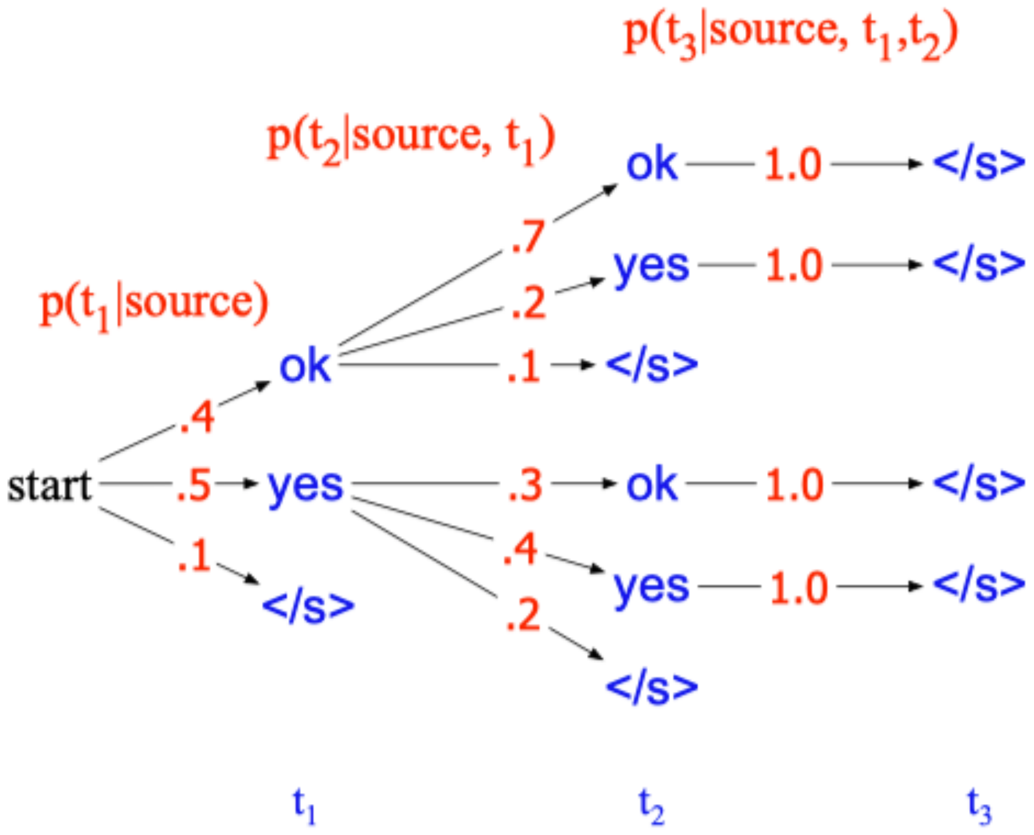




Beam Search

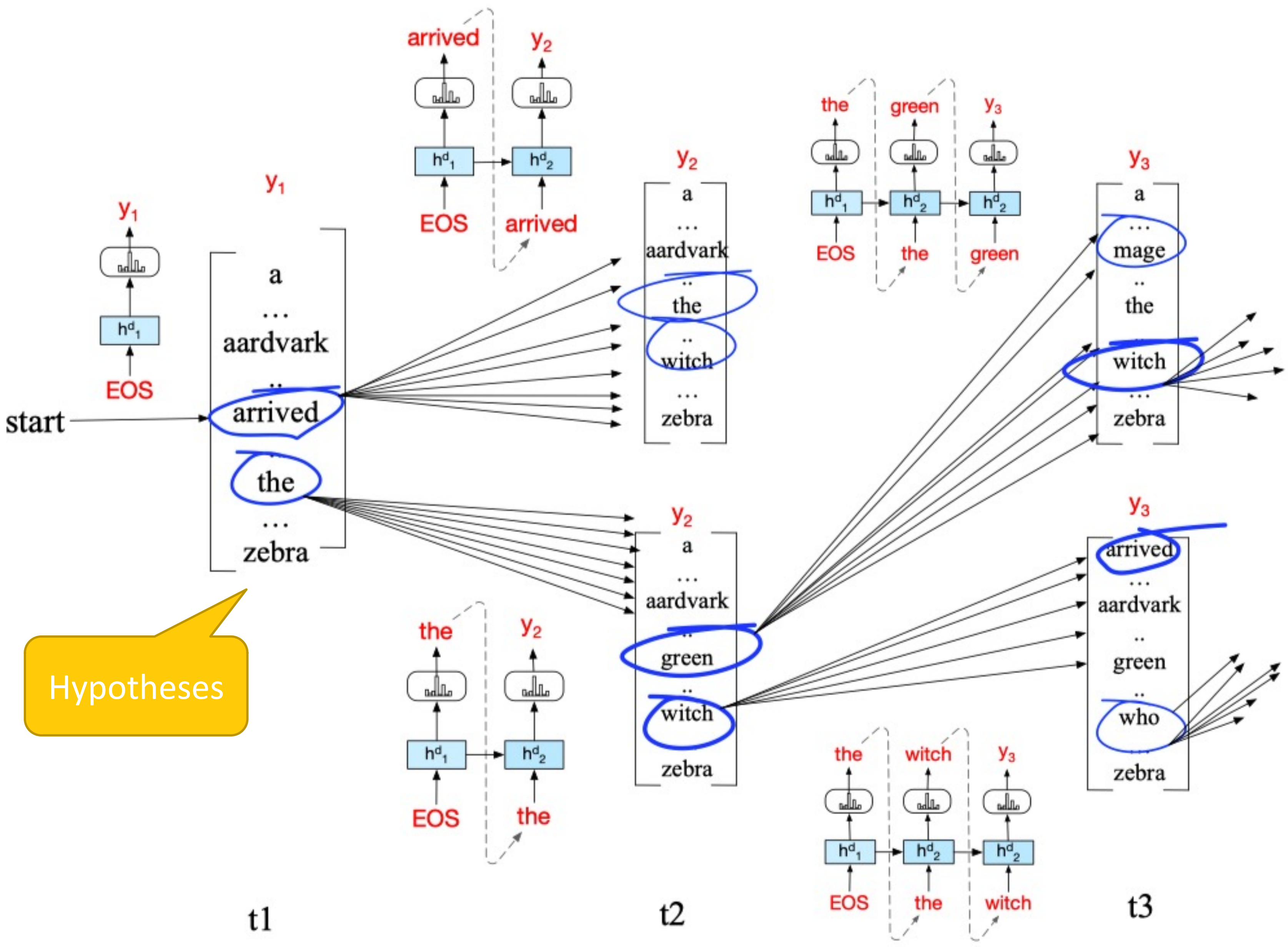
- Instead of choosing the best token to generate at each timestep, *k* possible tokens at each step are kept.
  - *k* = beam width
- In the decoding stage, we compute a softmax over the entire vocabulary (assigning a probability to each word).
- The k-best options from this softmax output is then selected.

Search Tree



p(t3|source, t1,t2)
p(t2|source, t1)
p(t1|source)
start
.4
.5
.1
ok
yes
</s>
.7
.2
.1
.3
.4
.2
ok
yes
</s>
ok
yes
</s>
1.0
</s>
t1
t2
t3

Beam Search Encoding



Beam Search Scoring

$$
\begin{aligned}
score(y) &= \log P(y|x) \\
&= \log\left(P(y_1|x)P(y_2|y_1,x)P(y_3|y_1,y_2,x)...P(y_t|y_1,...,y_{t-1},x)\right) \\
&= \sum_{i=1}^{t} \log P(y_i|y_1,...,y_{i-1},x)
\end{aligned}
$$

Normalization

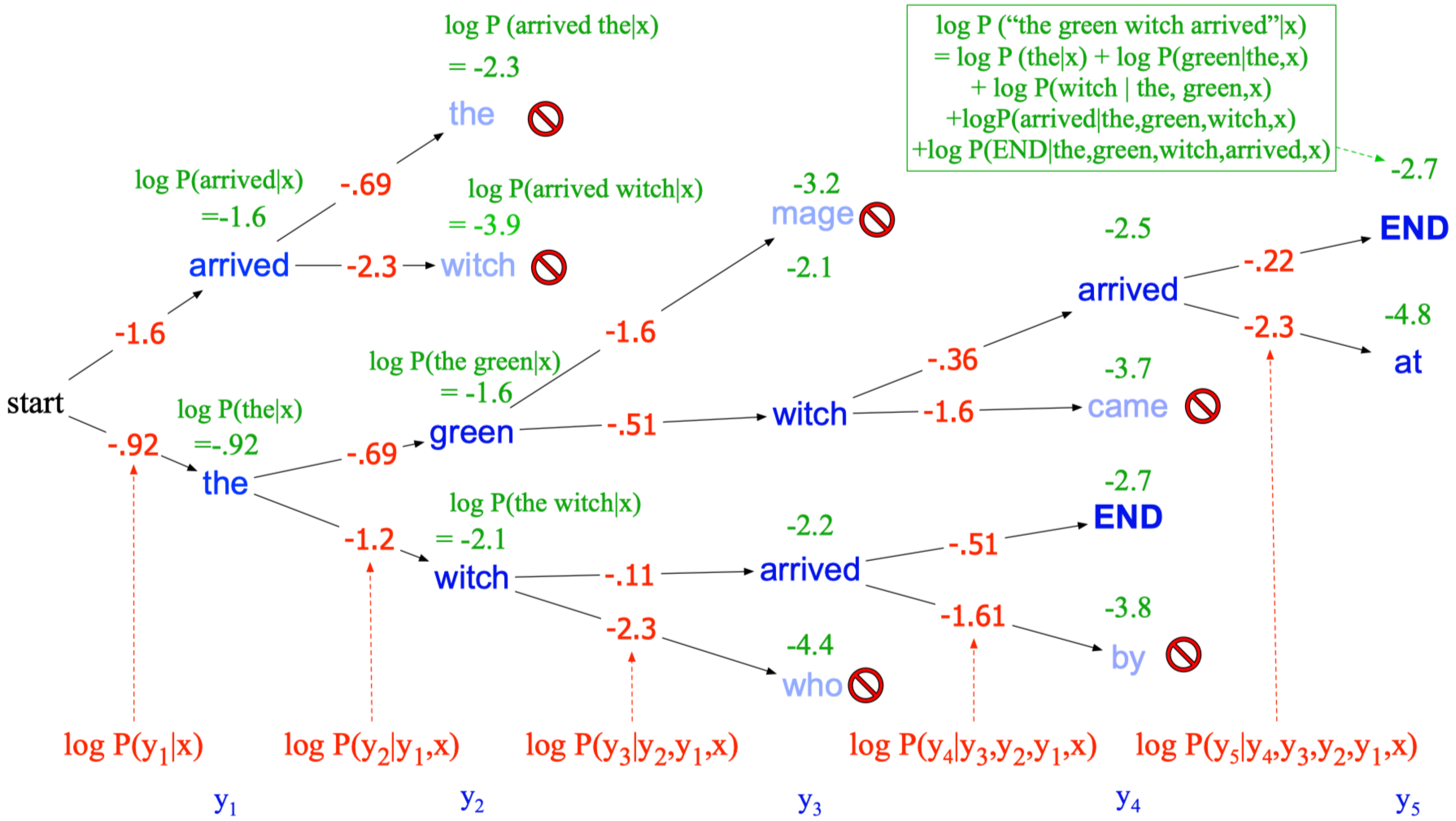
$$score(y) = -\log P(y|x) = \frac{1}{T}\sum_{i=1}^{t} -\log P(y_i|y_1,\ldots,y_{i-1},x)$$

The number of words

- Scoring without normalization

$$\begin{aligned} score(y) &= \log P(y|x) \\ &= \log\left(P(y_1|x)P(y_2|y_1,x)P(y_3|y_1,y_2,x)\ldots P(y_t|y_1,\ldots,y_{t-1},x)\right) \\ &= \sum_{i=1}^{t} \log P(y_i|y_1,\ldots,y_{i-1},x) \end{aligned}$$

Beam Search Encoding k = 2



Building Machine Translation Systems

Tokenization

- BPE or wordpiece is are used to generate the (fixed) vocabulary.
- Special symbol is added at the beginning of each token.

Words:	Jet makers feud over seat width with big orders at stake
Wordpieces:	_J et _makers _fe ud _over _seat _width _ with _big _orders _at _stake

- Wordpiece/BPE lexicon that contains both the source and the target language is built.

Wordpiece Algorithm

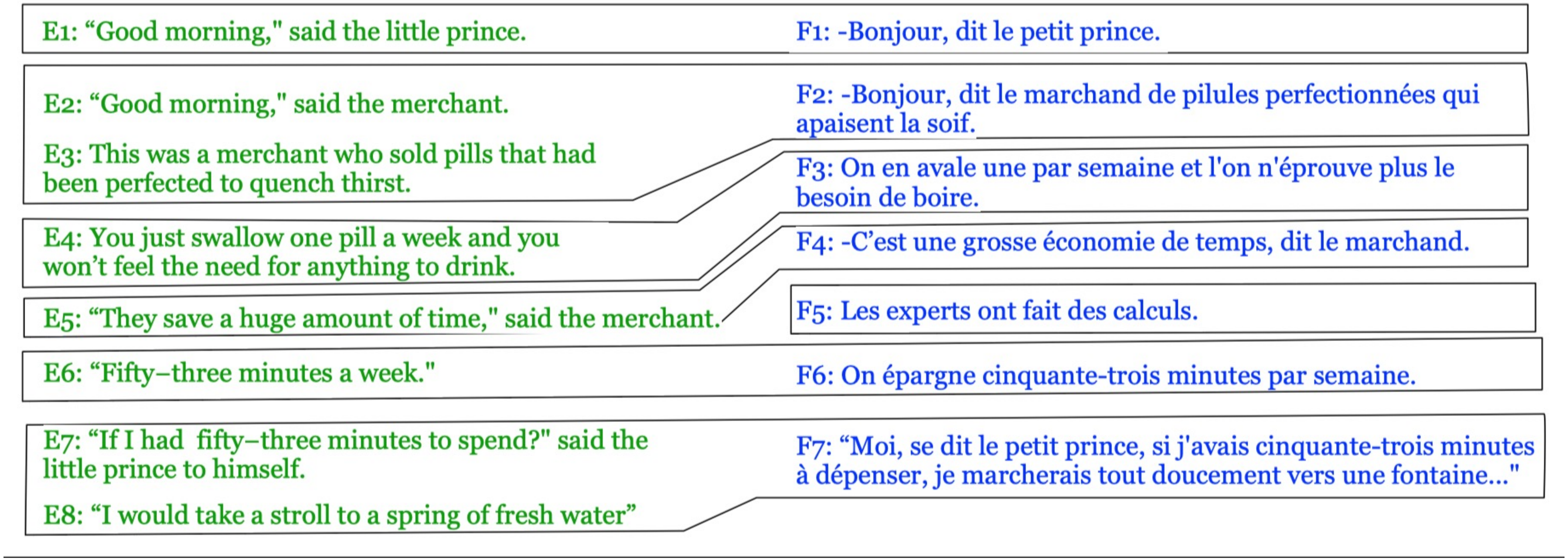
Vocabulary of 8K to 32K word pieces is commonly used

- 1. Initialize the wordpiece lexicon with characters (for example a subset of Unicode characters, collapsing all the remaining characters to a special unknown character token).
- 2. Repeat until there are V wordpieces:
  - Train an n-gram language model on the training corpus, using the current set of wordpieces.
  - Consider the set of possible new wordpieces made by concatenating two wordpieces from the current lexicon.
  - Choose the one new wordpiece that most increases the language model probability of the training corpus.

Machine Translation Corpora

- Machine translation models are trained on a parallel corpus (or bitext).
- Europarl: European Parliament
  - 400k to 2 million sentences
  - 21 European languages
- UN Parallel Corpus
  - 10 million sentences
  - Arabic, Chinese, English, French, Russian, Spanish
- OpenSubtitles: Movie and TV subtitles
- ParaCrawl: CommonCrawl
  - 223 million sentences
  - 23 EU Languages + English

Sentence Alignment



Backtranslation

- We're often short of data for training MT models
  - parallel corpora may be limited for particular languages or domains.
- However, we can find a large monolingual corpus, to add to the smaller parallel corpora that are available.
- Backtranslation is a way of making use of monolingual corpora in the target language by creating synthetic bitexts.
  - We train an intermediate target-to-source MT system on the small bitext to translate the monolingual target data to the source language.
  - This synthetic bitext is then added to the training data.

MT Evaluation

- Human Raters
- Automatic Evaluation
  - Metrics
    - BLEU (BiLingual Evaluation Understudy)
    - NIST
    - TER
    - Precision and Recall
    - METEOR
  - Embedding-Based Methods

BLEU

Source

la verdad, cuya madre es la historia, émula del tiempo, depósito de las acciones,
testigo de lo pasado, ejemplo y aviso de lo presente, advertencia de lo por venir.

Reference

truth, whose mother is history, rival of time, storehouse of deeds,
witness for the past, example and counsel for the present, and warning for the future.

Candidate 1

truth, whose mother is history, voice of time, deposit of actions,
witness for the past, example and warning for the present, and warning for the future

Candidate 2

the truth, which mother is the history, émula of the time, deposition of the shares,
witness of the past, example and notice of the present, warning of it for coming

BERTSCORE Recall