Machine Translation and Advanced Topics on LSTMs

COSC 7336: Advanced Natural Language Processing
Fall 2017

UNIVERSITY of HOUSTON

Some content on these slides was borrowed from Riloff, Money, and Socher and Manning.

UNIVERSIDAD NACIONAL DE COLOMBIA

Announcements

- ★ Reminder: Paper presentation sign up coming up
  - Presentation slides due Nov. 9th 11:59pm
  - Link: https://www.dropbox.com/request/2cWaeglMImqO5DQGYMWp
- ★ Final Project Proposals due Nov. 10th!
  - What is the problem
  - What kind of data do you have available
  - What approach you plan to use
  - Link: https://www.dropbox.com/request/YFkWhgS0c22iqzLETjEa

Today’s lecture

- ★ Short intro to Machine Translation (MT)
- ★ Challenges in MT
- ★ Pre-Deep Learning Era
- ★ Sequence to Sequence models with RNN
- ★ Attention
- ★ Translation using seq2seq models

Machine Translation (MT)

UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

MT Definition

- ★ Transform input text *s*, in source language *a*, into an equivalent text *t* in target language *b*.
- ★ Good translation:
  - Faithful
  - Natural
- ★ Many practical reasons for MT

Example Translations from Google Translate

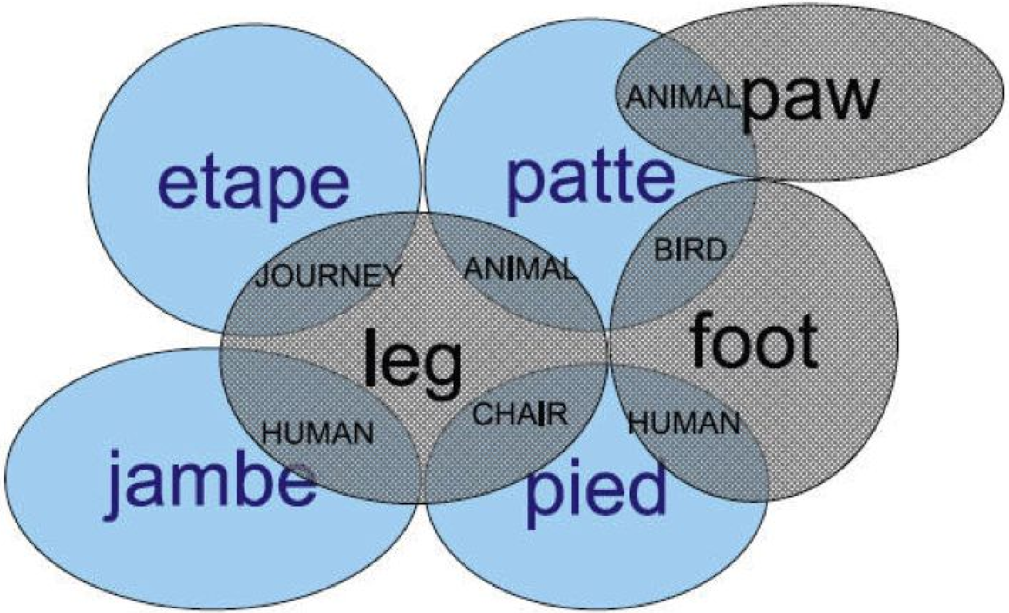
There is a lot at night. The oil lamps, which hang from a nail in front of the door, but the light floats like a bright almond tree, it is difficult to shake, it is terrible, unstable, to keep the dark deposit around it and the house up and down. until the last corners, where the darkness is so thick that it seems solid.

The night has much to last. The oil lamp, hanging from a nail next to the door, is lit, but the flame, like a luminous almond tree floating, barely manages, tremulous, unstable, to hold the dark mass that surrounds it and fills the house from top to bottom, until the last corners, where the darkness, so thick, seems to have become solid.

Example Translations from Google Translate

La noche tiene aún mucho que durar. El candil de aceite, colgado de un clavo al lado de la puerta, está encendido, pero la llama, como una almendrilla luminosa flotante, apenas consigue, trémula, inestable, sostener la masa oscura que la rodea y llena de arriba abajo la casa, hasta los últimos rincones, allí donde las tinieblas, de tan espesas, parecen haberse vuelto sólidas.

What makes MT difficult?



What makes MT difficult?

Differences between languages

★ Morphological differences

- From isolating like Cantonese to polysynthetic languages like Eskimo
- From agglutinative, like Turkish to fusion languages like Russian

What makes MT difficult?

Differences between languages (2)

- Syntactic divergences
  - Subject-Verb-Object (SVO) like English
  - SOV like Hindi and Japanese
  - VSO languages like Irish and Arabic

What makes MT difficult?

Differences between languages (2)

★ Allowable omissions

- *Pro-drop* languages regularly omit subjects that must be inferred
- [Tu madre]$_i$ llamó en la tarde. q$_i$ Dijo que te esperaba a comer mañana.
- Your mother] called this afternoon. [She] said she will see you tomorrow for lunch.

What makes MT difficult?

Differences between languages (3)

★ Lexical divergences that require specification

“John *plays* the guitar.” → “John *toca* la guitarra.”

“John *plays* tennis.” → “John *juega* tennis.”

“The singer wore a purple attire” → “*La* cantante usó un traje morado” | *El* cantante usó un traje morado”.

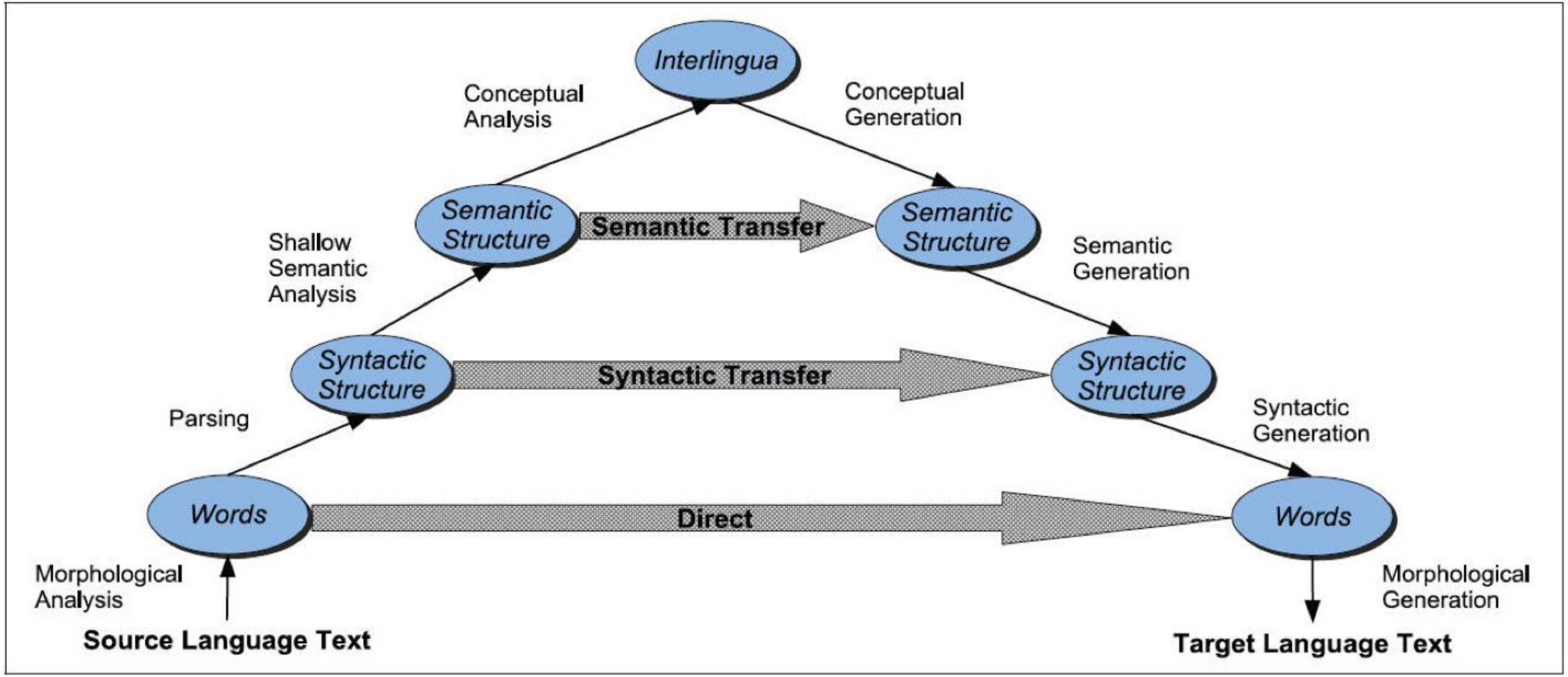
What makes MT difficult?

Differences between languages (4)

★ Lexical gaps

- Rivière (river that flows into ocean) and fleuve (river that does not flow into ocean) in French
- Schedenfraude (feeling good about another's pain) in German.
- Oyakoko (filial piety) in Japanese
- Xiào in Chinese

MT Approaches



UNIVERSITY of HOUSTON

UNIVERSIDAD NACIONAL DE COLOMBIA

Statistical MT (SMT)

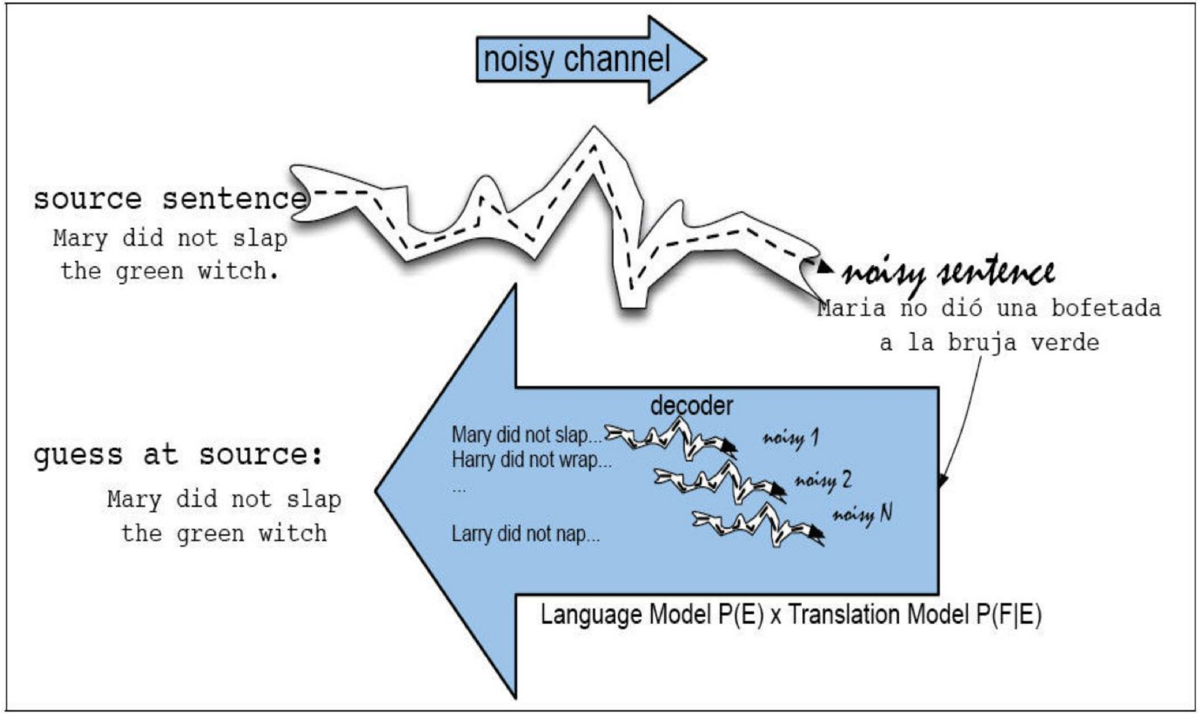
- Before DL, best methods were SMT
  - Trained on large amounts of parallel data
    - Canadian Hansard
    - European parliament corpora
  - But:
    - Corresponding sentences are not marked.
    - Paragraph boundaries may not be consistent.
    - Entire sentences or even paragraphs may be present in one but missing in the other!

SMT

A good translation should be ***faithful*** and ***fluent***, Final objective:

$$T_{best} = \underset{T \in \mathrm{Target}}{\operatorname{argmax}} \ \mathrm{faithfulness}(T, S)\ \mathrm{fluency}(T)$$

Noisy Channel Model for SMT



SMT

- Formulation following Bayes rule:

$$
\begin{aligned}
\hat{E} &= \operatorname*{argmax}_{E \in English} P(E \mid F) \\
&= \operatorname*{argmax}_{E \in English} \frac{P(F \mid E) P(E)}{P(F)} \\
&= \operatorname*{argmax}_{E \in English} \underbrace{P(F \mid E)}_{\text{Translation Model}} \underbrace{P(E)}_{\text{Language Model}}
\end{aligned}
$$

SMT

French → Translation Model $p(f|e)$ → Pieces of English → Language Model $p(e)$

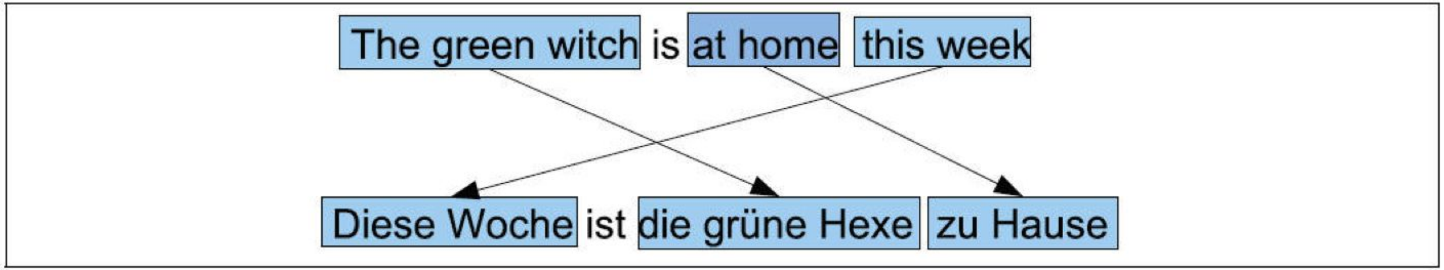
Translation Model $p(f|e)$ → Decoder

Language Model $p(e)$ → Decoder

Decoder $\mathrm{argmax}\ p(f|e)p(e)$ → Proper English

Phrase-Based SMT

A good way to compute P(F|E) is by considering the behavior of *phrases*



Phrase-Based SMT

- Base P(*F* | *E*) on translating phrases in *E* to phrases in *F*.
- First segment *E* into a sequence of phrases $\bar{e}_1, \bar{e}_1, \ldots, \bar{e}_l$
- Then translate each phrase $\bar{e}_i$, into f_i, based on *translation probability* $\Phi(f_i \mid \bar{e}_i)$
- Then reorder translated phrases based on *distortion probability* *d*(*i*) for the *i*th phrase.

$$P(F \mid E) = \prod_{i=1}^{I} \phi(\bar{f}_i, \bar{e}_i) d(i)$$

Translation Probabilities

- Assume a *phrase aligned* parallel corpus is available or constructed that shows matching between phrases in *E* and *F*.
- Then compute (MLE) estimate of f based on simple frequency counts.

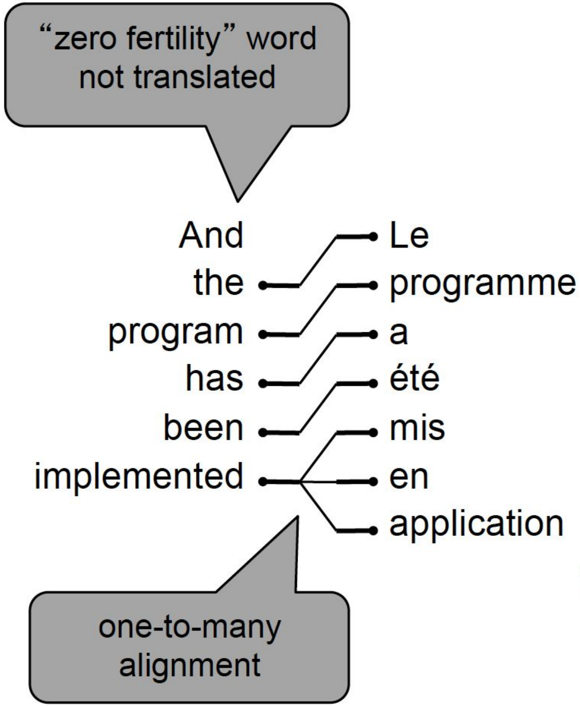
$$\phi(\bar{f},\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{f} \text{count}(\bar{f},\bar{e})}$$

Alignment

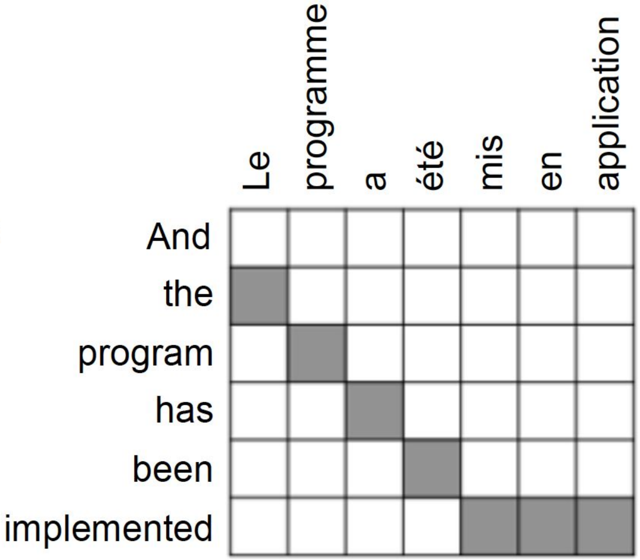
To train the translation model we need to know which words belong to which other words in the target language

★ It’s a really hard problem!

Alignment (2)

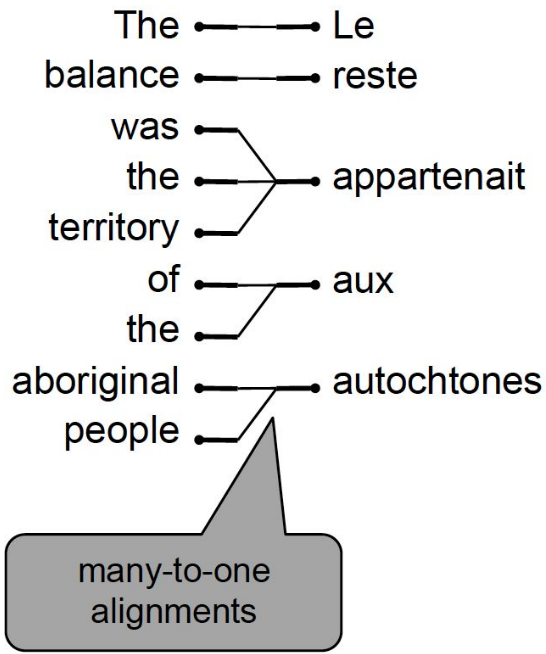


“zero fertility” word not translated
And
the
program
has
been
implemented
Le
programme
a
été
mis
en
application
one-to-many alignment

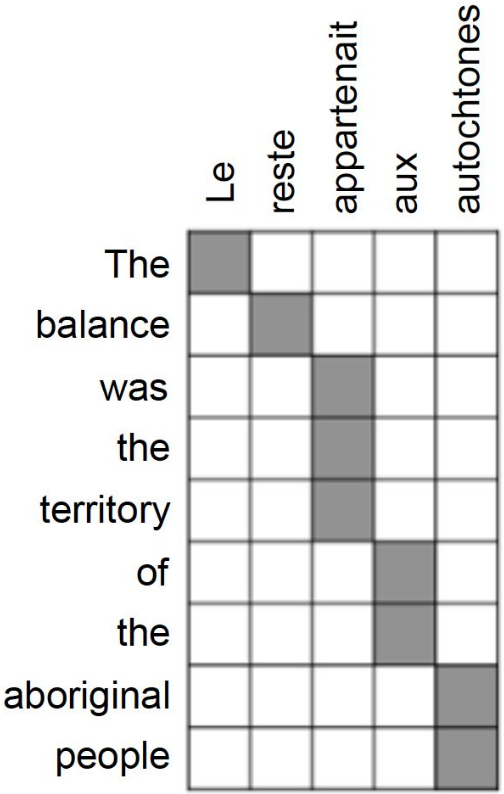


Le
programme
a
été
mis
en
application
And
the
program
has
been
implemented

Alignment (3)



The
balance
was
the
territory
of
the
aboriginal
people
Le
reste
appartenait
aux
autochtones
many-to-one alignments



Le
reste
appartenait
aux
autochtones
The
balance
was
the
territory
of
the
aboriginal
people

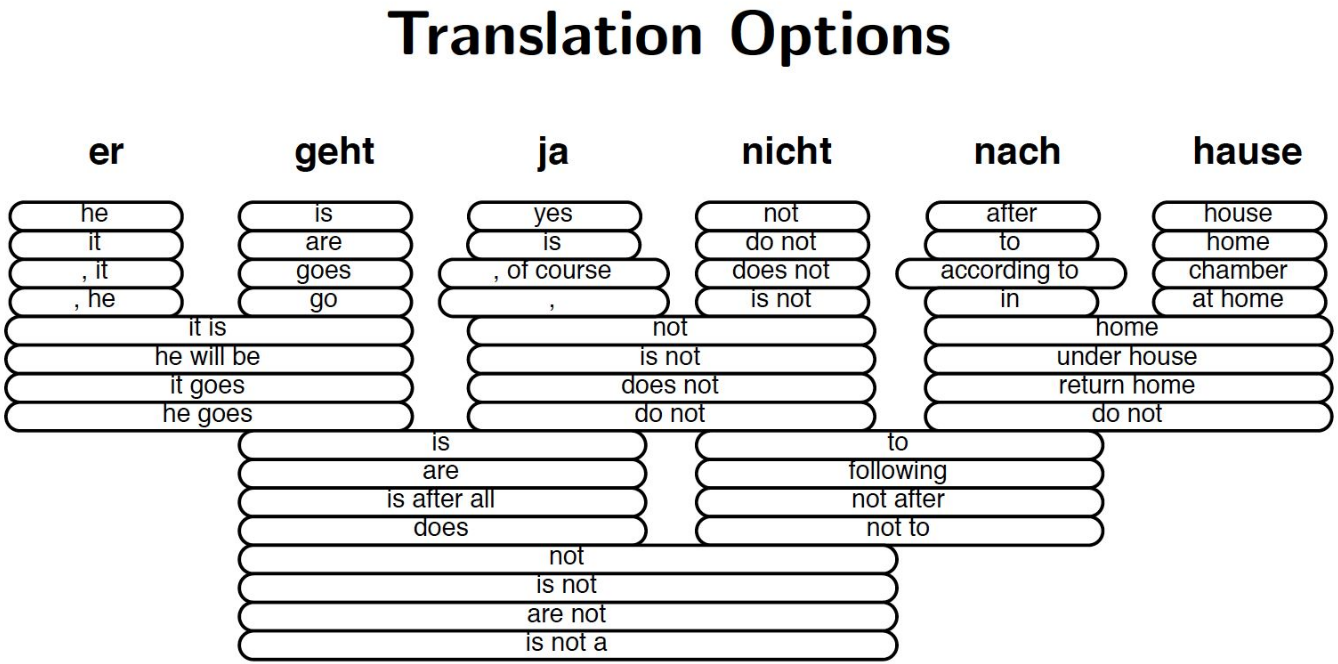
Decoding

- Assuming we have solved the alignment problem we can then estimate phrase translation probabilities'
- What's next?

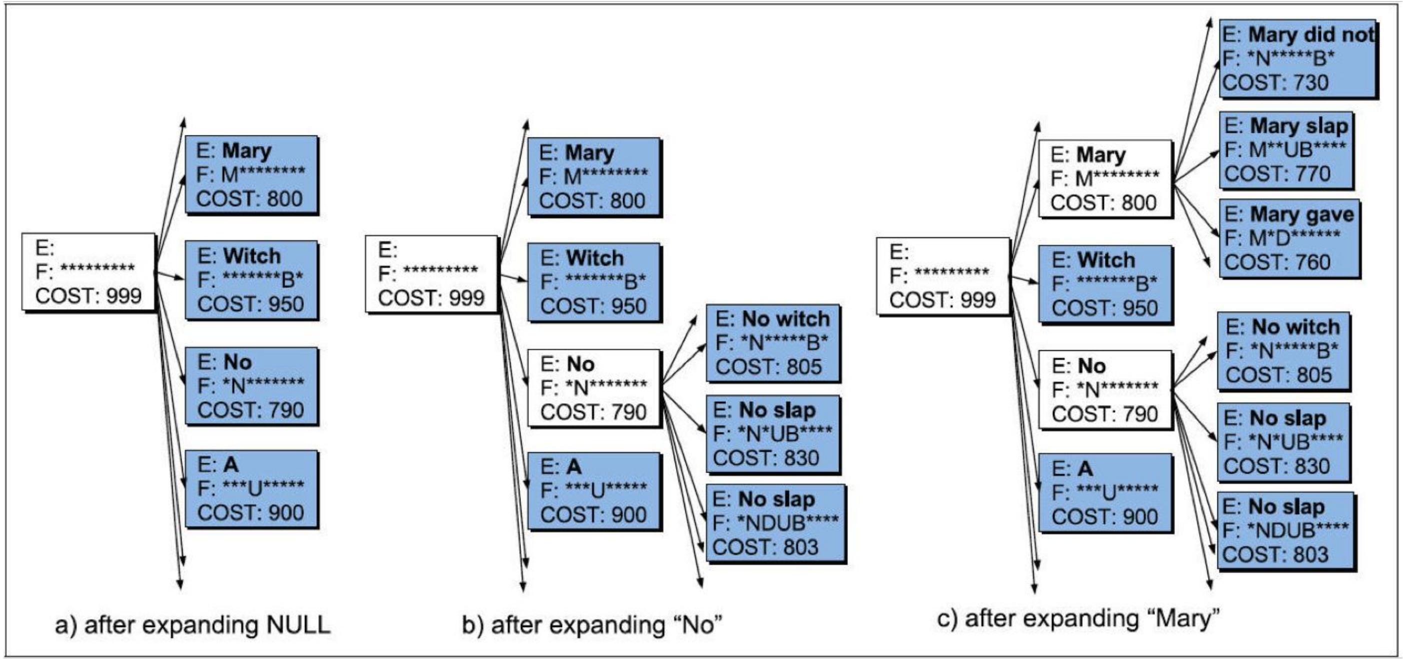
Decoder

$$\arg\max p(f|e)p(e)$$

After Alignment There's a Lot More!



After Alignment There's a Lot More!



Evaluation of MT Systems

Evaluation of MT Systems

- Human subjective evaluation is the best but is time-consuming and expensive.
- Automated evaluation comparing the output to multiple human reference translations is cheaper and correlates with human judgments.

Automatic Evaluation of MT

- Collect one or more human ***reference translations*** of the source.
- Compare MT output to these reference translations.
- Score result based on similarity to the reference translations.
  - BLEU
  - NIST
  - TER
  - METEOR

BLEU

- ★ Determine number of *n*-grams of various sizes that the MT output shares with the reference translations.
- ★ Compute a modified precision measure of the *n*-grams in MT result.

BLUE Example

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

BLUE Example

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 Bigram Precision: 1/5

Modified N-gram Precision

Average *n*-gram precision over all *n*-grams up to size *N* (typically 4) using geometric mean.

$$p_n = \frac{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{C \in corpus} \sum_{\text{n-gram} \in C} \text{count}(\text{n-gram})} \qquad p = \sqrt[N]{\prod_{n=1}^{N} p_n}$$

Brevity Penalty

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

c = total length of the candidate translation corpus

r = effective reference length

BLEU Score

Final BLEU Score: BLEU = *BP* x *p*

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$$c = 6, \quad r = 7, \qquad BP = e^{(1-7/6)} = 0.846$$
$$BLEU = 0.846 \times 0.408 = 0.345$$

Discussion Points

★ SMT was state-of-the-art before Deep NLP

★ Evaluation metrics can be improved

★ SMT relies heavily on parallel corpora