PHRASE-BASED MACHINE TRANSLATION

David Talbot

Spring 2017

Yandex School of Data Analysis

Given foreign sentence f and a set of possible translations E, choose translation e^* s.t.

$$\begin{aligned} e^* &= \operatorname*{argmax}_{e \in E} \Pr(e|f) \\ &= \operatorname*{argmax}_{e \in E} \Pr(e)\Pr(f|e) \end{aligned}$$

Why might the second line be easier to deal with?

$$e^* = \operatorname*{argmax}_{e \in E} \Pr(e)\Pr(f|e)$$

- Pr(e) models the *fluency* of the translation
- Pr(f|e) models the *adequacy* of the translation
- argmax is the search problem implemented by a *decoder*

Modelling Pr(e|f) directly, we would need to handle fluency and adequacy simultaneously which is hard.

- Language models Pr(e) help us choose translations that sound good in the target language.
- Goal 1: Assign high probability to well formed candidates:
  - ”The cat in the hat.”
  - ”Green eggs and ham.”
- Goal 2: Assign low probability to malformed candidates:
  - ”Cat the hat in the.”
  - ”Eggs ham green and.”

”I don’t need to remember everything to predict the next ...”

- *N*-gram models assume each word is conditionally independent given previous $n-1$ words, e.g.

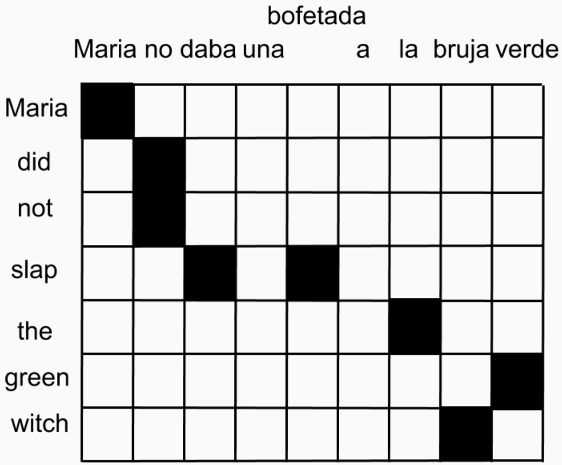
$$\mathrm{Pr}(\mathrm{e}) \quad \approx \quad \prod_i \mathrm{Pr}(e_i | e_{i-1}, \mathrm{e}_{i-2})$$

- What parameters does this model have?
- How could we estimate them?
- What problems will we have with this model?

- Not so obvious how to factorize $\Pr(f|e)$
- Would be easier if we could see how the translator worked...
- IBM researchers introduced *word alignments* (1990)

- Alignments provide a *generative* story for the data
- Source words *generate* target words aligned to them
- Alignments can be one-to-one, one-to-many, many-to-one

WORD ALIGNMENT MATRIX



How well can this model represent the data ?

- Choose $a_1 = 1$, generate "*Maria*" given "*Maria*"
- Choose $a_2 = 3$, generate "*no*" given "*not*"
- Choose $a_3 = 2$, generate "*daba*" given "*did*" $\ldots$

Maria (did) not slap the green witch.
Maria no daba una bofetada (a) la bruja verde.

These models differ only in the prior over alignments

- IBM Model 1 (uniform)

$$\Pr(a_j = i|\mathrm{e}) \approx \epsilon$$

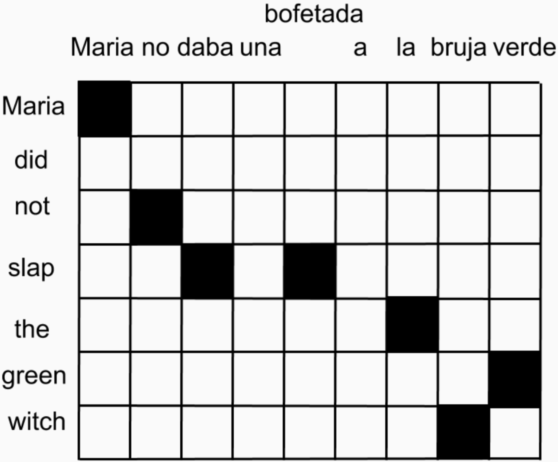
- IBM Model 2 (independent with positional bias)

$$\Pr(a_j = i|\mathrm{e}) \approx p(a_j = i|j, I, J)$$

- HMM (Markov dependency with relative bias)

$$\Pr(a_j = i|\mathrm{e}) \approx h(a_j = i|a_{j-1} = i', I, J)$$

Only one source word aligned to each target word



New generative story

1. Choose how many target words ϕ_i to generate from each source word e_i
2. Choose whether to insert NULL token
3. Choose how to order each group of words
4. Choose list of target words τ to generate

Why is not possible to train this model with full EM?
What other algorithms could we use?


Maria did not slap the green witch
1. Choose ‘fertility’
Maria not slap slap slap the green witch
2. Insert NULL
Maria not slap slap slap NULL the green witch
3. Reorder
Maria not slap slap slap NULL the witch green
4. Generate
Maria no daba una bofetada a la bruja verde

PROBLEMS WITH WORD BASED MODELS

- Word based models are still used for alignment
- Rarely used for translation
- They make unrealistic independence assumptions
- Translations don’t consider context
- Reordering model is very weak
- Generating the target sentence requires many steps

phrase based machine translation (koehn 2003)

1. Estimate translation probabilities for *phrases* extracted from word aligned data
2. Add *feature functions* for *length* and *reordering*
3. Decode using a simple stack based algorithm

Basis for popularization of MT (Google, Yandex, Bing)

Optimization

$$
\begin{aligned}
e^* &= \operatorname*{argmax}_{e} \Pr(e|f) && (1)\\
&= \operatorname*{argmax}_{e} \Pr(f|e)\Pr_{LM}(e)\omega^{length_e} && (2)
\end{aligned}
$$

where ω is a new free parameter.

Translation model defined over phrases $(\hat{e}, \hat{f})$ rather than words

$$\Pr(\mathrm{f}|\mathrm{e}) = \prod_i \phi(\hat{f}_i|\hat{e}_i) d(a_i - b_{i-1})$$

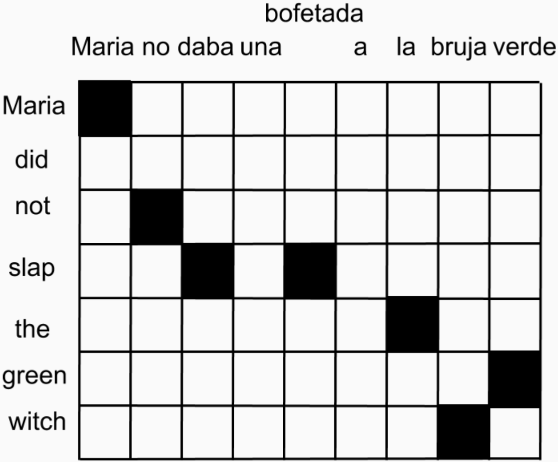
where a_i is the start index of the source phrase translated as the i-th phrase and b_{i-1} is end index of the previously translated source phrase.

Translation model $\phi(\cdot)$ defined over phrases $(\hat{e}, \hat{f})$ rather than words is estimated using word aligned text.

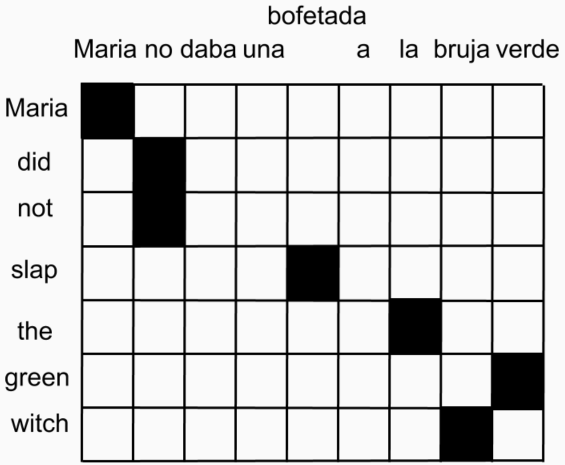
$$\phi(\hat{f}|\hat{e}) = \frac{count(\hat{f}, \hat{e})}{\sum_{\hat{f}'} count(\hat{f}', \hat{e})}$$

	Maria	no	daba	una	bofetada	a	la	bruja	verde
Maria									
did									
not									
slap									
the									
green									
witch									

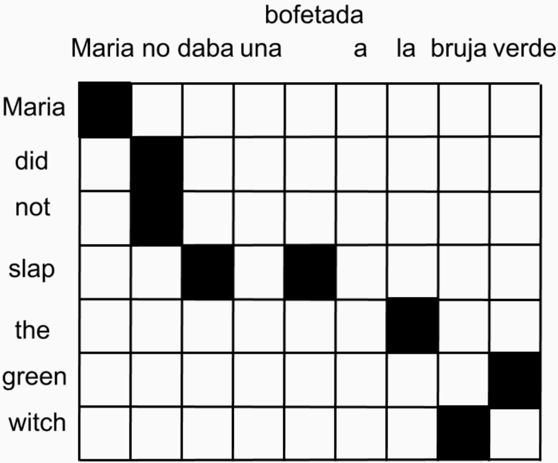
WORD ALIGNMENT MATRIX: pr(f|e)

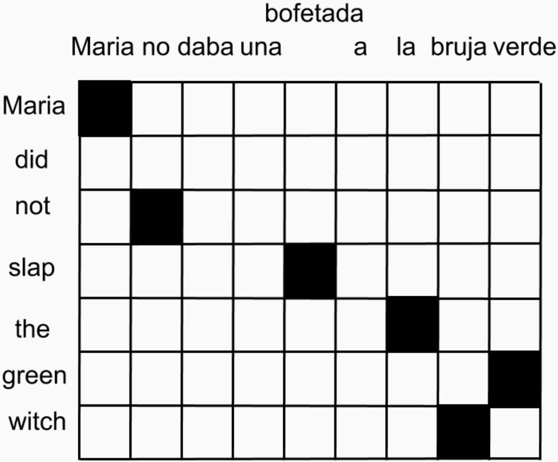


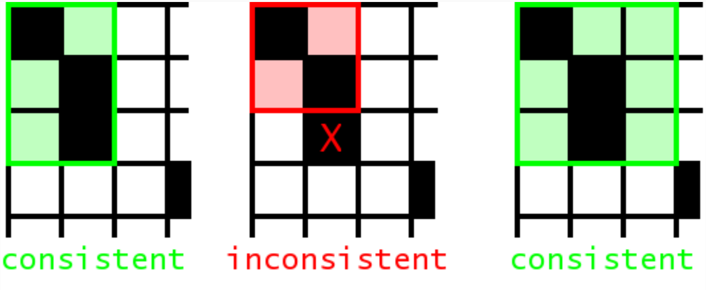
WORD ALIGNMENT MATRIX: pr(e|f)



WORD ALIGNMENT MATRIX: UNION

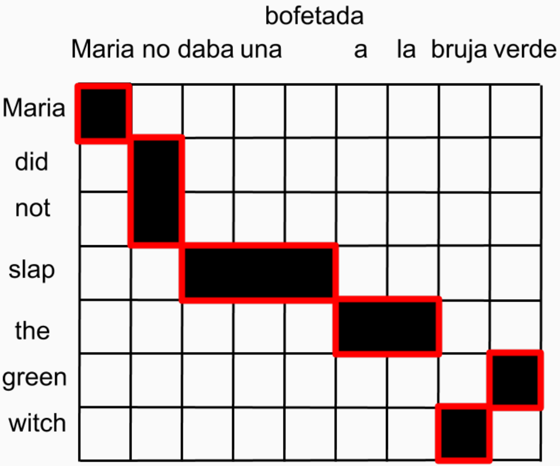



bofetada
Maria no daba una a la bruja verde
Maria
did
not
slap
the
green
witch

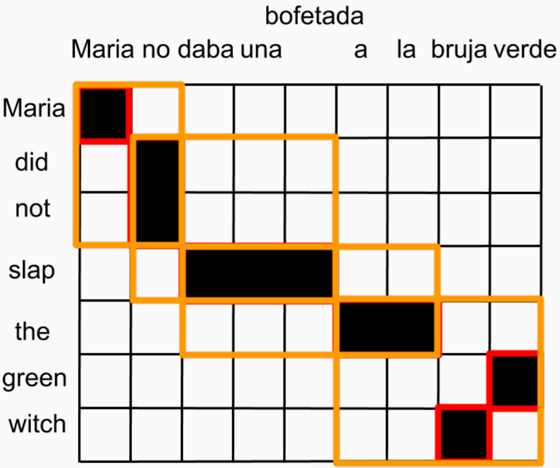


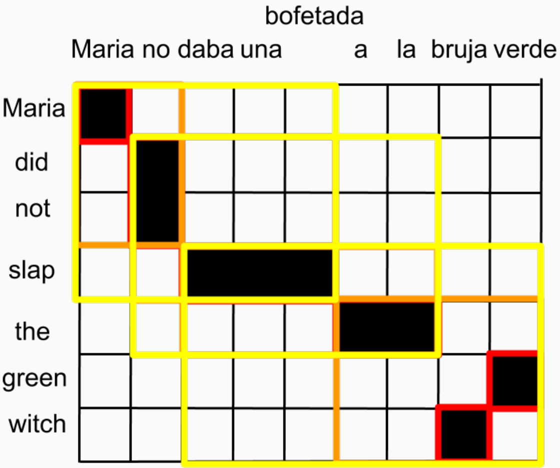
Word alignments constrain the set of possible phrase pairs.

WORD ALIGNMENT MATRIX: INITIAL PHRASES



WORD ALIGNMENT MATRIX: EXTENSIONS




bofetada
Maria no daba una a la bruja verde
Maria
did
not
slap
the
green
witch

- Intersection: high confidence but sparse
- Union: more direct phrases, but also more constraints
- Null aligned words aren't a huge problem
- Many different ways of segmenting the translation
- Not a generative model

Which will produce the most phrase pairs?