Word Alignment

David Kauchak
CS159 – Fall 2020

Some slides adapted from

Philipp Koehn
School of Informatics
University of Edinburgh

Kevin Knight
USC/Information Sciences Institute
USC/Computer Science Department

Dan Klein
Computer Science Department
UC Berkeley

1

Quiz 2

Mean 23.7 (84.6%)

Quartile 1: 22 (79%)
Quartile 2: 24 (85.7%)
Quartile 3: 26 (92.8%)

2

Admin

Assignment 5

3

Language translation

4

Problems for Statistical MT

Preprocessing
- How do we get aligned bilingual text?
- Tokenization
- Segmentation (document, sentence, word)

Language modeling
- Given an English string e, assigns P(e) by formula

Translation modeling
- Given a pair of strings <f,e>, assigns P(f | e) by formula

Decoding
- Given a language model, a translation model, and a new sentence f … find translation e maximizing P(e) * P(f | e)

Parameter optimization
- Given a model with multiple feature functions, how are they related? What are the optimal parameters?

Evaluation
- How well is a system doing? How can we compare two systems?

5

Translation Model

Want: probabilistic model gives us how likely one sentence is to be a translation of another, i.e *p(foreign | english)*

Mary did not slap the green witch

↓

Maria no dió una botefada a la bruja verde

Can we just model this directly, i.e. p(foreign | english) ?
How would we estimate these probabilities, e.g.
p("Maria …" | "Mary …")?

6

Translation Model

Want: probabilistic model gives us how likely one sentence is to be a translation of another, i.e *p(foreign | english)*

Mary did not slap the green witch

↓

Maria no dió una botefada a la bruja verde

$$p(\text{"Maria…"} \mid \text{"Mary…"}) = \frac{\text{count("Mary…" aligned-to "Maria…")}}{\text{count("Mary…")}}$$

Not enough data for most sentences!

7

Translation Model

Key: break up process into smaller steps

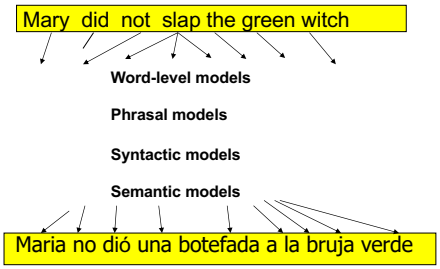
Mary did not slap the green witch

sufficient statistics for smaller steps

Maria no dió una botefada a la bruja verde

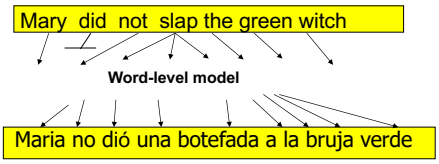
8

What kind of Translation Model?



9

IBM Word-level models

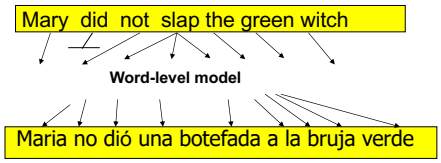


Generative story: description of how the translation happens

1. Each English word gets translated as 0 or more Foreign words
2. Some additional foreign words get inserted
3. Foreign words then get shuffled

10

IBM Word-level models



Each foreign word is *aligned* to exactly one English word.

Key idea: decompose *p(foreign | english)* into word translation probabilities of the form *p(foreign_word | english_word)*

IBM described 5 different levels of models with increasing complexity (and decreasing independence assumptions)

11

Some notation

$E = e_1 e_2 ... e_{|E|}$ English sentence with length |E|

$F = f_1 f_2 ... f_{|F|}$ Foreign sentence with length |F|

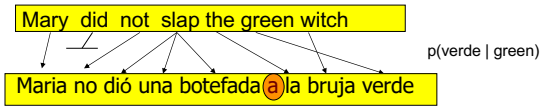
Mary did not slap the green witch
e_1 e_2 e_3 e_4 e_5 e_6 e_7

f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9
Maria no dió una botefada a la bruja verde

Translation model: $p(F \mid E) = p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|})$

12

Word models: IBM Model 1



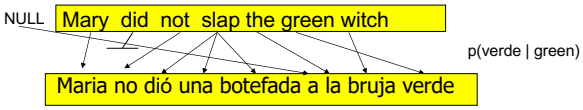
Each foreign word is aligned to exactly one English word

This is the **ONLY** thing we model!

Does the model handle foreign words that are not aligned, e.g. "a"?

13

Word models: IBM Model 1



Each foreign word is aligned to exactly one English word

This is the **ONLY** thing we model!

Include a "NULL" English word and align to this to account for deletion

14

Word models: IBM Model 1

generative story -> probabilistic model

– Key idea: introduce "hidden variables" to model the word alignment

$$p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|})$$

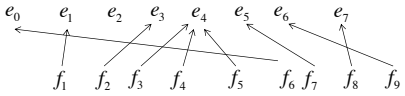
$$p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|})$$

- one variable for each foreign word
- a_i corresponds to the ith foreign word
- each a_i can take a value 0…|E|

15

Alignment variables

Mary did not slap the green witch



Maria no dió una botefada a la bruja verde

a_1	1
a_2	3
a_3	4
a_4	4
a_5	4
a_6	0
a_7	5
a_8	7
a_9	6

16

Alignment variables

And the program has been implemented

e_0 e_1 e_2 e_3 e_4 e_5 e_6

Alignment?

f_1 f_2 f_3 f_4 f_5 f_6 f_7

Le programme a ete mis en application

17

Alignment variables

And the program has been implemented

e_0 e_1 e_2 e_3 e_4 e_5 e_6

f_1 f_2 f_3 f_4 f_5 f_6 f_7

Le programme a ete mis en application

a1	?
a2	?
a3	?
a4	?
a5	?
a6	?
a7	?

18

Alignment variables

And the program has been implemented

e_0 e_1 e_2 e_3 e_4 e_5 e_6

f_1 f_2 f_3 f_4 f_5 f_6 f_7

Le programme a ete mis en application

a1	2
a2	3
a3	4
a4	5
a5	6
a6	6
a7	6

19

Probabilistic model

$$p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|}) \quad \overset{?}{=} \quad p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|})$$

NO!

$$p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|}) \Rightarrow p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|})$$

How do we get rid of variables?

20

Joint distribution

NLPPass, EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is P(ENGPass)?

21

Joint distribution

NLPPass, EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

0.92

How did you figure that out?

22

Joint distribution

$$P(x) = \sum_{y \in Y} p(x, y)$$

Called "marginalization", aka summing over a variable

NLPPass, EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

23

Probabilistic model

$$p(f_1 f_2 \ldots f_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \sum_{a_1} \sum_{a_2} \ldots \sum_{a_{|F|}} p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|})$$

Sum over all possible values, i.e. marginalize out the alignment variables

24

Independence assumptions

IBM Model 1:

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

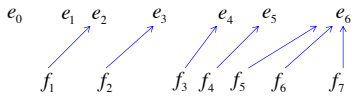
What independence assumptions are we making?

What information is lost?

25

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

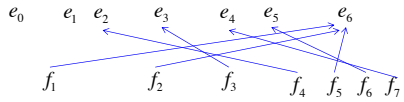
And the program has been implemented



Le programme a ete mis en application

Are the probabilities any different under model 1?

And the program has been implemented

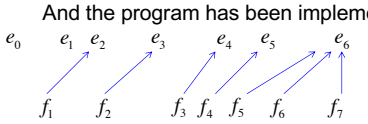


application en programme Le mis ete a

26

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

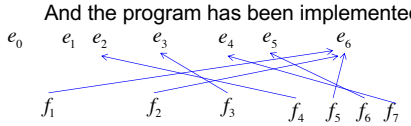
And the program has been implemented



Le programme a ete mis en application

No. Model 1 ignores word order!

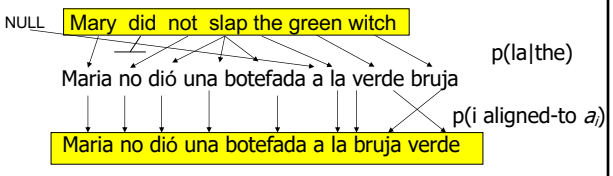
And the program has been implemented



application en programme Le mis ete a

27

IBM Model 2



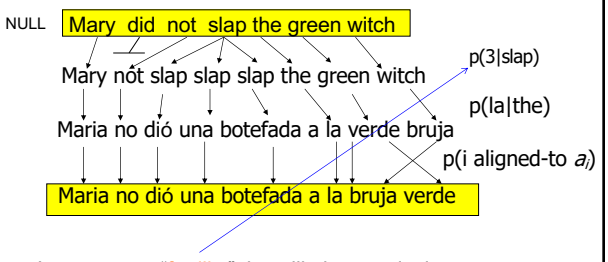
$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(i \text{ aligned-to } a_i) p(f_i \mid e_{a_i})$$

Models word movement by position, e.g.

- Words don't tend to move too much
- Words at the beginning move less than words at the end

28

IBM Model 3



Incorporates "fertility": how likely a particular English word is to produce multiple foreign words

29

Word-level models

Problems/concerns?

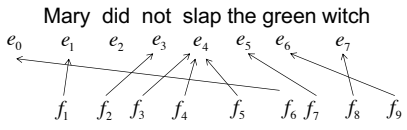
- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - "real estate", "note that", "interest in"
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

30

Benefits of word-level model

Rarely used in practice for modern MT systems

Why talk about them?



Two key side effects of training a word-level model:

- Word-level alignment
- p(f | e): translation dictionary

31

Training a word-level model

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

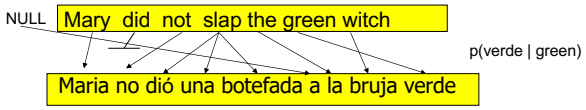
Where do these come from?

Have to learn them!

The old man is happy. He has fished many times.	———	El viejo está feliz porque ha pescado muchos veces.
His wife talks to him.	———	Su mujer habla con él.
The sharks await.	———	Los tiburones esperan.
…		…

32

Word models: IBM Model 1



Each foreign word is aligned to exactly one English word

This is the **ONLY** thing we model!

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

33

Training a word-level model

The old man is happy. He has fished many times.	El viejo está feliz porque ha pescado muchos veces.
His wife talks to him.	Su mujer habla con él.
The sharks await.	Los tiburones esperan.
…	…

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

$p(f_i \mid e_{a_i})$: probability that *e* is translated as *f*

How do we learn these?

What data would be useful?

34

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

$$p(f_i \mid e_{a_i}) = \; ?$$

35

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

$$p(f_i \mid e_{a_i}) = \frac{count(f \text{ aligned-to } e)}{count(e)}$$

p(el | the) = 0.5
p(Los | the) = 0.5

Any problems concerns?

36

Thought experiment

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

His wife talks to him.

Su mujer habla con él.

The sharks await.

Los tiburones esperan.

Getting data like this is expensive!

Even if we had it, what happens when we switch to a new domain/corpus

37

Thought experiment #2

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Annotator 1

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

Annotator 2

$$p(f_i \mid e_{a_i}) = \frac{count(f \text{ aligned-to } e)}{count(e)}$$

What do we do?

38

Thought experiment #2

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

80 annotators

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

20 annotators

$$p(f_i \mid e_{a_i}) = \frac{count(f \text{ aligned-to } e)}{count(e)}$$

What do we do?

39

Thought experiment #2

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

80 annotators

The old man is happy. He has fished many times.

El viejo está feliz porque ha pescado muchos veces.

20 annotators

$$p(f_i \mid e_{a_i}) = \frac{count(f \text{ aligned-to } e)}{count(e)}$$

Use partial counts:

- count(viejo | man) 0.8
- count(viejo | old) 0.2

40

Training without alignments

How should these be aligned?

There is some information!
(Think of the alien translation task last time)

41