Translation

Statistical Machine Translation

Martin Kay

Stanford University

with thanks to Kevin Knight

Elementary Probability

The probability of an event *e* occurring in a given *trial*. *A* number in the range 0 .. 1 giving the proportion of the trials in which *e* is expected to occur.

0 — Never

.5 — half the time

1 — *Always*

Elementary Probability

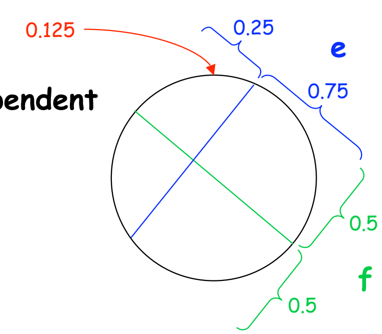
P(*e*) — *a priori* probability. The chance that *e* happens

P(*f* | *e*) — *conditional probability*. The probability of *f* happening in a situation where *e* happens.

P(e, f) — *joint probability*. The probability of *e* and *f* happening together.

Elementary Probability

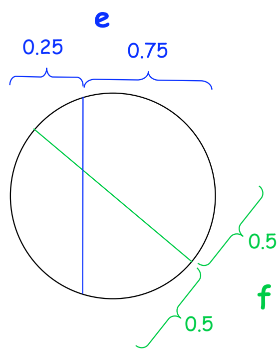
Independent



$$P(e \mid f) = P(f \mid e) = P(e) * P(f)$$

Elementary Probability

Dependent



P(e | f) P(f) = P(e, f) = P(f | e) P(e) = P(f, e) P(e)

Bayes' Rule $P(e, f) = \frac{P(f, e)\, P(e)}{P(f)}$

Bayes' Rule $P(e, f) = \frac{P(f, e)\, P(e)}{P(f)}$

Statistical Machine Translation

P(*e* | *f*) — The probability that *e* is an English translation of the given French sentence *f*.

$\underset{e}{\text{argmax}}\ P(e \mid f)$ The *e* that gives the maximum value of P(*e* | *f*)

$= \frac{P(f \mid e)\, P(e)}{P(f)}$

Does not effect the maximum

Statistical Machine Translation

P(*e* | *f*) — The probability that *e* is an English translation of the given French sentence *f*.

Translation model Language model

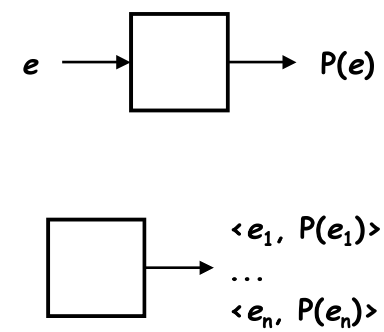
$= \frac{P(f \mid e)\, P(e)}{P(f)}$

The *Noisy Channel* Model

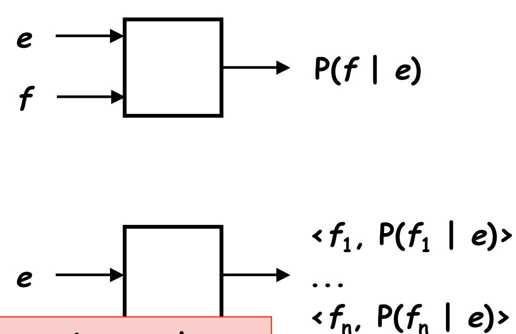
A person wants to say *e* but, by the time it comes out, it has been corrupted by noise to become *f*. To make our best guess as to what was intended we reason about

1. The things English speakers are likely to say, and
2. The statistics of the corruption process

P(*e*) as a program

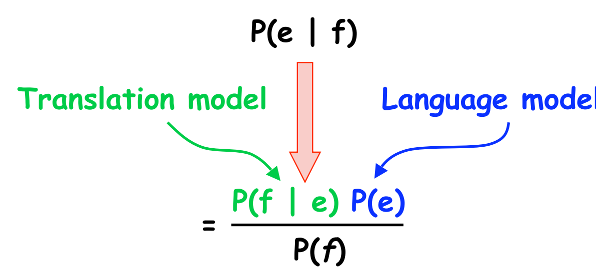


P(*f* | *e*) as a program



<u>Note</u>: Arrows point to the *right* because this is a theory of how *French* sentences are generated

Two models are better than one ...



... because they constrain one another, so neither has to take as much responsibility

For example

P(f | e): e and f contain words that are translations of one another in any order

P(e): gives a high value to e iff it is grammatical.

Language Models — 1

Find all the n English sentences on the web. If a sentence occurs k times, assign it a probability of k/n.

Problems

- Big data base
- Still does not contain many sentences

Language Models — 2

Use a probabilistic grammar

- Morphology
- Syntax
- ...

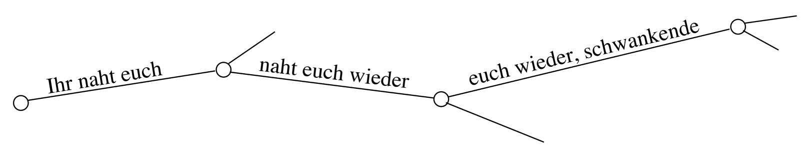
Language Models — 3

Use N-grams

Ihr naht euch wieder, schwankende Gestalten
Die früh sich einst dem trüben Blick gezeigt.
Versuch ich wohl, euch diesmal festzuhalten?
Fühl ich

Ihr naht euch
naht euch wieder
euch wieder, schwankende
wieder, schwankende Gestalten
schwankende Gestalten Die
früh sich einst
sich einst dem
...

Other models can be used in place of this



Smoothing

- What happens when N-grams appear that have never been seen before?
- Answer: <u>smoothing</u>
- Construct, say 3-gram, 2-gram and 1-gram (2nd-order, 1st-order, and 0-th order) models and take a certain proportion of the probability estimate from each.
- let $n_k(s)$ be the probability estimate of *s* in the *(k+1)*-st order model. Estimate probability of "abc" as $\lambda_3 n_3$("abc") + $\lambda_2 n_2$("bc") + $\lambda_1 n_1$("c") + λ_0 where $\lambda_3 + \lambda_2 + \lambda_1 + \lambda_0 = 1$.

How does English become French?

- IBM Model 3
- Replace English word by French words that appear opposite them in a bilingual dictionary and then scramble their order

Translation can change length

- Each English word e_i has a fertility ϕ_i which gives the sequence number of the French word that will be generated for it.
- Each French has a target position in its sentence which is a function of the position in the English sentence of the word it translates.

Translation as string rewriting

Hans ist nicht in dem Esszimmer gegangen

Assign fertilities

1 0 1 1 1 2 2

Apply fertilities

Hans nicht into dem Esszimmer Esszimmer gegangen gegangen

Translate words

Hans not in the dining room did go

Permute

Hans did not go into the dining room

Parameters

- t(not, nicht): The probablity that German *nicht* will become English *not*.
- n(5 | 2) The probability that the English for a German word in position 2 of the sentence will be placed in position 5.
- p_1 The probability of a adding a *spurious* word, Add a word NULL at the beginning of the source sentence that can give rise to new (spurious) words in the target. These can be inserted anywhere after the other words have been arranged.

The Model-3 procedure

1. For each English word e_i choose a fertility ϕ_i with probability $n(\phi_i \mid e_i)$.
2. Choose the number ϕ_o of NULL words to insert with probability $p1 + \Sigma\phi_i$.
3. Let m = sum of all fertilities (including ϕ_o)
4. For *i* in (1..*n*) and *k in* in (1..ϕ_i), choose a French word τ_{ik} with probability $t(\tau_{ik} \mid e_i)$.
5. For *i* in (1..ϕ_o) choose a French position π_{ik} for a NULL translation with a total probability $1/\phi_o$.
6. Arrange and output the sentence

Parameters

• n — fertilities • t — translations	2 dimensons
• d — position	1 dimension
• p — NULL translations	Scaler

Parameter Values

Parameter values could be estimated easily on the basis of an English text and its translation into French where corresponding sentences have been aligned. An alignment of a pair of word strings is simply a mapping of the words of one string onto the words of the other. It can be represented by a vector *A* where $A_i = j$ if the *i*-th English word is translated by the *j*-th French word. The frequncy of a translation, ordering, etc. is simply the number of alignments in which it is observed.

Alignments

Since alignments are not given, consider all *k* alignments for a given sentence pair, but add 1/*k* instead of 1 for the count for a given parameter.

Given values for the parameters, we can estimate probabilities of the aligments.

Given alignments, we can make better estimates of the parameters

The EM-algorithm

(Estimation Maximization)

Problems with Model 3

- Distortions are a lousy model of word order.
- Long sentences have to many alignments for training. Maybe use only good alignments but, how to find them (and only them).