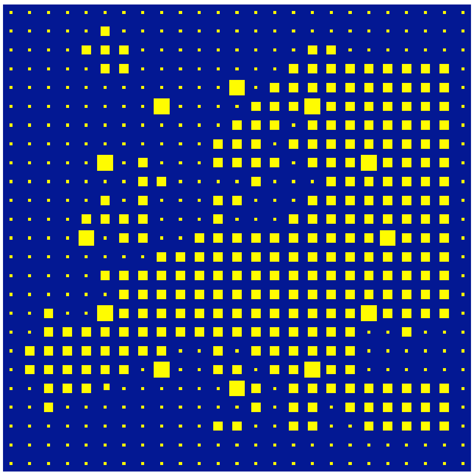
Intro to Statistical MT



EuroMatrix

MT Marathon
Chris Callison-Burch

Various approaches

- Word-for-word translation
- Syntactic transfer
- Interlingual approaches
- Controlled language
- Example-based translation
- Statistical translation

Advantages of SMT

- Data driven
- Language independent
- No need for staff of linguists of language experts
- Can prototype a new system quickly and at a very low cost

Statistical machine translation

- Find most probable English sentence given a foreign language sentence
- Automatically align words and phrases within sentence pairs in a parallel corpus
- Probabilities are determined automatically by training a statistical model using the parallel corpus

Parallel corpus

what is more , the relevant cost dynamic is completely under control.	im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle .
sooner or later we will have to be sufficiently progressive in terms of own resources as a basis for this fair tax system .	früher oder später müssen wir die notwendige progressivität der eigenmittel als grundlage dieses gerechten steuersystems zur sprache bringen .
we plan to submit the first accession partnership in the autumn of this year .	wir planen , die erste beitrittspartnerschaft im herbst dieses jahres vorzulegen .
it is a question of equality and solidarity .	hier geht es um gleichberechtigung und solidarität .
the recommendation for the year 1999 has been formulated at a time of favourable developments and optimistic prospects for the european economy .	die empfehlung für das jahr 1999 wurde vor dem hintergrund günstiger entwicklungen und einer für den kurs der europäischen wirtschaft positiven perspektive abgegeben .
that does not , however , detract from the deep appreciation which we have for this report .	im übrigen tut das unserer hohen wertschätzung für den vorliegenden bericht keinen abbruch .

Probabilities

- Find most probable English sentence given a foreign language sentence

$$p(e|f)$$

$$\hat{e} = \arg\max_e \; p(e|f)$$

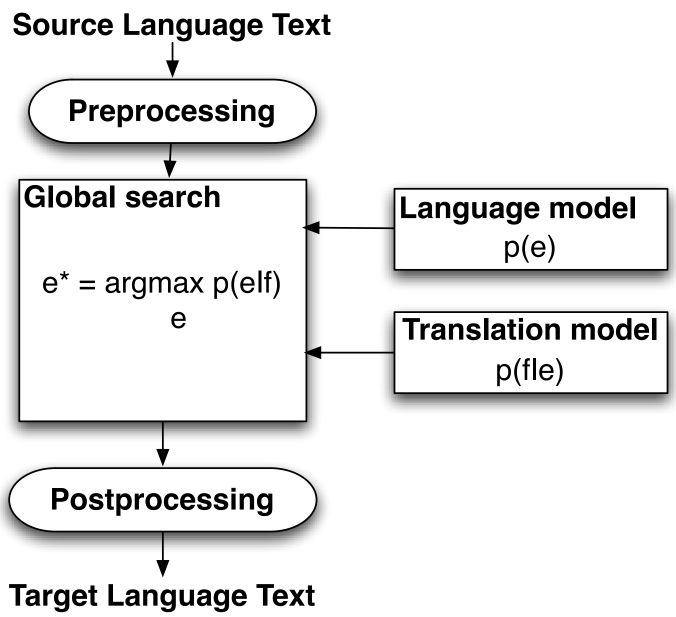
$$p(e|f) = \frac{p(e)p(f|e)}{p(f)}$$

$$\hat{e} = \arg\max_e \; p(e)p(f|e)$$

What the probabilities represent

- p(e) is the "Language model"
  - Assigns a higher probability to fluent / grammatical sentences
  - Estimated using monolingual corpora
- p(f|e) is the "Translation model"
  - Assigns higher probability to sentences that have corresponding meaning
  - Estimated using bilingual corpora

For people who don't like equations



Language Model

- Component that tries to ensure that words come in the right order
- Some notion of grammaticality
- Standardly calculated with a trigram language model, as in speech recognition
- Could be calculated with a statistical grammar such as a PCFG

Trigram language model

- p(I like bungee jumping off high bridges) =
  p(I | <s> <s>) *
  p(like | I <s>) *
  p(bungee | I like) *
  p(jumping | like bungee) *
  p(off | bungee jumping) *
  p(high | jumping off) *
  p(bridges | off high) *
  p(</s> | high bridges) *
  p(</s> | bridges </s>)

Calculating Language Model Probabilities

- Unigram probabilities

$$p(w_1) = \frac{count(w_1)}{total\ words\ observed}$$

Calculating Language Model Probabilities

- Bigram probabilities

$$p(w_2|w_1) = \frac{count(w_1 w_2)}{count(w_1)}$$

Calculating Language Model Probabilities

- Trigram probabilities

$$p(w_3|w_1 w_2) = \frac{count(w_1 w_2 w_3)}{count(w_1 w_2)}$$

Calculating Language Model Probabilities

- Can take this to increasingly long sequences of n-grams
- As we get longer sequences it's less likely that we'll have ever observed them

Backing off

- Sparse counts are a big problem
- If we haven't observed a sequence of words then the count = 0
- Because we're multiplying the n-gram probabilities to get the probability of a sentence the whole probability = 0

Backing off

$$
\begin{aligned}
&.8 * p(w_3|w_1 w_2) + \\
&.15 * p(w_3|w_2) + \\
&.049 * p(w_3) + \\
&.001
\end{aligned}
$$

- Avoids zero probs

Translation model

- p(f|e)... the probability of some foreign language string given a hypothesis English translation
- f = Ces gens ont grandi, vécu et oeuvré des dizaines d'années dans le domaine agricole.
- e = *Those people have grown up, lived and worked many years in a farming district.*
- e = *I like bungee jumping off high bridges.*

Translation model

- How do we assign values to p(f|e)?

$$p(f|e) = \frac{count(f, e)}{count(e)}$$

- Impossible because sentences are novel, so we'd never have enough data to find values for all sentences.

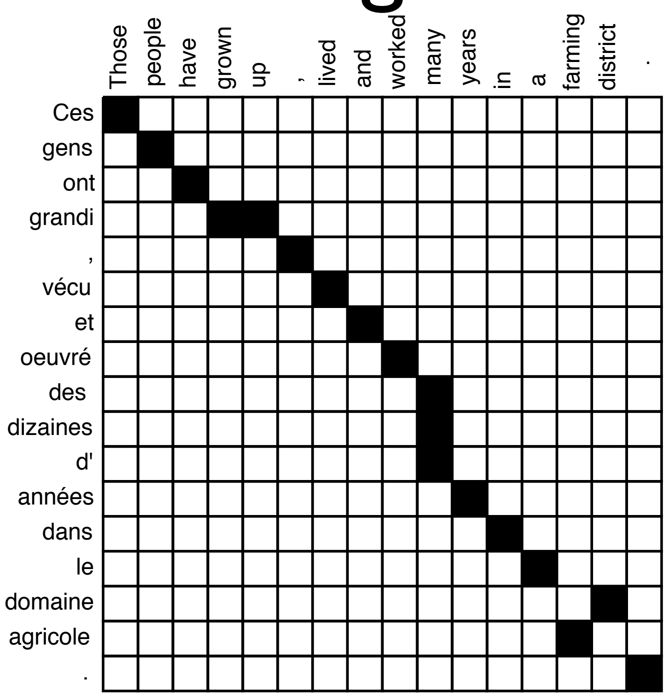
Translation model

- Decompose the sentences into smaller chunks, like in language modeling

$$p(f|e) = \sum_{a} p(a, f|e)$$

- Introduce another variable *a* that represents alignments between the individual words in the sentence pair

Word alignment



Alignment probabilities

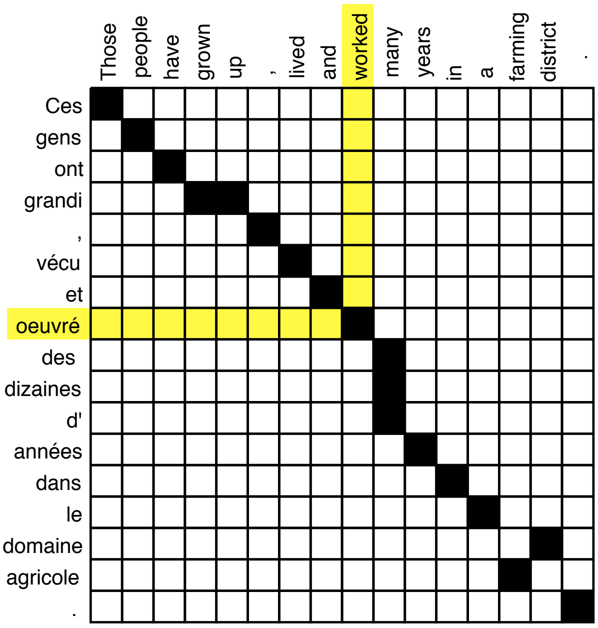
- So we can calculate translation probabilities by way of these alignment probabilities

$$p(f|e) = \sum_a p(a, f|e)$$

- Now we need to define p(a, f | e)

$$p(a, f|e) = \prod_{j=1}^{m} t(f_j|e_i)$$

Calculating $t(f_j|e_i)$



- Counting! I told you probabilities were easy!

$$= \frac{count(f_j, e_i)}{count(e_i)}$$

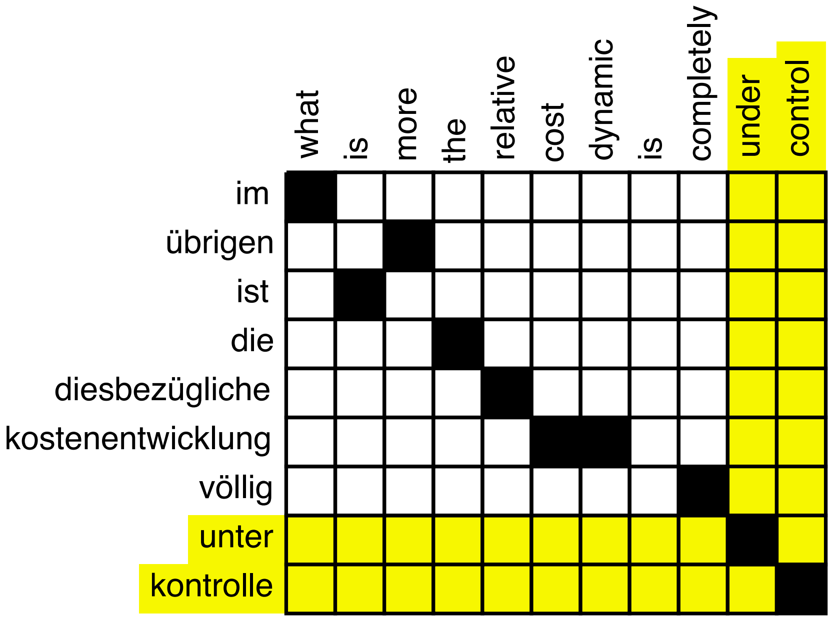
- **worked... fonctionné, travaillé, marché, oeuvré**

- 100 times total 13 with this f. 13%

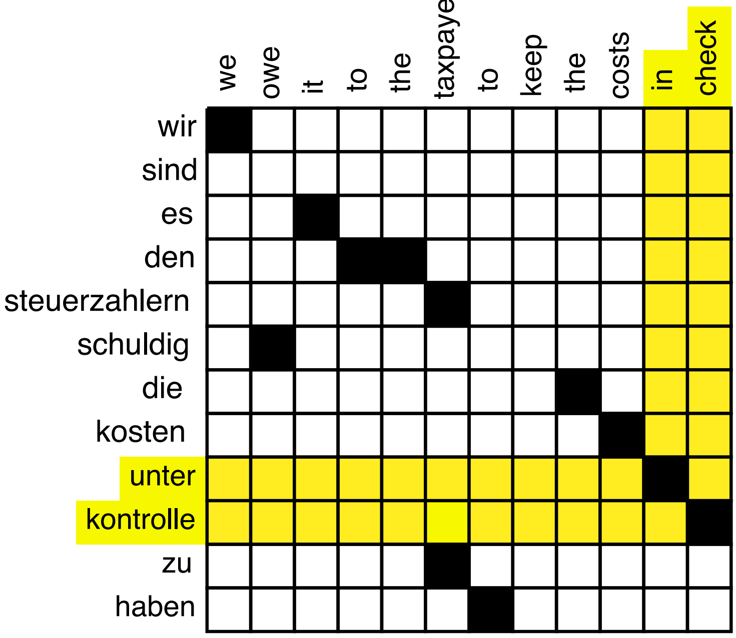
Calculating $t(f_j|e_i)$

- Unfortunately we don't have word aligned data, so we can't do this directly.
- OK, so it's not quite as easy as I said.
- There will be another lecture on how to do word alignments later in the week.

Phrase Translation Probabilities



Phrase Translation Probabilities

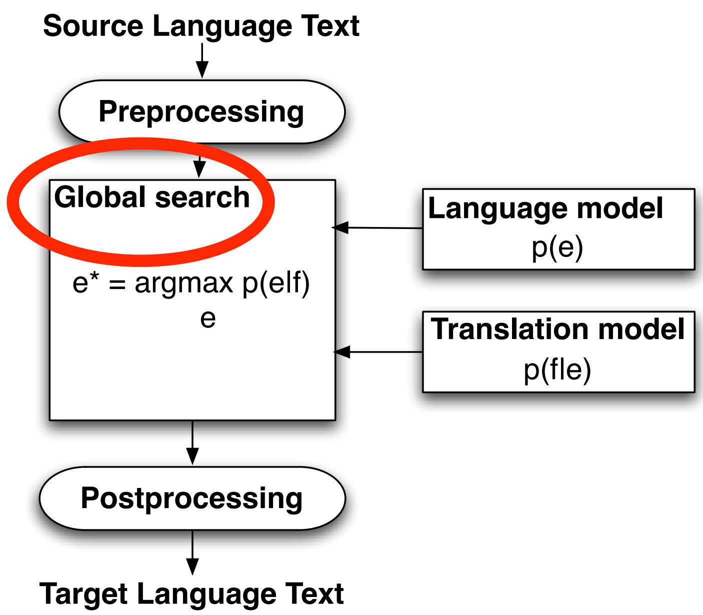


Phrase Table

- Exhaustive table of source language phrases paired with their possible translations into the target language, along with probabilities

das thema	the issue	.51
	the point	.38
	the subject	.21

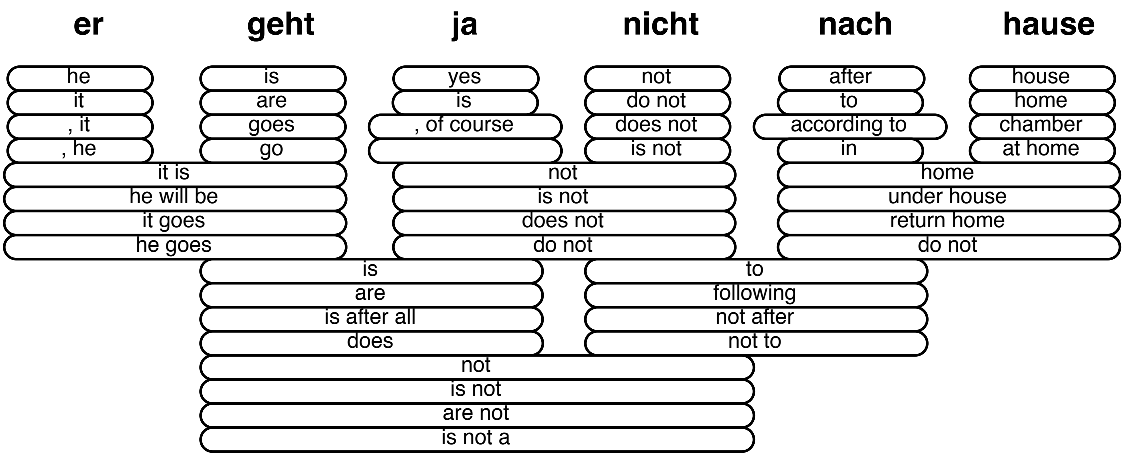
``Diagram Number 1"



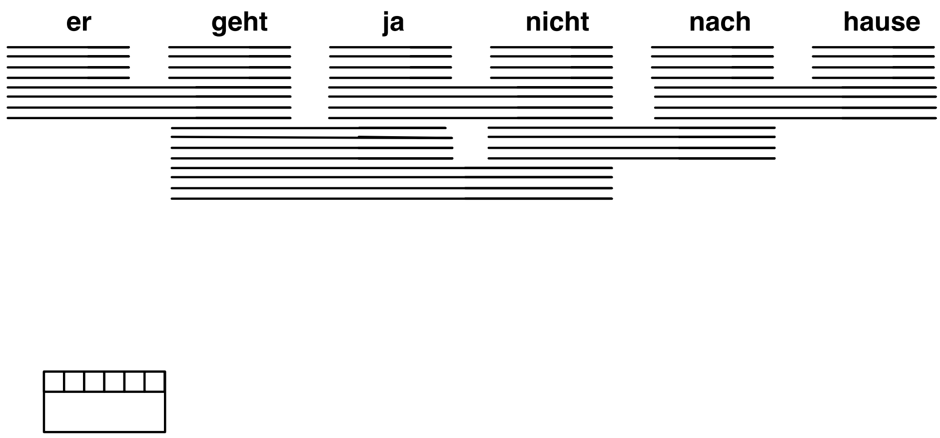
The Search Process AKA ``Decoding"

- Look up all translations of every source phrase, using the phrase table
- Recombine the target language phrases that maximizes the translation model probability * the language model probability
- This search over all possible combinations can get very large so we need to find ways of limiting the search space

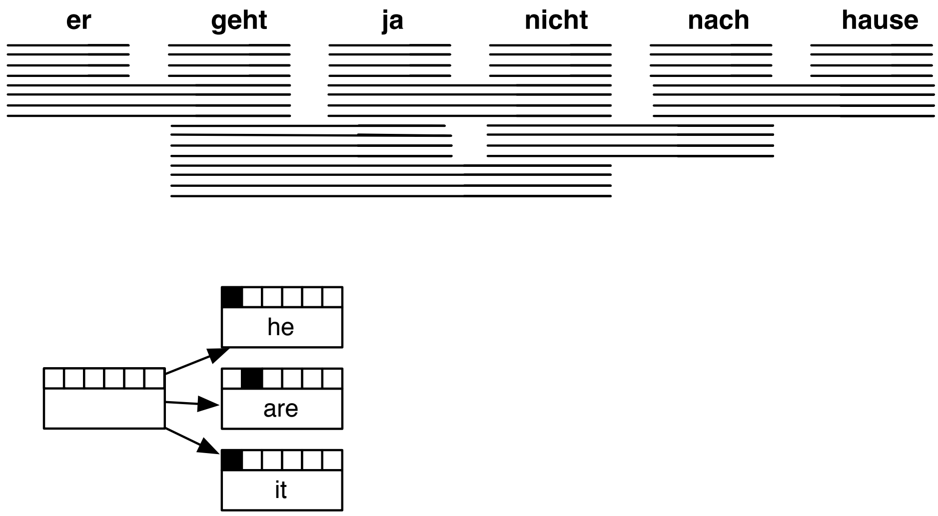
Translation Options



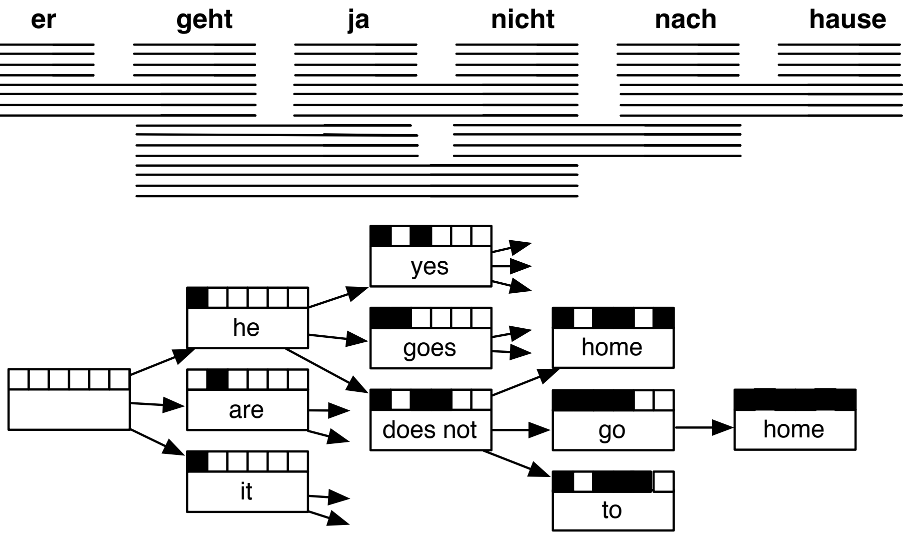
Search



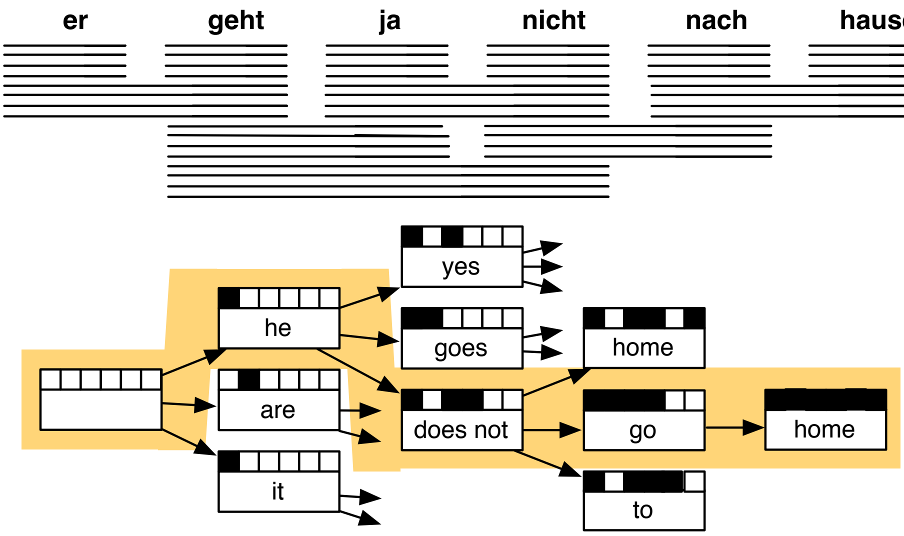
Search



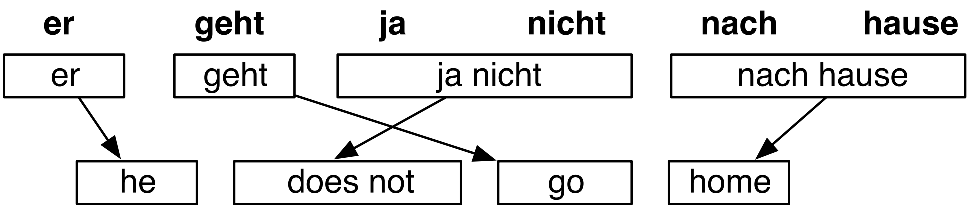
Search



Search



Best Translation



The Search Space

- In the end the item which covers all of the source words and which has the highest probability wins!
- That's our best translation

$$\hat{e} = \arg\max_{e} \; p(e)p(f|e)$$

- And there was much rejoicing!

Wrap-up: SMT is data driven

- Learns translations of words and phrases from parallel corpora
- Associate probabilities with translations empirically by counting co-occurrences in the data
- Estimates of probabilities get more accurate as size of the data increases

Wrap-up: SMT is language independent

- Can be applied to any language pairs that we have a parallel corpus for
- The only linguistic thing that we need to know is how to split into sentences, words
- Don't need linguists and language experts to hand craft rules because it's all derived from the data

Wrap-up: SMT is cheap and quick to produce

- Low overhead since we aren't employing anyone
- Computers do all the heavy lifting / statistical analysis of the data for us
- Can build a system in hours or days rather than months or years