Machine Learning in Statistical Machine Translation

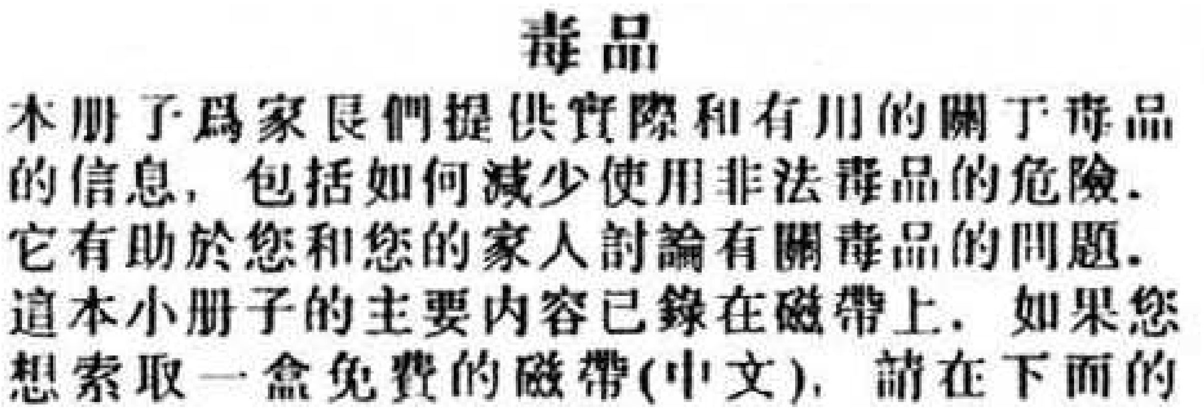
Phil Blunsom
Philipp Koehn

26 November 2008

School of informatics

Machine Translation

- Task: make sense of foreign text like



毒品
本册子為家長們提供實際和有用的關于毒品
的信息，包括如何減少使用非法毒品的危險.
它有助於您和您的家人討論有關毒品的問題.
這本小册子的主要内容已錄在磁帶上. 如果您
想索取一盒免費的磁帶(中文), 請在下面的

- AI-hard: ultimately reasoning and world knowledge required

- Statistical machine translation: Learn how to translate from data

Prediction Problem

- Given an input sentence, we have to predict an output translation

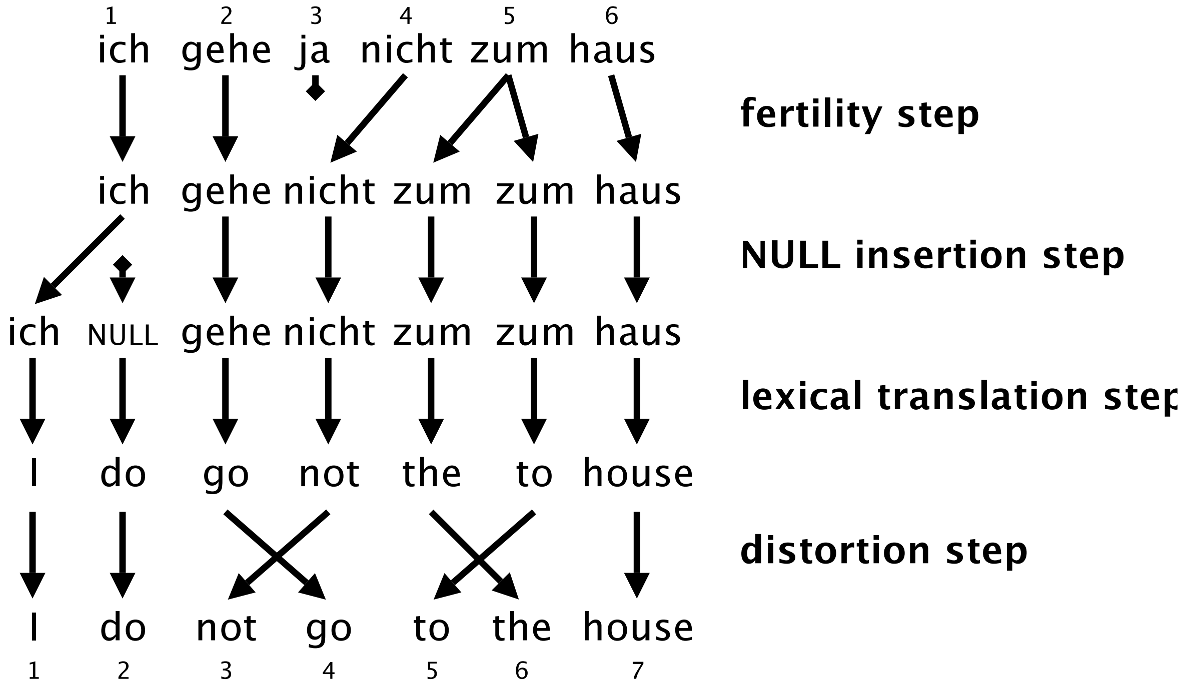
Ich gehe ja nicht zum Haus.

$\Downarrow$

I do not go to the house.

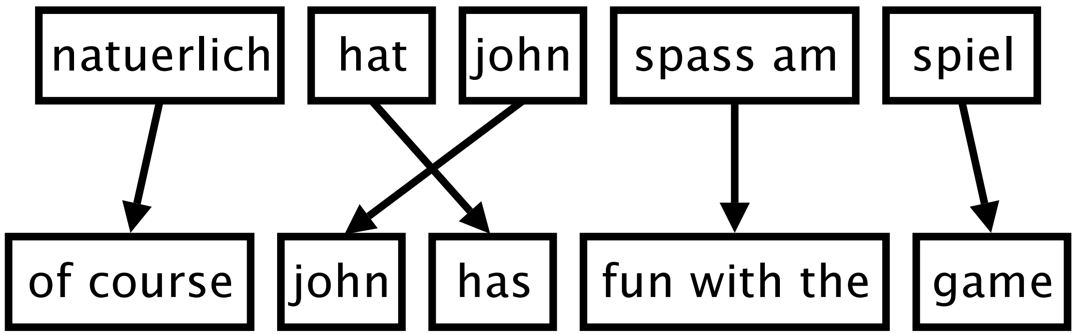
- Since the set of possible output sentences is too large, we need to construct the translation according to some decomposition of the translation process

Word-Based Model



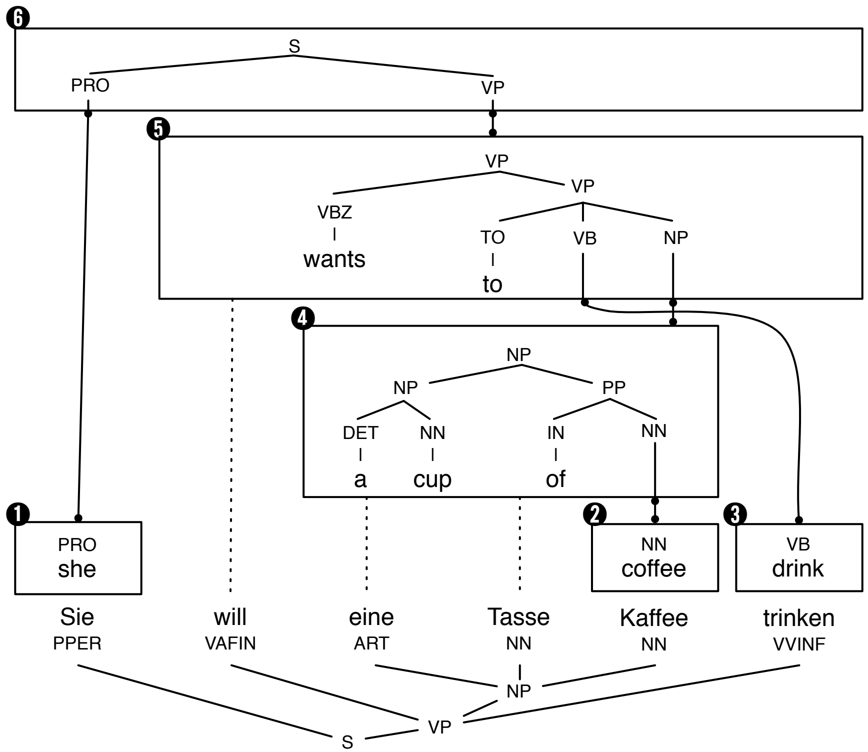
Original statistical machine translation models (1990s):
break down translation to the word level

Phrase-Based Model



Current state of the art:
map larger chunks of words (huge mapping tables)

Tree-Based Model



One way forward: generate translation with syntactic structure

Structured Prediction

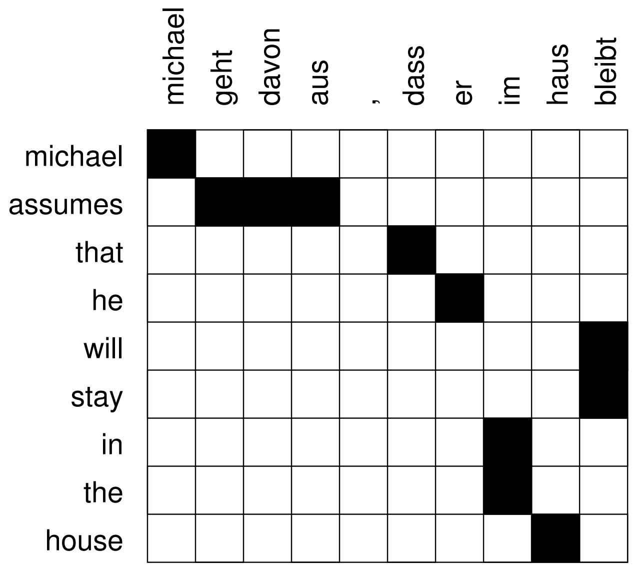
- A prediction problem
  - given an input
  - predict an output
  - many example (input, output) pairs available

- But: space of possible outputs too large
  - prediction has to be broken down into steps
  - decomposition of the problem is a hidden variable
  - search space too large to explore exhaustively

- Additional trouble
  - there is not a *single* right translation, many are possible
  - evaluation of machine translation unclear

Learning Problem: Word Alignment

- For many models, an essential first step is establishing the word alignment in the training data



- Very little labeled data available
  → typically treated as unsupervised learning problem

Learning Problem: Model Parameters

- The output translation from an input sentence is derived over several steps
  - segmentation of the input
  - word and phrase translation
  - reordering
- Each of the steps is modeled by probability distributions or features
- How do we learn the parameters for these models?

Heuristic Generative Model

- The decomposition of the translation process breaks down into steps
- Each step is modeled with a probability distribution
- Phrase translation probability distributions are estimated by maximum likelihood estimation:

$$p(\text{house}|\text{Haus}) = \frac{\text{count}(\text{house},\text{Haus})}{\text{count}(\text{Haus})}$$

- This is a biased ML estimator, we'd like to replace it:
  - Bayesian approach [Blunsom, Cohn and Osborne, 2008]

Discriminatively Combining Local Models

- Sentence translation is a combination of several component models

$$p_{LM} \times p_{TM} \times p_D$$

- These may be weighted

$$p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D}$$

- Many components p_i with weights λ_i

$$\prod_i p_i^{\lambda_i} = \exp \sum_i \lambda_i \log(p_i)$$

- Optimizing the weights λ_i to directly optimize translation performance

Global Discriminative Model

- Where we are now: a unsatisfying mix of local models and global models
- Grand goal: train all parameters discriminatively to optimize translation
- Note:
  - hidden derivation
  - millions of sentence pairs
  - millions of features

  $\rightarrow$ heavy computational problem
- Ongoing work
  - Perceptron, MIRA [Arun and Koehn, 2007]
  - probabilistic model [Blunsom and Osborne, 2008]

Deluge of Data

- Parallel texts: 100s millions of words

$\longrightarrow$ translation models take up giga-bytes on disk

- Monolingual texts: trillions of words

$\longrightarrow$ much more than we can currently handle

- Need for efficient data structures and training methods
  - suffix arrays for on-the-fly translation model [Lopez et al., 2008]
  - randomized language models [Talbot and Osborne, 2008]

Related Task: Tools for Translators

<< [4] Hamburg - Sechs Stunden sprachen sie miteinander. >>

Hamburg -

enter six

Hamburg	-	Sechs	Stunden	sprachen	sie	miteinander	.
Hamburg	-	six	hours	it		with each other .	
Hamburg ,	,	Six	hours ,	it would be		.	
of Hamburg ,	-	six ,	hours of	it would		to each other .	
Hamburg is	Mr President ,	then	few hours	they were speaking		together	.
of Hamburg	Mr President , ladies and gentlemen ,	6	a few hours	spoke		with each other	
Hamburg market	.	six members	working hours	you spoke		with	
Hamburg accounts	- the	six governments	hours ago	were	they		
	- and	concurred	time	have	they are		
	as	all concurred	hour	there were	it	each other	
	- to	six leaders	hours in	talked	they	work with each other .	

Learning task: predicting the next user input

Machine Translaton at Edinburgh

- People
  - 2 faculty: Philipp Koehn and Miles Osborne
  - 3 postdocs, 1 research programmer, 7 PhD students
- Funding
  - European projects: EuroMatrix, EuroMatrixPlus
  - DARPA project: GALE
  - EPSRC project: Demeter
  - Industry: Google, Systran
- Resources for the community
  - our open source Moses decoder is standard benchmark for MT community
  - we organize MT evaluation campaigns, open source conventions, workshops
- Online demo: `http://demo.statmt.org/webtrans/`