UPPSALA UNIVERSITET

Scalability in Statistical Machine Translation Research

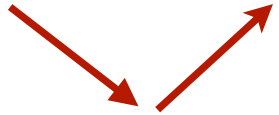
Jörg Tiedemann
Department of Linguistics and Philology
Uppsala University

UPPSALA UNIVERSITET

Computational Linguistics and NLP

Understand Language
- find linguistic building blocks
- structure, patterns, preferences
- differences and universals

Language Understanding
- real-world applications
- human-computer interaction
- speech, text, multimedia



Computational Models

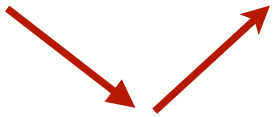
UPPSALA UNIVERSITET

Computational Linguistics and NLP

Understand Language
- find linguisti
- structure, p
- differences and universals

Language Understanding
- cations
- r interaction
- speech, text, multimedia

Human Experts



Computational Models

UPPSALA UNIVERSITET

Computational Linguistics and NLP

Understand Language
- find linguisti
- structure, p
- differences and univ

Language Understanding
- cations
- r interaction
- h, text, multimedia

Human Experts

Big Data

Computational Models

UPPSALA UNIVERSITET

Computational Linguistics and NLP

Understand Language — Language Understanding

- find linguisti… …cations
- structure, p… …er interaction
- differences and univ… …h, text, multimedia

Human Experts

Big Data

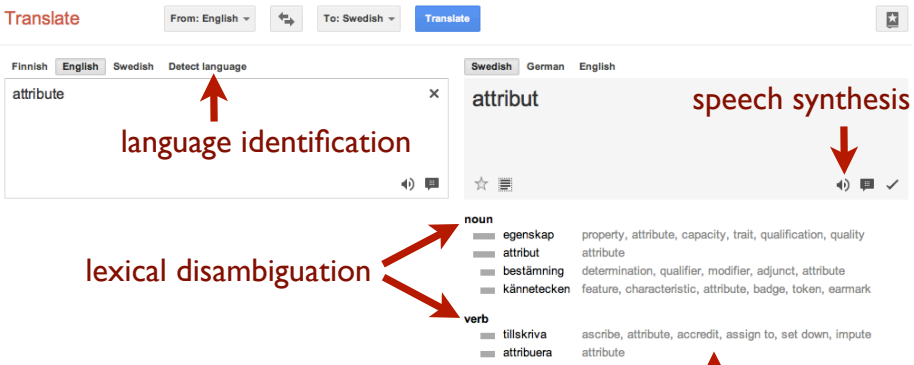
Machine Learning

Computational Models

UPPSALA UNIVERSITET

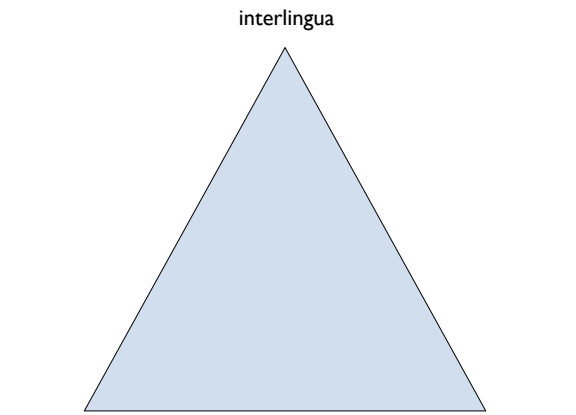
NLP and Real-World Applications

machine translation (MT)



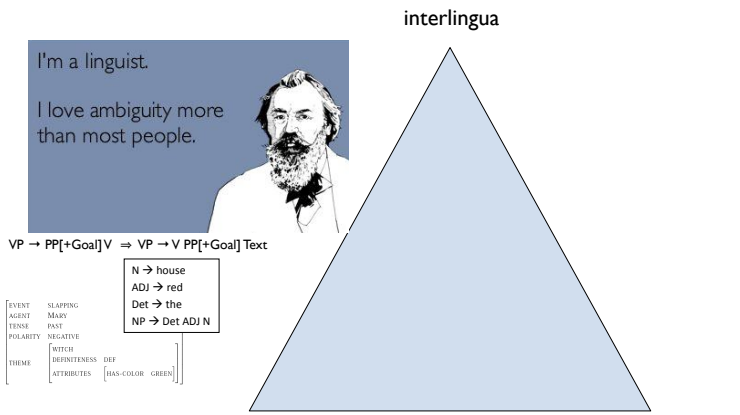
UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT



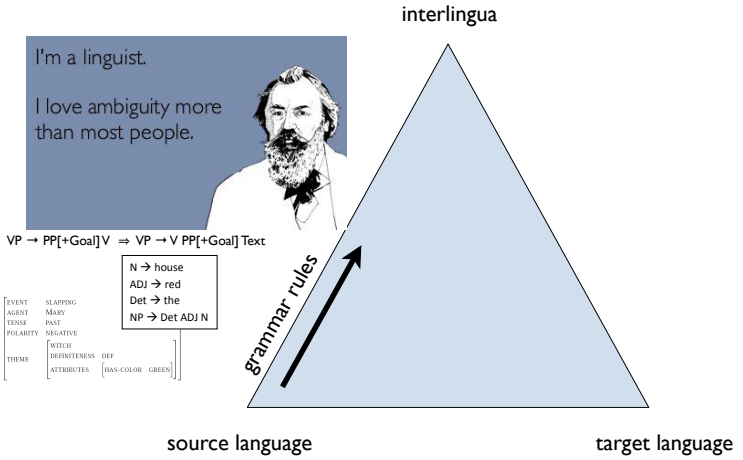
UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT



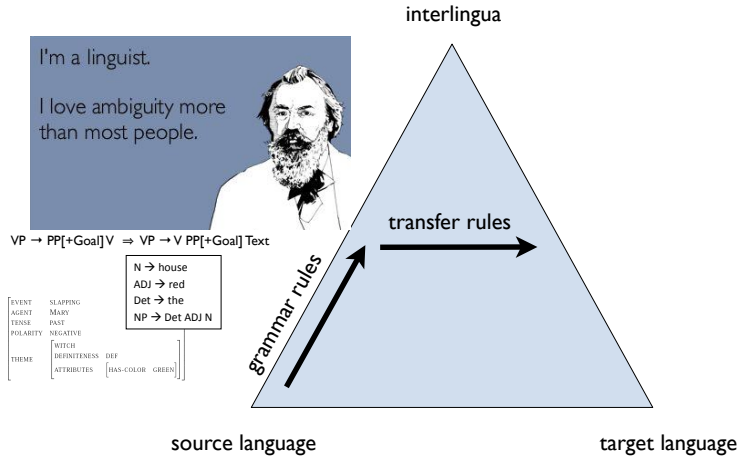
UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT



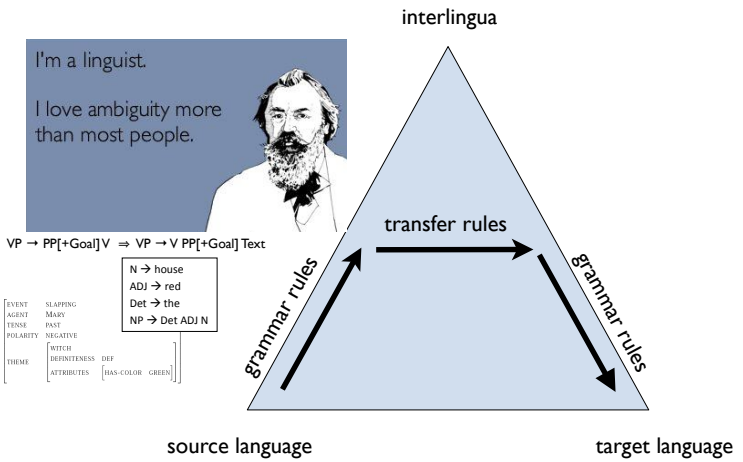
UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT



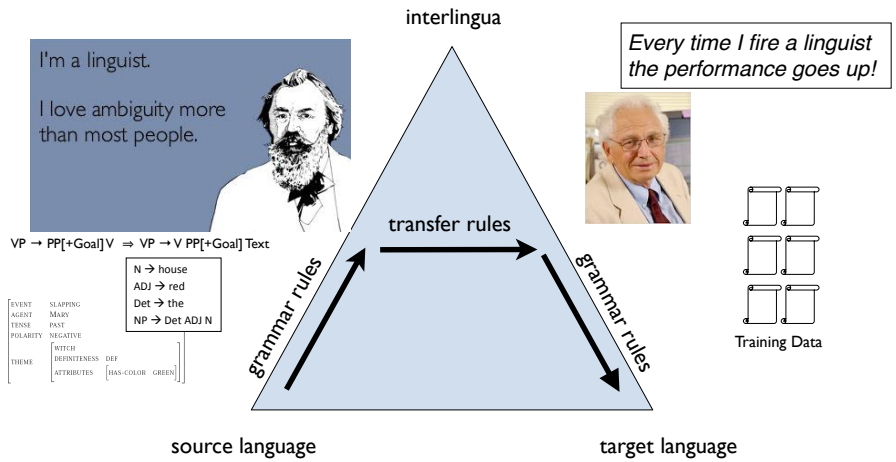
UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT



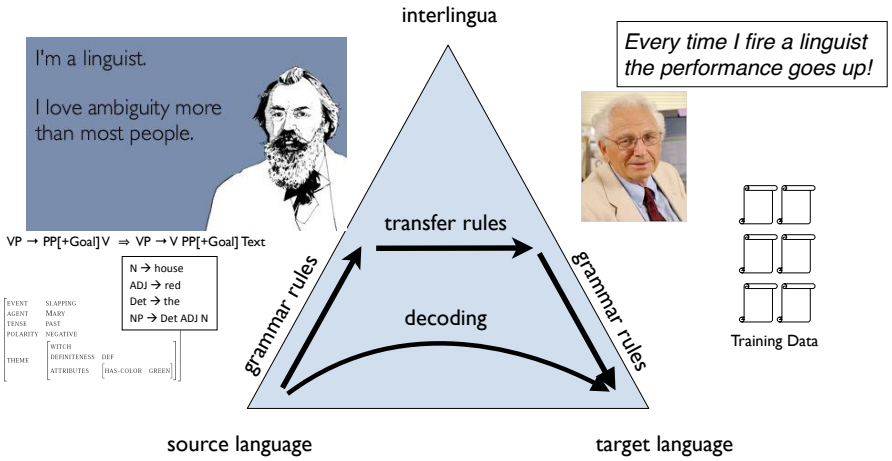
UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT



UPPSALA UNIVERSITET

Expert-Driven vs. Data-Driven MT

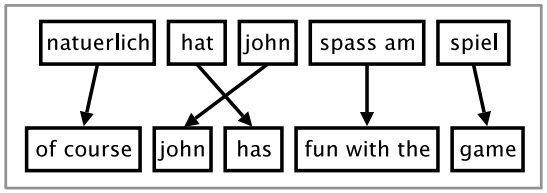


UPPSALA UNIVERSITET

Phrase-Based Statistical MT in a Nutshell

UPPSALA UNIVERSITET

Phrase-Based Statistical MT in a Nutshell

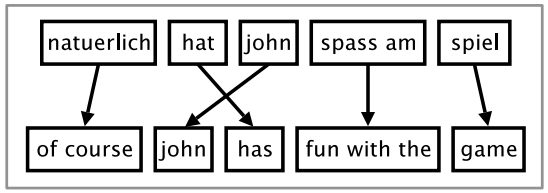


Translation Model

Reordering Model

UPPSALA UNIVERSITET

Phrase-Based Statistical MT in a Nutshell



Translation Model

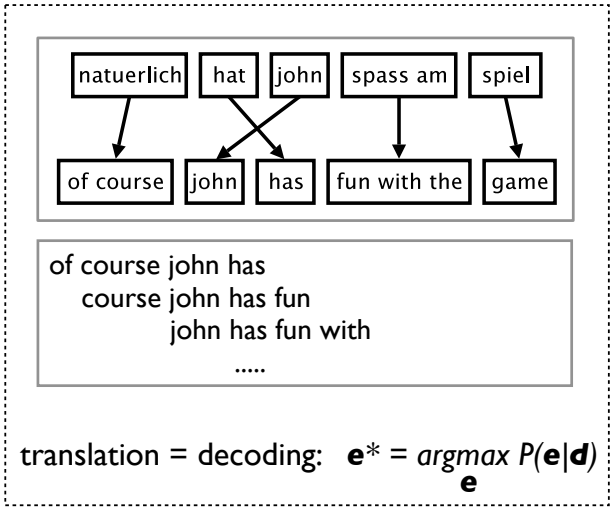
Reordering Model

of course john has
course john has fun
john has fun with
.....

Target Language Model

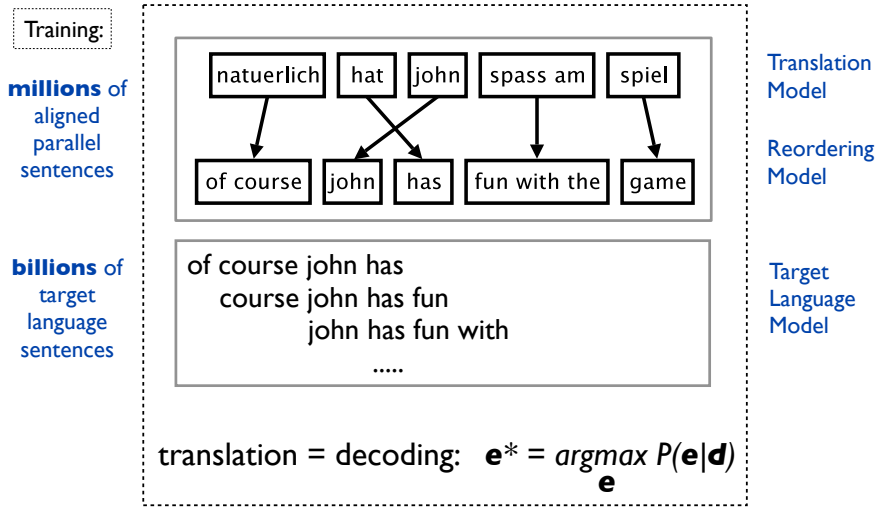
UPPSALA UNIVERSITET

Phrase-Based Statistical MT in a Nutshell



UPPSALA UNIVERSITET

Phrase-Based Statistical MT in a Nutshell



UPPSALA UNIVERSITET

Why High-Performance Computing?

UPPSALA UNIVERSITET

Why High-Performance Computing?

Training statistical models

- natural languages are highly ambiguous and productive
- billions of model parameters
- complex numeric optimization problems
- growing data sets
- many languages and textual domains

UPPSALA UNIVERSITET

Why High-Performance Computing?

Training statistical models

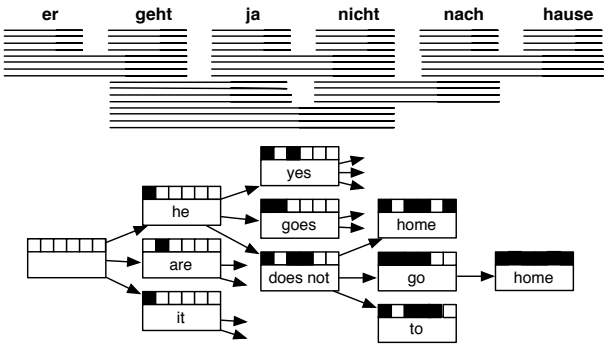
- natural languages are highly ambiguous and productive
- billions of model parameters
- complex numeric optimization problems
- growing data sets
- many languages and textual domains

Translation as decoding

- ... is a gigantic search problem

UPPSALA UNIVERSITET

Translation as Search

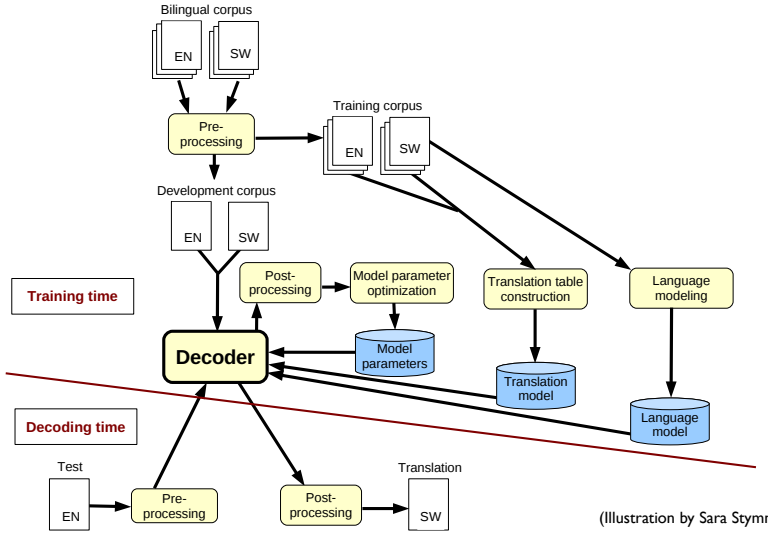


- $O(2^N)$ phrase segmentations,
- $O(T^N)$ sets of phrase translations, and
- $O(N!)$ word reordering permutations.

(Illustration by Philipp Koehn)

UPPSALA UNIVERSITET

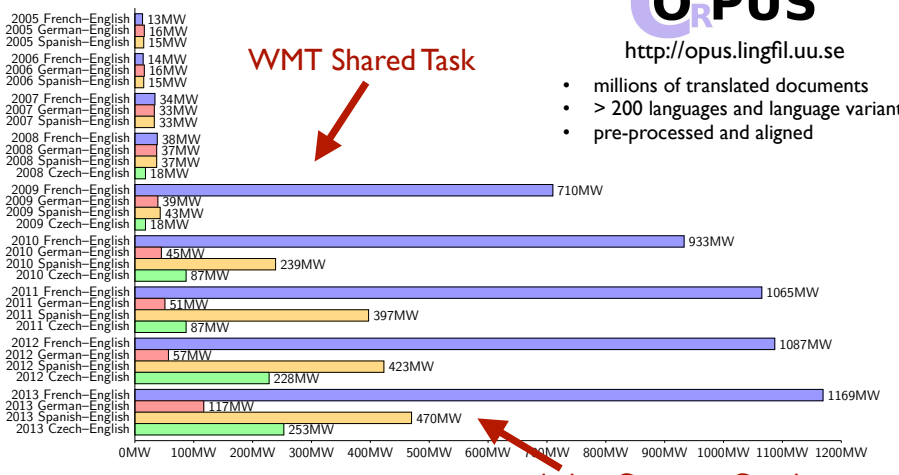
Typical Architecture in Statistical MT



(Illustration by Sara Stymne)

UPPSALA UNIVERSITET

Growing Data Sets



OPUS

http://opus.lingfil.uu.se

- millions of translated documents
- > 200 languages and language variants
- pre-processed and aligned

(Source: Philipp Koehn)

UPPSALA UNIVERSITET

Model Size and Training Time

A English-French baseline model for WMT has

- a 60-90 GB binarized phrase table
- a 60 GB binarized language model

UPPSALA UNIVERSITET

Model Size and Training Time

A English-French baseline model for WMT has

- a 60-90 GB binarized phrase table
- a 60 GB binarized language model

Training and tuning takes time and memory

- several days to word-align and to extract phrase tables
- several days to tune feature weights
- a lot of temporary disk space (TB) with fast I/O

UPPSALA UNIVERSITET

Model Size and Training Time

A English-French baseline model for WMT has

- a 60-90 GB binarized phrase table
- a 60 GB binarized language model

Training and tuning takes time and memory

- several days to word-align and to extract phrase tables
- several days to tune feature weights
- a lot of temporary disk space (TB) with fast I/O

Research in MT

- re-train & re-tune several models with various features
- new feature functions, new domains, new test data

UPPSALA UNIVERSITET

Scaling Up

Good News: Parallelization

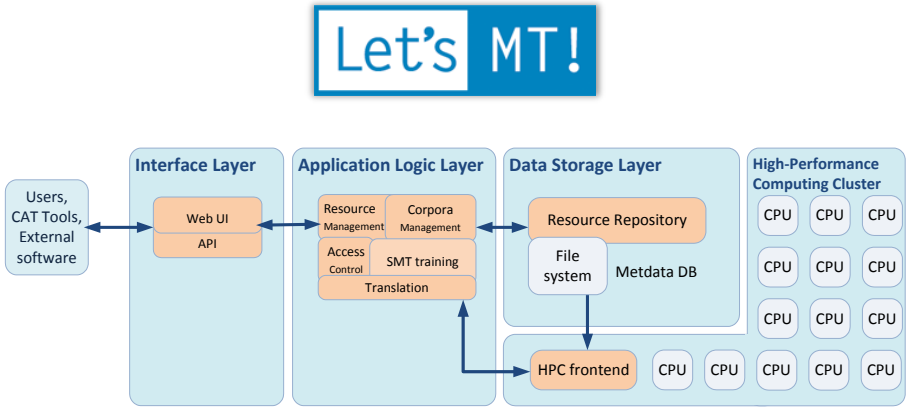
- data pre-processing
- alignment (paragraphs, sentences, words)
- phrase and rule extraction / scoring
- parameter tuning (requires decoding)
- translation

HPC infrastructure is very useful!

UppsalaMT@HPC: > 500,000 h used in 2013/14

UPPSALA UNIVERSITET

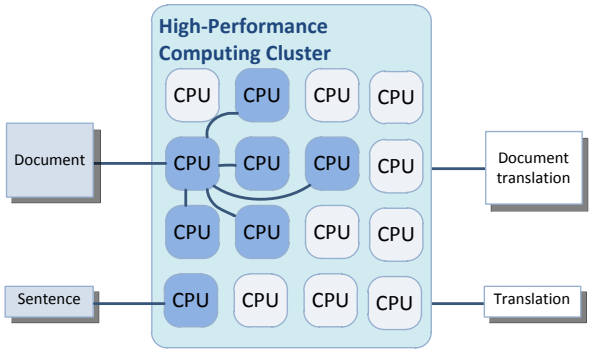
A Collaborative MT Platform: Let'sMT!



UPPSALA UNIVERSITET

Let's MT!

Distributed Document Translation



UPPSALA UNIVERSITET

Scaling Up Limitations

Example: Huge Language Models (Edinburgh WMT'14)

English LM:
- size: 3.5 TB
- required filtering!
- tuning: 1 TB RAM

Lang	Lines (B)	Tokens (B)	Bytes
en	59.13	975.63	5.14 TiB
de	3.87	51.93	317.46 GiB
fr	3.04	49.31	273.96 GiB
ru	1.79	21.41	220.62 GiB
cs	0.47	5.79	34.67 GiB
hi	0.01	0.28	3.39 GiB

2012/2013 CommonCrawl

Not all language pairs could be run!

Pair	Baseline		+ Huge LM's	
newstest	2013	2014	2013	2014
en-de	20.85	20.10	–	20.61 +0.51
en-cs	19.39	21.00	20.03 +0.64	21.60 +0.60
en-ru	19.90	28.70	20.80 +0.90	29.90 +1.20
en-hi	11.43	11.10	12.83 +1.40	12.50 +1.40
hi-en	15.48	13.90	–	14.80 +0.90

BLEU score gains

UPPSALA UNIVERSITET

Conclusions

Real-World Applications use SMT

Data-Driven MT requires Big Data

SMT needs a lot of Computing Power