EuroMatrix

Machine Translation for all European Languages

Philipp Koehn, University of Edinburgh

13 September 2007

The European Challenge

Many languages

- 11 official languages in EU-15
- 23 official languages in EU-27
- many more minority languages

Challenge

- European reports, meetings, laws, etc.
- develop technology to **enable use of local languages** as much as possible

Existing MT systems for EU languages

[from Hutchins, 2005]

	Cze	Dan	Dut	Eng	Est	Fin	Fre	Ger	Gre	Hun	Ita	Lat	Lit	Mal	Pol	Por	Slo	Slo	Spa	Swe	
Czech	–	.	.	1	.	.	1	1	.	.	1	.	.	.	.	.	.	.	.	.	4
Danish	.	–	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	1
Dutch	.	.	–	6	.	.	2	1	.	.	.	.	.	.	.	.	.	.	.	.	9
English	2	.	6	–	.	.	42	48	3	3	29	1	.	.	7	30	2	.	48	1	222
Estonian	.	.	.	.	–	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0
Finnish	.	.	.	2	.	–	.	1	.	.	.	.	.	.	.	.	.	.	.	.	3
French	1	.	2	38	.	.	–	22	3	.	9	.	.	.	1	5	.	.	10	.	91
German	1	1	1	49	.	1	23	–	.	1	8	.	.	.	4	3	2	.	8	1	103
Greek	.	.	.	2	.	.	3	.	–	.	.	.	.	.	.	.	.	.	.	.	5
Hungarian	.	.	.	1	.	.	.	1	.	–	.	.	.	.	.	.	.	.	.	.	2
Italian	1	.	.	25	.	.	9	8	.	.	–	.	.	.	1	3	.	.	7	.	54
Latvian	.	.	.	1	.	.	.	.	.	.	.	–	.	.	.	.	.	.	.	.	1
Lithuanian	.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	.	.	.	.	.	0
Maltese	.	.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	.	.	.	.	0
Polish	.	.	.	6	.	.	1	3	.	.	1	.	.	.	–	2	.	.	1	.	14
Portuguese	.	.	.	25	.	.	4	4	.	.	3	.	.	.	1	–	.	.	6	.	43
Slovak	.	.	.	1	.	.	.	1	.	.	.	.	.	.	.	.	–	.	.	.	2
Slovene	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	0
Spanish	1	.	.	42	.	.	8	7	.	.	7	.	.	.	1	6	.	.	–	.	72
Swedish	.	.	.	2	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	–	3
	6	1	9	201	0	1	93	99	6	4	58	1	0	0	15	49	4	0	80	2	

Goals of the EuroMatrix Project

- Machine translation between **all EU language pairs**
  - baseline machine translation performance for all pairs

  → starting point for national research efforts
  - more intensive effort on specific language pairs
- Creating an **open research** environment
  - open source **tools** for baseline machine translation system
  - collection of open data **resources**
  - open **evaluation campaigns** and **research workshops** ("marathons")
- Scientific **approaches**
  - **statistical** phrase-based, extended by factored approach
  - **hybrid** statistical/rule-based
  - tree-transfer based on **tecto-grammatic** probabilistic models

EuroMatrix **Project**

- **Participants**
  - University of Saarbrücken (coordinator)
  - University of Edinburgh (scientific lead)
  - Charles University, Prague
  - CELCT, Italy
  - GROUP Technologies, Germany
  - Morphologic, Hungary
- Time: 30 months, September 2006–February 2009

EuroMatrix **Events**

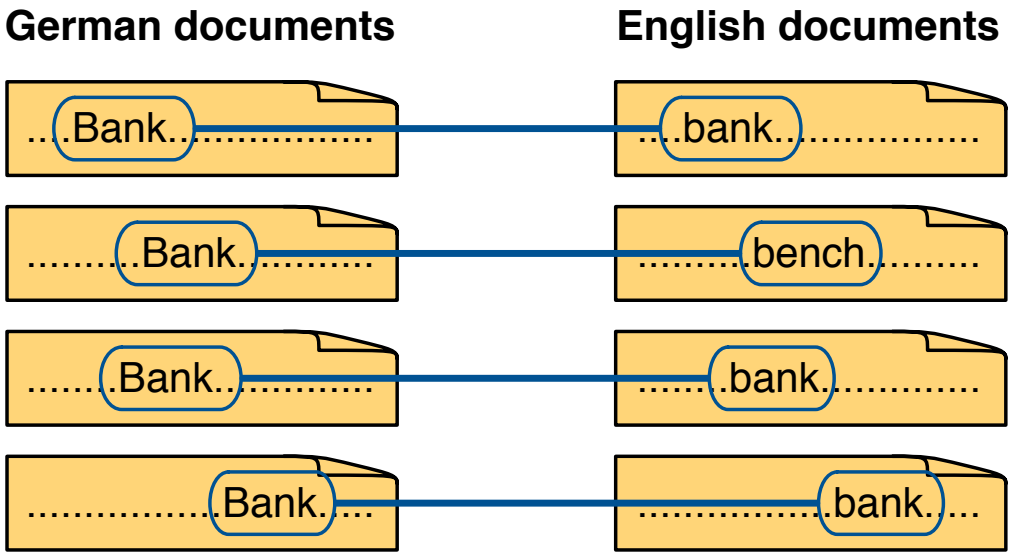
- Evaluation Campaign
  - dry-run March 2007, meeting at ACL 2007
  - first campaign February 2008, meeting at ACL 2008(?)
  - second campaign early 2009
- Machine Translation Marathon
  - April 2007 in Edinburgh: summer school and research showcase
  - May 2008 in Berlin
    - summer school
    - open source conference
    - research showcase
    - early analysis of evaluation campaign results

Progress in machine translation

- **Data revolution**
  - **immense text resources** available in digital form (trillion of words)
  - large ammounts of **translated text** become increasingly available (today 10-100s millions of words, maybe soon billions)
- **Statistical machine translation (SMT)**
  - development of **data-driven** statistical approach to MT
  - **competitive** with traditional approaches
- **Favorable research environment**
  - **US DARPA funding** for Arabic–English and Chinese–English SMT
  - **open** competitions, **sharing** of resources
  - however: limited academic research in **European languages**

Statistical Machine Translation

- Learning from data (sentence-aligned translated texts)



$\Rightarrow$ p(bank|Bank) = 0.75, p(bench|Bank) = 0.25

- New machine translation systems can be built **automatically**

Translating between all EU-15 languages

- Statistical methods allow the rapid development of MT systems
- BLEU scores for 110 statistical machine translation systems

	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

[from Koehn, 2005]

Good translation quality

French

Nous savons très bien que les Traitès actuels ne suffisent pas et qu'il sera nécessaire à l'avenir de développer une structure plus efficace et différente pour l'Union, une structure plus constitutionnelle qui indique clairement quelles sont les compétences des États membres et quelles sont les compétences de l'Union.

French–English MT

We know very well that the current Treaties are not enough and that in the future it will be necessary to develop a different and more effective structure for the Union, a constitutional structure which clearly indicates what are the responsibilities of the Member States and what are the competences of the Union.

More examples: `http://www.statmt.org/matrix/`

Commitment to open resources

- A lot of **infrastructure** required to build a statistical MT system
  - parallel corpora
  - word alignment
  - language modeling
  - basic linguistic tools (tokenizers, taggers, morph. analyzers, parsers)
  - training statistical models
  - decoding

→ MT systems become **too large** to be built completely by a small group

- **Sharing** of resources
  - avoids rebuilding the wheel everywhere
  - allows everybody to work on the state of the art
  - focus on novel solutions to current problems

Moses: Open Source Toolkit

- **Open source** statistical machine translation system (developed from scratch 2006)
  - state-of-the-art **phrase-based** approach
  - novel methods: **factored translation models**, **confusion network decoding**
  - support for **very large models** through **memory-efficient** data structures

- Documentation, source code, binaries **available at** `http://www.statmt.org/moses/`
- Development also **supported by**
  - EC-funded **TC-STAR** project
  - **US** funding agencies DARPA, NSF
  - universities (Edinburgh, Maryland, MIT, ITC-irst, RWTH Aachen, ...)

Parallel Corpora

- **Europarl** corpus: proceedings of the European Parliament
  - Release of v3 imminent
  - 30-40 million word per language, all 11 official languages of the EU-15
- **News Commentary**: from `http://www.project-syndicate.com/`
  - Used in ACL WMT 2007 Shared Task
  - 1-2 million words in English, French, Spanish, German, Czech, Arabic, ...
- **Other** corpus projects
  - Acquis Communitaire: includes all 23 languages of EU-25
  - more data from European Union / European Commission?
  - patent translation data?

→ good translation quality possible with this data

Open evaluation

- Website hosted by the EuroMatrix project
- Machine translation for **all EU-25** languages
  - extending the matrix of MT systems to 23x22=**506 language pairs**
  - information about **resources** for each language pair
  - **example output** to demonstrate translation performance
- **Ongoing evaluation** of translation quality
  - test sets from the last annual competition
  - **anybody can upload** their system's translations
  - **automatic** scoring with BLEU, NIST, METEOR, and other metrics
  - facilitate **manual** evaluations (also a good **teaching tool**)
- Will go online this Fall

Research efforts in EuroMatrix

- **Statistical** phrase-based, extended by factored approach
  - builds on state-of-the art phrase-based approach
  - idea: add additional annotation at the word level (POS, morphology, ...)
  - effort centered at the University of Edinburgh
- **Hybrid** statistical/rule-based
  - integration of the Logos system with statistical methods
  - system combination / deep integration of components
  - effort centered at the University of Saarland, Saarbrücken
- Tree-transfer based on **tecto-grammatic** probabilistic models
  - based on long-term efforts, builds on parallel Czech–English treebank
  - transfer at the level of enriched dependency structures
  - effort centered at Charles University, Prague

Why is Machine Translation Hard?

- Languages **differ** in
  - lexical items
  - syntactic structure
  - morphology
  - word order
  - concepts, especially connotations
  - metaphorical use
  - degree of redundancy
  - millions of exceptions
- What has **changed** with the advent of statistical MT?

What is Not Hard Anymore

- **Knowledge acquisition**
  - **long tail** of rare words, special uses, exceptions, ...

  → learn from data
- **Word sense disambiguation**
  - a word in one language has many translations into another
  - ambiguity is often **resolved in local context** (language models)
  - ambiguity may also resolved by establishing **domain**

  → all this fits nicely into the statistical framework

What is Still Hard

- **Morphology**
  - some languages express a lot with morphology

  $\longrightarrow$ large vocabularies, sparser data
  - generative: **new words** may be made up
  - especially **generating rich morphology** is hard
- **Word order**
  - SVO, SOV, VSO, free word order
  - especially **long-range movement** is difficult when relying only on LM
- **Different means of syntactic representation**
  - German expresses argument structure of verbs by **morphology**
  - English expresses argument structure of verbs by **word order**

What is Still Hard (2)

- **Sentence-level coherence**
  - a sentence should have a **verb**
  - **agreement** in number, case, argument structure, etc.
- **Document-level coherence**
  - how do you translate the English *it* into German?
  - **co-reference** / anaphora resolution

Solutions: Engineering

- Always: bigger, better, faster, more
  - vast possibilities with better **modeling and machine learning**
  - more efficient algorithms allow for more data and more complex models
  - from the titles of papers at this conference:

    online learning, discriminative training, noisy channel, unsupervised, discontiguous, noisy channel, correction model, n-gram-based, recursive acquisition, bootstrapping, domain adaptation, iterative refinement, log-linear, beam-search decoding, structural phrase alignment, hypothesis re-ranking, parallel fragment extraction, finite state

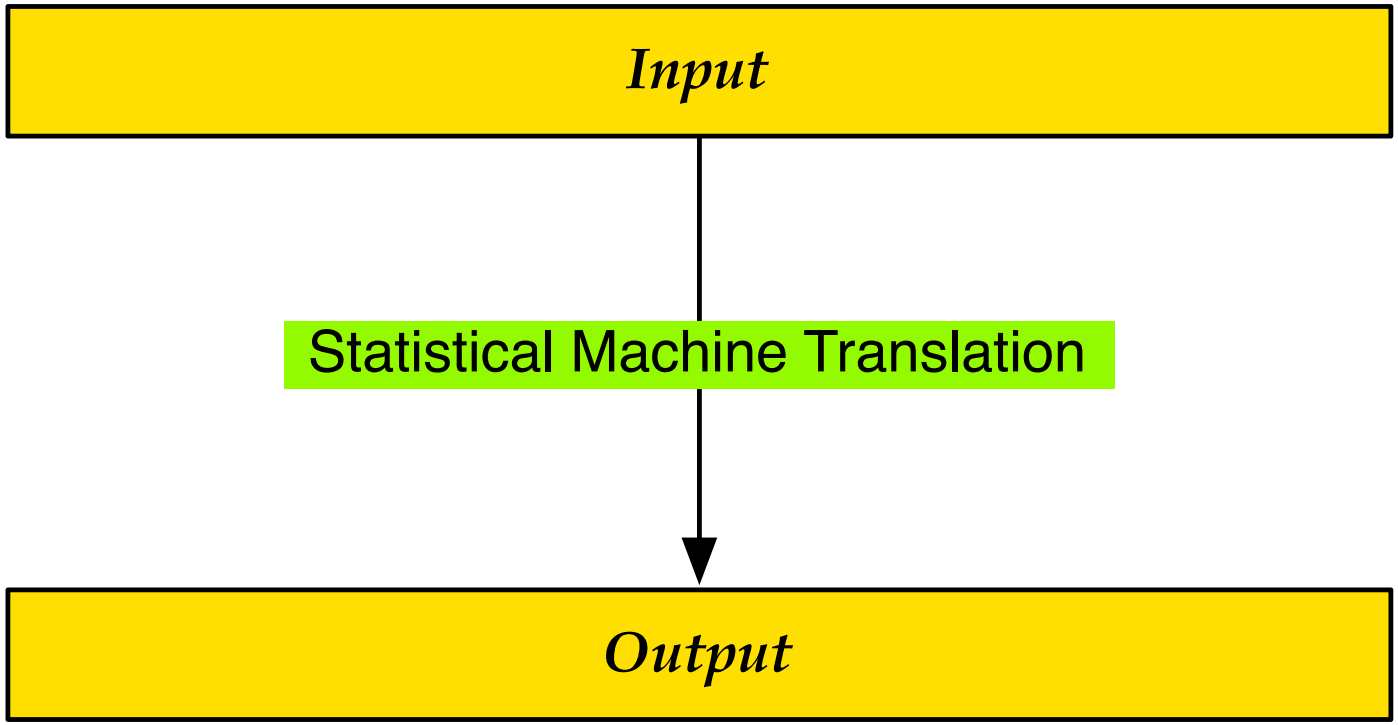
We do not need syntax, we need a lower-perplexity language model.

[from Franz Och]

Hybrid machine translation

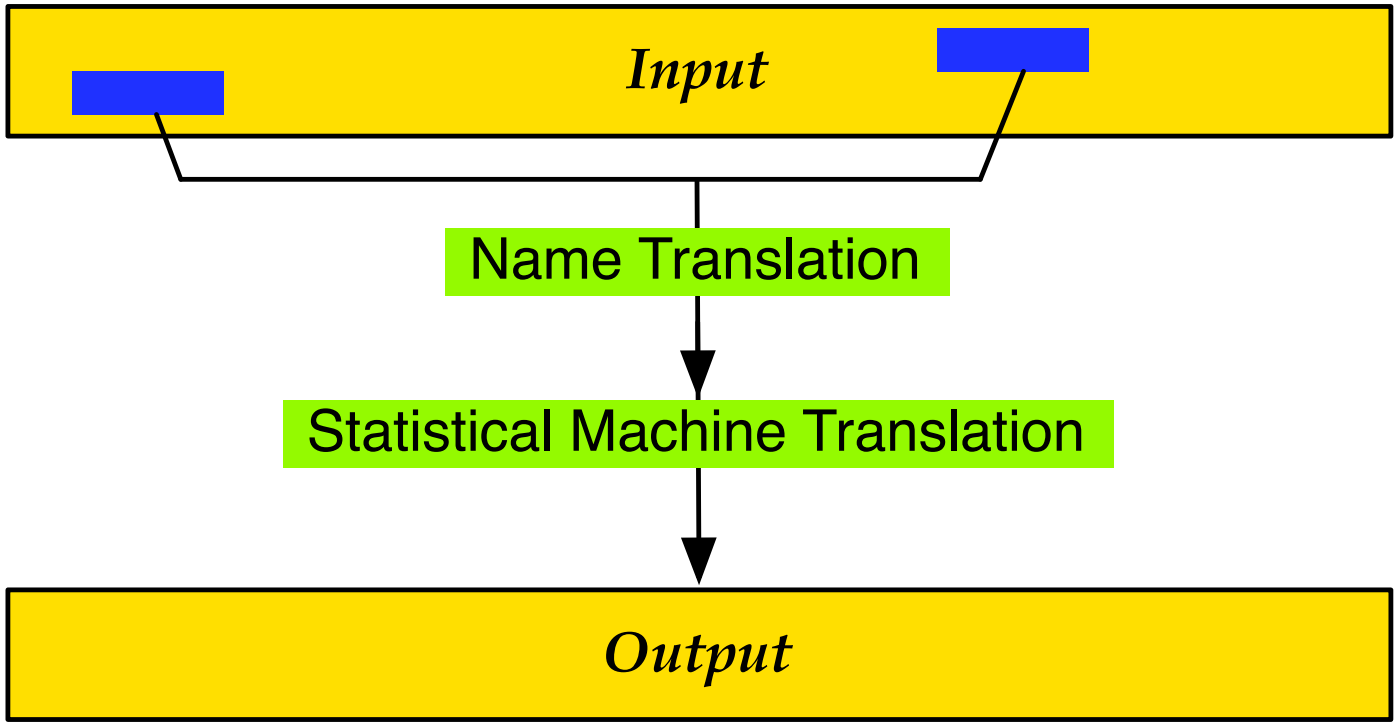
- Big idea: **combine the strengths** of rule-based and statistical MT
- Shallow integration
  - creating training data with rule-based systems
  - using the rule-base system's lexicon as training data
  - consensus translation of system outputs
- Deep integration
  - sharing of components
  - adding statistics for ambiguous rules

Hybrid machine translation



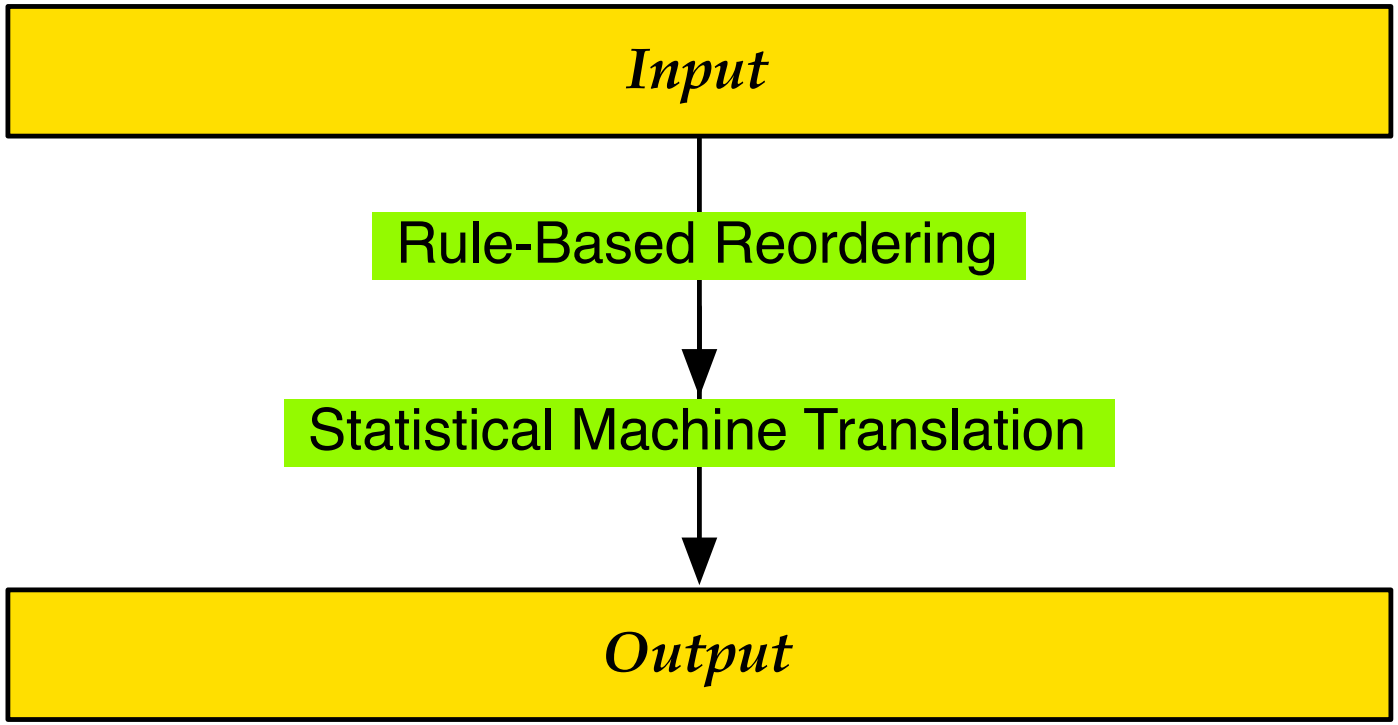
Purist's statistical machine translation system

Hybrid machine translation



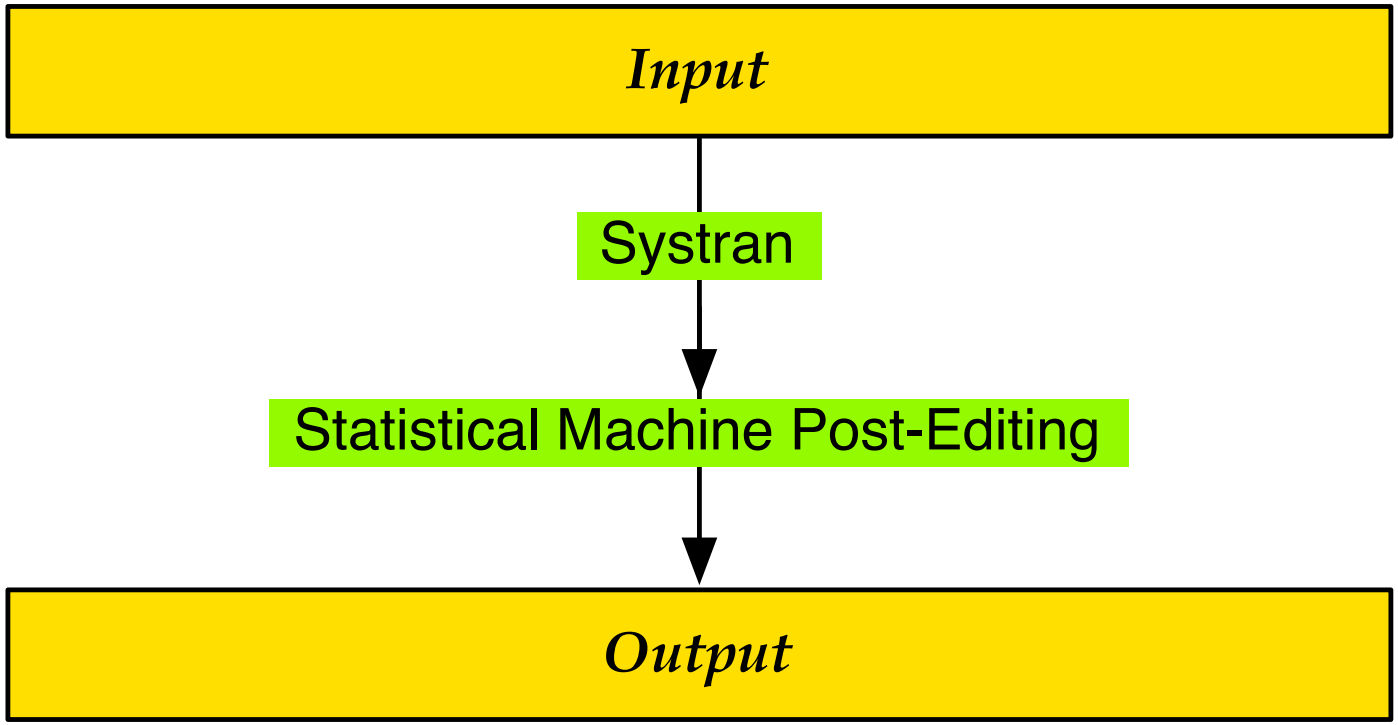
Name translation as rule-based special component

Hybrid machine translation



Rule-based reordering based on syntax [Collins et al., 2005, 2006]

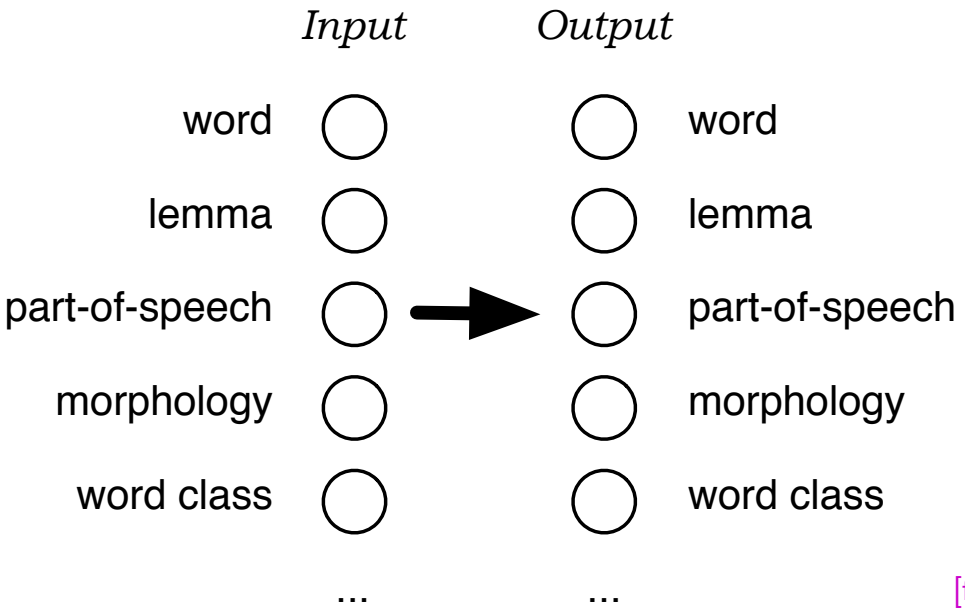
Hybrid machine translation



Systran as preprocessing [Simard et al., 2007; Dugast et al., 2007]

Factored Translation Models

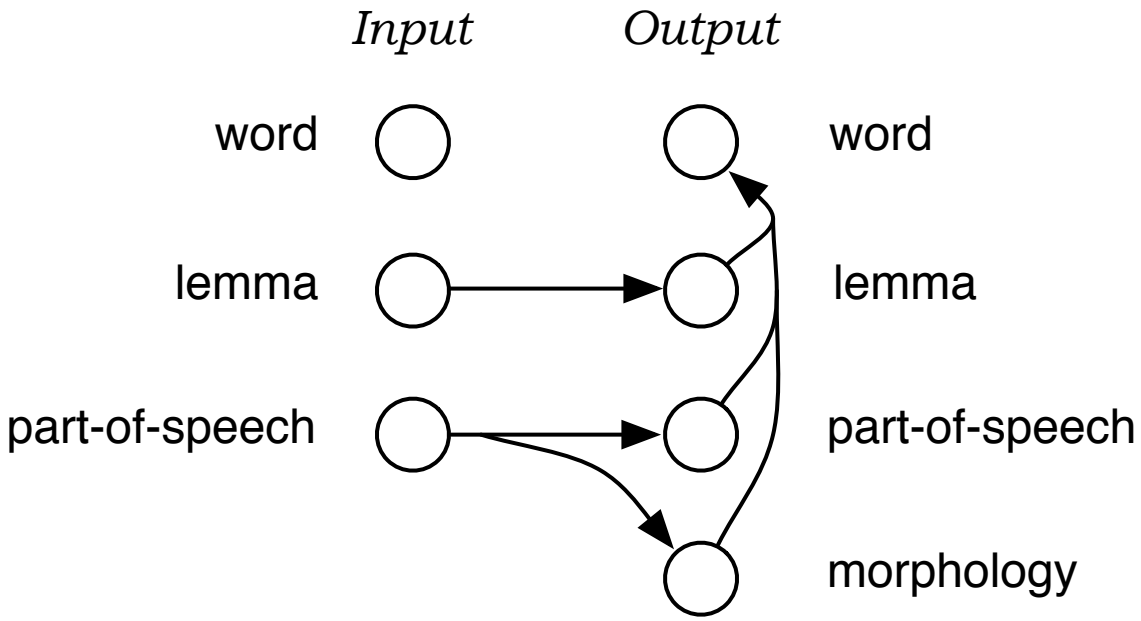
- **Factored represention** of words



[from Koehn and Hoang, 2007]

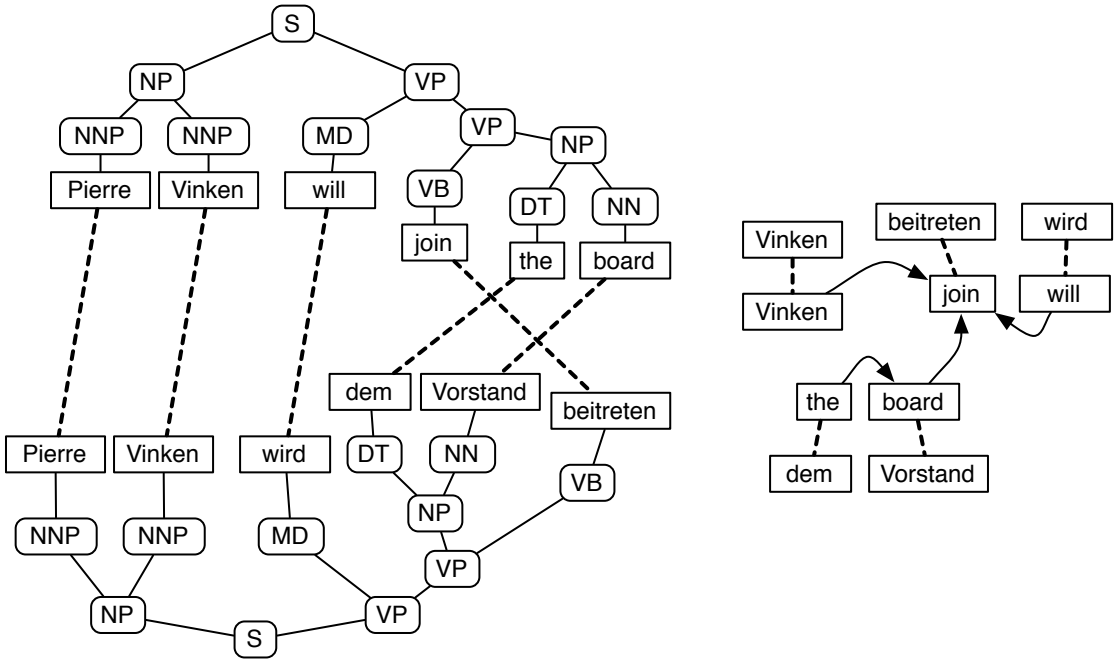
- Benefits
  - **generalization**, e.g. by translating lemmas, not surface forms
  - **richer model**, e.g. using syntactic info for reordering, language modeling

Morphological analysis and generation model



- addresses data sparseness through lemma and morphology mapping

Tree-based models

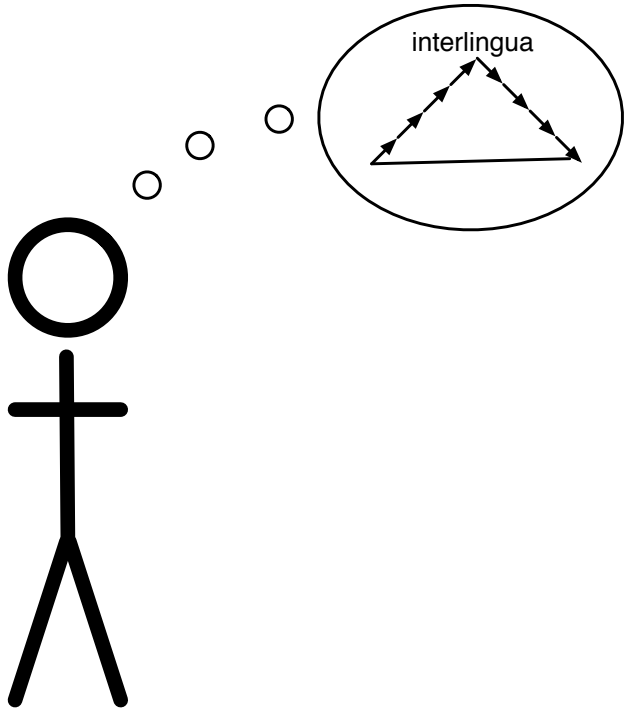


Word order difference are **explained by syntactic trees**

Often disappear in dependency structures

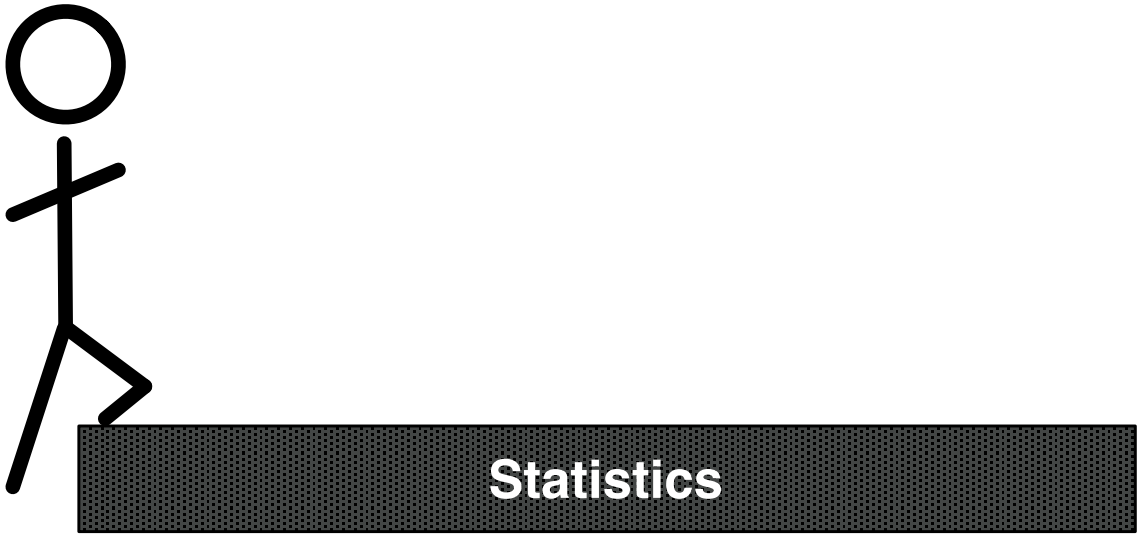
Tecto-grammatic: dependency structure with additional markup

Conclusion



Dreams of Interlingua

Conclusion

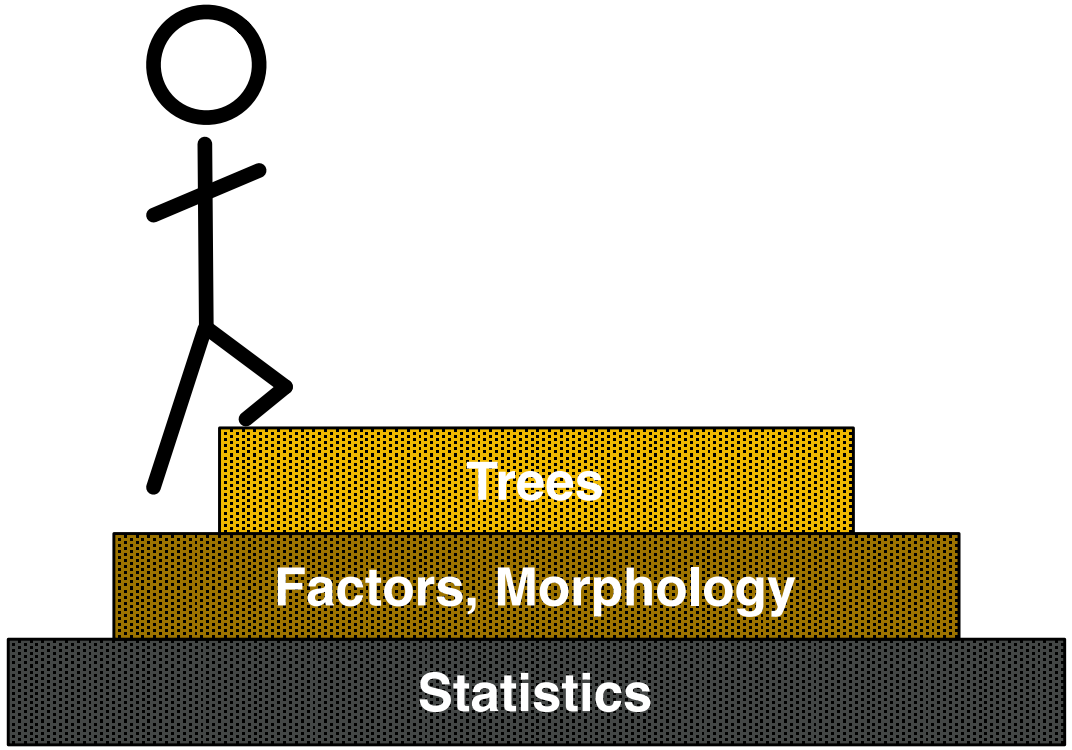


Building a strong statistical foundation

Conclusion

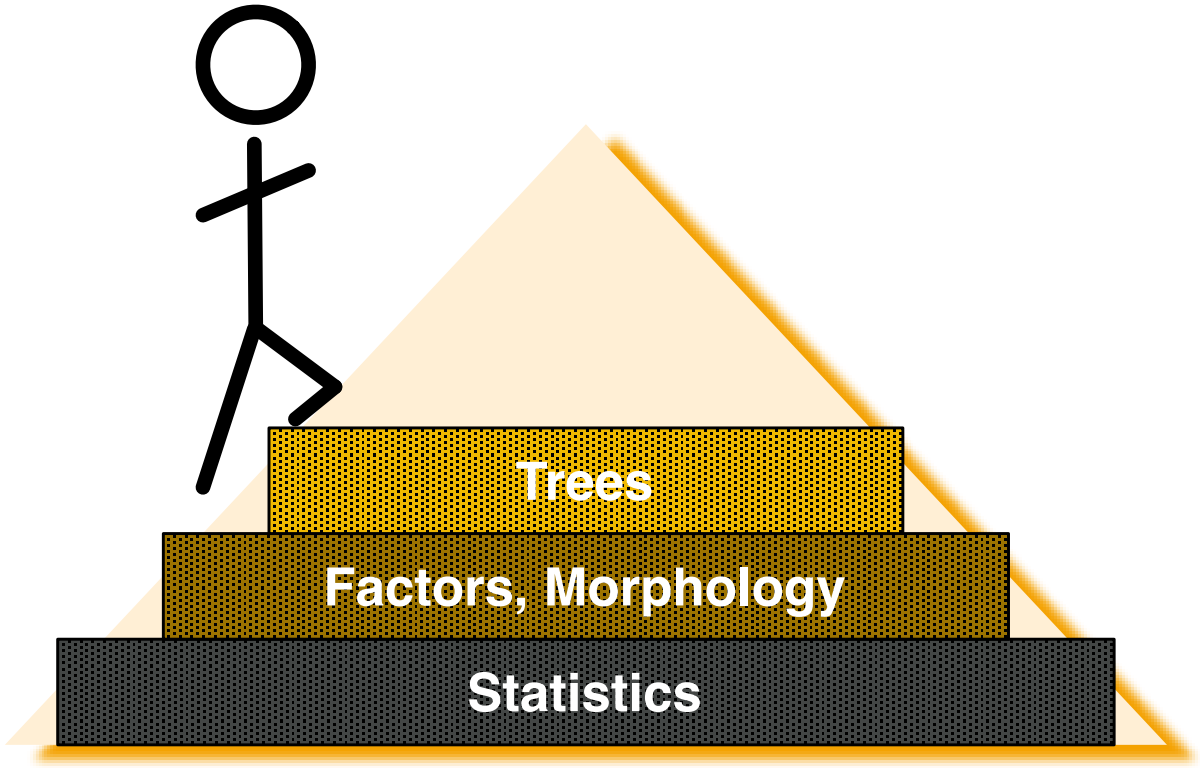
Adding linguistic annotation (POS, morphology, etc.)

Conclusion



Using tree-based transfer models

Conclusion



Statistical MT is getting so good, we need linguistics again.