

Machine Translation II: Statistical and Hybrid Methods

January 16, 2008



Andreas Eisele

UdS Computerlinguistik & DFKI

eisele@dfki.de

Language Technology I

WS 2007/8

- History of *Statistical* MT
- Distorted Channel Paradigm, again
- IBM Models and Variations
- Hybrid MT Architectures
- Evaluation techniques

- 1949: Warren Weaver: *the translation problem can be largely solved by “statistical semantic studies”*
- 1950s..1970s: Predominance of rule-based approaches
- 1966: ALPAC report: general discouragement for MT (in the US)
- 1980s: example-based MT proposed in Japan (Nagao), statistical approaches to speech recognition (Jelinek e.a. at IBM)
- Late 80s: Statistical POS taggers, SMT models at IBM, work on translation alignment at Xerox (M. Kay)
- Early 90s: many statistical approaches to NLP in general, IBM's Candide claimed to be as good as Systran
- Late 90s: Statistical MT successful as a fallback approach within Verbmobil System (Ney, Och). Wide distribution of translation memory technology (Trados) indicates big commercial potential of SMT

- 1999 Johns Hopkins workshop: open source re-implementation of IBM's SMT methods (GIZA)
- since 2001: DARPA/NIST evaluation campaign (XYZ -> English), uses BLEU score for automatic evaluation
- Various companies start marketing/exploring SMT:
language weaver, aixplain GmbH, Linear B Ltd., esteam, Google Labs
- 2002: Philipp Koehn (ISI) makes EuroParl corpus available
- 2003: Koehn, Och & Marcu propose *Statistical Phrase-Based MT*
- 2004: ISI publishes Philipp Koehn's SMT decoder *Pharaoh*
- 2005: First SMT workshop with shared task
- 2006: Johns Hopkins workshop on OS factored SMT decoder Moses, Start of EuroMatrix project for MT between all EU languages, Acquis Communautaire (EU laws in 20+ languages) made available
- 2007: Google abandons Systran and switches to own SMT technology

Existing MT systems for EU languages

Situation in early 2005, almost all systems are rule-based

From Hutchins: Compendium of Translation Software, 2005

	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Pol.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40	23		11	13	8	4	1	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-
Spanish	41	7	11		9	8	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Italian	29	10	13	9		4	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Portuguese	29	5	7	8	4		1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dutch	10	2	4	1	1	1		-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Polish	7	2	1	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Latvian	2	1	1	1	1	1	1	-		-	-	-	-	-	-	-	-	-	-	-	-	-	-
Greek	3	-	3	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-
Czech	1	1	1	-	1	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-
Hungarian	2	2	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-
Swedish	2	1	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-
Finnish	2	1	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-
Slovak	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
Romanian	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Existing MT systems for EU languages

Situation in early 2005, almost all systems are rule-based

From Hutchins: Compendium of Translation Software, 2005

	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40																						
Spanish	41																						
Italian	29	Amikai; Babelfish; Click2Translate; Dictionary.com																					
Portuguese	29	Translator; Easy Translator; e- Translation Server;																					
Dutch	10	FB-Active; FB-Win; FJWSpylltrans; FreeTranslation;																					
Polish	7	GETrans; Google; Hypertrans; IM Translator;																					
Latvian	2	iTranslator On-line; JxEuro; Korya Eiwa Ippatu																					
Greek	3	Honyaku; Language Weaver SMTS; LocalTranslation;																					
Czech	1	LogoMedia; Lycos; MZ-Win Translator; NeuroTran;																					
Hungarian	2	Palm Translator; PC Translator 2005; Personal																					
Swedish	2	Translator PT; PocketPROMT; Power Translator																					
Finnish	2	Global; Pragma; Pragma Online; @prompt;																					
Slovak	-	PROMT-Online; PT-SMS; PT-WAP; Reverso [series];																					
Romanian	1	SDL Enterprise; Smart Translator; Systran; T1;																					
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Existing MT systems for EU languages

Situation in late 2005, after Philipp Koehn created 110 SMT systems from EuroParl

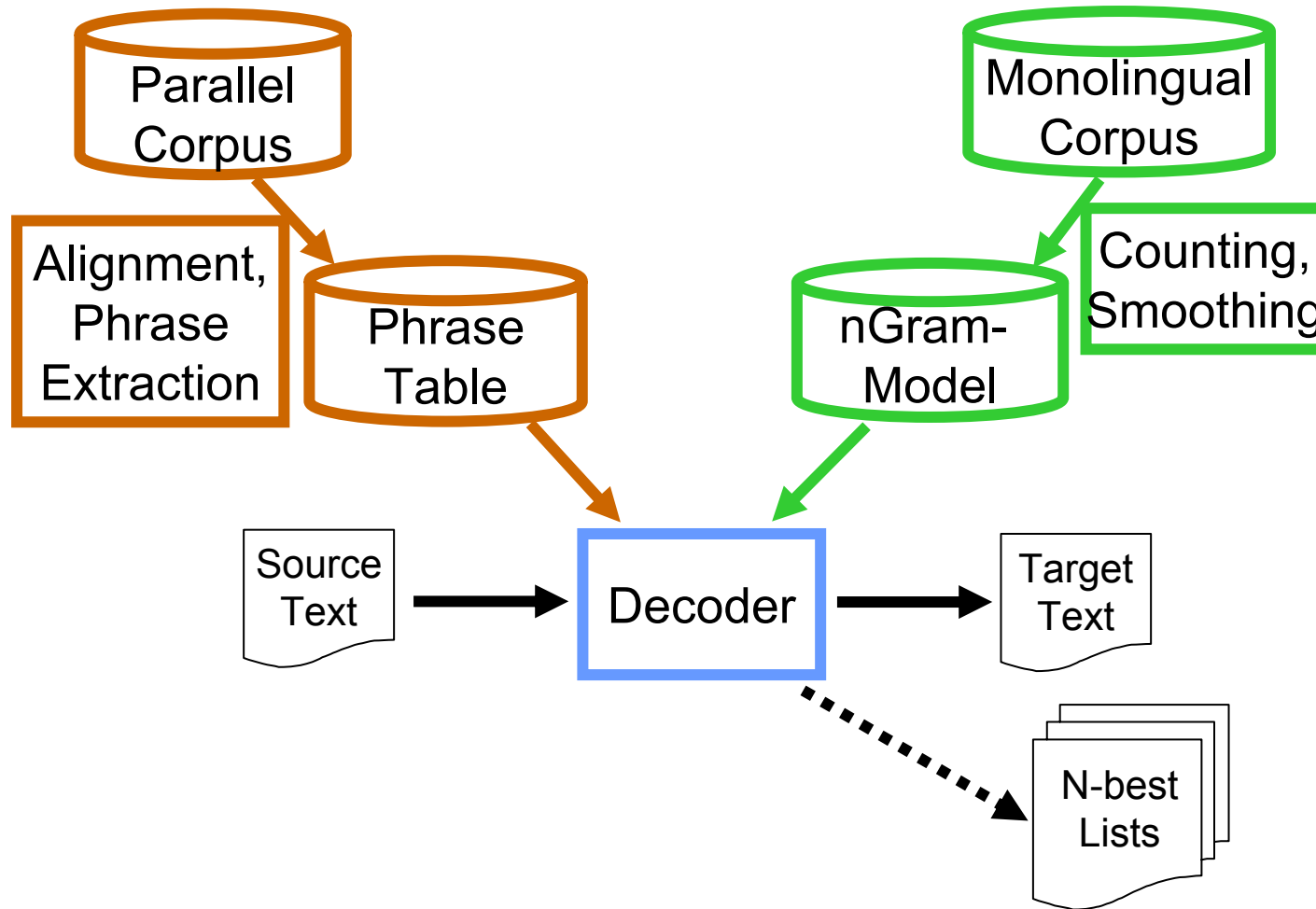
	En	Ge	Fr	Sp	It	Po	Du	Gr	Sw	Fi	Po	Da	La	Cz	Hu	Sl	Ro	Bu	Sl	Ma	Li	Ir	Es
English		48	42	45	31	31	11	5	2	1	8	1	2	1	4	1	1	2	-	-	-	-	-
German	49		25	9	11	5	3	1	2	2	3	2	1	1	2	1	-	-	-	-	-	-	-
French	41	24		12	14	9	5	4	1	1	1	1	1	1	-	-	-	-	-	-	-	-	-
Spanish	42	8	12		10	9	2	1	1	1	-	1	1	1	-	-	-	-	-	-	-	-	-
Italian	30	11	14	10		5	2	1	1	1	-	1	1	1	-	-	-	-	-	-	-	-	-
Portuguese	30	6	8	9	5		2	1	1	1	-	1	1	-	-	-	-	-	-	-	-	-	-
Dutch	11	3	5	2	2	2		1	1	1	-	1	1	-	-	-	-	-	-	-	-	-	-
Greek	4	1	4	1	1	1	1		1	1	-	1	-	-	-	-	-	-	-	-	-	-	-
Swedish	3	2	1	1	1	1	1	1		1	-	1	-	-	-	-	-	-	-	-	-	-	-
Finnish	3	2	1	1	1	1	1	1	1		-	1	-	-	-	-	-	-	-	-	-	-	-
Polish	7	2	1	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-
Danish	1	2	1	1	1	1	1	1	1	1	-		-	-	-	-	-	-	-	-	-	-	-
Latvian	2	1	1	1	1	1	1	-	-	-	-	-		-	-	-	-	-	-	-	-	-	-
Czech	1	1	1	-	1	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-	-
Hungarian	2	2	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-	-
Slovak	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-	-
Romanian	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

Existing MT systems for EU languages

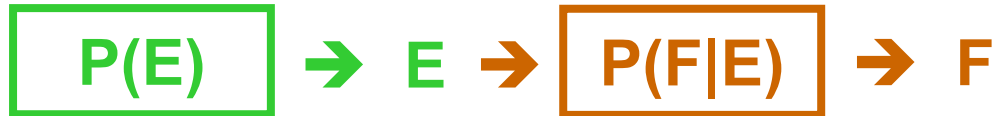
Situation in late 2007, if Acquis Communautaire is used for SMT training

	En	Ge	Fr	Sp	It	Po	Du	Gr	Sw	Fi	Po	Da	La	Cz	Hu	Sl	Ro	Sl	Ma	Li	Es	Bu	Ir
English		49	43	46	32	32	12	6	3	2	9	2	3	2	5	2	2	1	1	1	1	2	-
German	50		26	10	12	6	4	2	3	3	4	3	2	2	3	2	1	1	1	1	1	-	-
French	42	25		13	15	10	6	5	2	2	2	2	2	2	1	1	1	1	1	1	1	-	-
Spanish	43	9	13		11	10	3	2	2	2	1	2	2	2	1	1	1	1	1	1	1	-	-
Italian	31	12	15	11		6	3	2	2	2	1	2	2	2	1	1	1	1	1	1	1	-	-
Portuguese	31	7	9	10	6		3	2	2	2	1	2	2	1	1	1	1	1	1	1	1	-	-
Dutch	12	4	6	3	3	3		2	2	2	1	2	2	1	1	1	1	1	1	1	1	-	-
Greek	5	2	5	2	2	2	2		2	2	1	2	1	1	1	1	1	1	1	1	1	-	-
Swedish	4	3	2	2	2	2	2	2		2	1	2	1	1	1	1	1	1	1	1	1	-	-
Finnish	4	3	2	2	2	2	2	2	2		1	2	1	1	1	1	1	1	1	1	1	-	-
Polish	8	3	2	1	1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	1	-	-
Danish	2	3	2	2	2	2	2	2	2	2	1		1	1	1	1	1	1	1	1	1	-	-
Latvian	3	2	2	2	2	2	2	1	1	1	1	1		1	1	1	1	1	1	1	1	-	-
Czech	2	2	2	1	2	1	1	1	1	1	1	1	1		1	1	1	1	1	1	1	-	-
Hungarian	3	3	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	1	-	-
Slovak	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	1	-	-
Romanian	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	1	-	-
Slovene	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	1	-	-
Maltese	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1	-	-
Lithuanian	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1	-	-
Estonian	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

schematic architecture



- Based on „*distorted channel*“ Paradigm (successful for pattern- and speech recognition)



- Decoding: Given observation F , find most likely cause E^*

$$E^* = \operatorname{argmax}_E P(E|F) = \operatorname{argmax}_E P(E, F) = \operatorname{argmax}_E P(E) * P(F|E)$$

- Three subproblems

Model of $P(E)$

Model of $P(F|E)$

Search for E^*

each has approximative solutions

nGram-Models $P(e_1 \dots e_n) = \prod P(e_i | e_{i-2} \dots e_{i-1})$

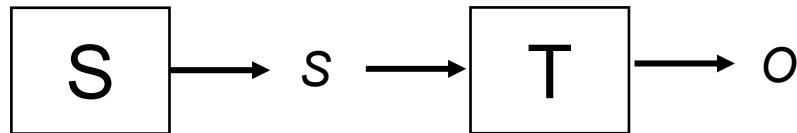
Transfer of „phrases“ $P(F|E) = \prod P(f_i | e_i) * P(d_i)$

Heuristic (*beam*) search

- Models are trained with (parallel) corpora, **correspondences (alignments)** between languages are estimated via EM-Algorithm (GIZA++, F.J.Och)

“Distorted Channel” Paradigm in General

Assume a signal that has to be transmitted through a channel that may add distortion, noise, or other modifications:



Both the source of the signal and the transmission channel can be characterized as probability distributions:

$P(s)$: probability that signal s is generated

$P(o|s)$: probability that observation o is made, *given* s

$P(o,s) = P(s) * P(o|s)$: probability that s is sent *and* o is observed

In typical applications, the most likely cause s' for a given observation o is sought, i.e.

$$s' = \operatorname{argmax}_s P(s|o) = \operatorname{argmax}_s P(s,o) = \operatorname{argmax}_s P(s) * P(o|s)$$

Communications Engineering:

S may be an input device, T a transmission line (modem line, audio/video transmission)

Speech recognition:

S is the speaker's brain, generating a string of words

T is the chain consisting of speakers articulatory device, sound transmission, microphone, signal processing up to morpheme hypotheses. The task is to reconstruct from a string of decoded sound events the intended chain of words.

Machine translation:

S is text in one language, T is translation to another
applying this model means to translate from the target language of the assumed "distortion" to the source, which can be confusing

Error correction

S is the intended (correct) text, T is the modification by introducing typing, spelling and other errors

OCR, ...

- $P(S)$ and $P(O|S)$ can be modeled independently
- $P(S)$ can be approximated using large amounts of monolingual text (e.g. using n-gram models)
- The same monolingual model $P(S)$ can be useful for ASR, OCR, and SMT
- Models based on grammars and other knowledge sources are harder to build, but will be superior in the long run
- $P(O|S)$ can be approximated from translated corpora, when correspondences between source and target language are known or can be estimated
- Models that take linguistic structure into account will be better in the long run, require more effort to be built & used

- Brown e.a. 1993 propose 5 different ways to define $P(F|E)$ and to train the parameters from a bilingual corpus
- There is a chicken-and-egg situation between translation models and alignments: given one, we can estimate the other. The standard approach to bootstrap reasonable models from partially hidden data is the Expectation-Maximization (EM-) Algorithm (as also used e.g. for HMMs)
- Model 1 assumes a one-to-one relation between individual words and a uniform distribution over all possible permutations
- Model 2 is similar, but prefers alignments that roughly preserve the original order

Word Alignment Example from Europarl

Frau Ludford , möchten Sie auch wirklich eine Anmerkung zum Protokoll machen ?

NULL	.	.	.	.	*	.	*	.	.	.	*	.
Mrs	*	.	.	.	.	.	.	.	.	.	.	.
Ludford	.	*	.	.	.	.	.	.	.	.	.	.
,	.	.	*	.	.	.	.	.	.	.	.	.
are	.	.	.	*	.	.	.	.	.	.	.	.
you	.	.	.	.	*	.	.	.	.	.	.	.
sure	.	.	.	.	.	*	.	.	.	.	.	.
your	.	.	.	.	.	.	.	.	.	.	.	.
point	.	.	.	.	.	.	.	*	.	.	.	.
of	.	.	.	.	.	.	.	.	.	.	.	.
order	.	.	.	.	.	.	.	.	.	.	.	.
is	.	.	.	.	.	.	.	.	.	.	.	.
related	.	.	.	.	.	.	.	.	.	.	.	.
to	.	.	.	.	.	.	.	.	.	.	.	.
the	.	.	.	.	.	.	.	.	*	.	.	.
Minutes	.	.	.	.	.	.	.	.	.	*	.	.
?	.	.	.	.	.	.	.	.	.	.	.	*

- Model 3 assumes that one English word can give rise to multiple French words by introducing “fertilities”, i.e. distributions over the number of words in the translation of a given word. Exact calculation of EM-estimates becomes infeasible and is replaced with approximations restricted to plausible subsets of all possible alignments.
- Model 4 introduces a distinction between groups of words (derived from one source word) that tend to stay together (like: *implemented* → *mis en application*) and groups that tend to get separated (like: not → *ne ... pas*).
- Model 5 is similar to Model 4, but avoids to distribute probability mass over impossible word sequences, e.g. sequences where words are missing or positions are simultaneously occupied with more than one word.
- Formulas in the CL'93 paper look heavy, but there are many tutorials and even an open-source implementation available.

- Bootstrapping also works across models of increasing complexity (i.e. alignment from Model i is used to estimate parameters for Model $i+1$)
- Development of the IBM models was based on about 1.8 million sentence pairs from the Canadian parliament debates (Hansards)
- Decoding (i.e. search for $\text{argmax}_s P(s) * P(o|s)$) was computationally challenging for long sentences, hence various heuristics for sentence splitting were used
- The performance was evaluated in a '94 ARPA test; Candide translations were judged as more fluent, but less adequate than those of Systran
- All models assume that correspondences are triggered by single words on the source level side, i.e. there is no support for phrase-to-phrase alignments

- Parallel text
- Sentence segmentation and tokenization
- Sentence alignment
- Make sure you will have unseen test data
- Word alignment
- Phrase-table construction
- More text from target language
- Stochastic (target) language model
- Decoding
- Inspect/evaluate results

De-facto standard: EUROPARL corpus

“Successor” of Canadian Hansards used by IBM

Free, no legal constraints

11 languages -> 110 directed language pairs

includes some “important” languages

somewhat representative for EU publications

future versions will have 20++ languages

But:

does not cover the most difficult/interesting languages (Chinese, Arabic, Japanese, Walpiri, Inuktitut, ...)

not very technical

not overly interesting from stylistic point of view

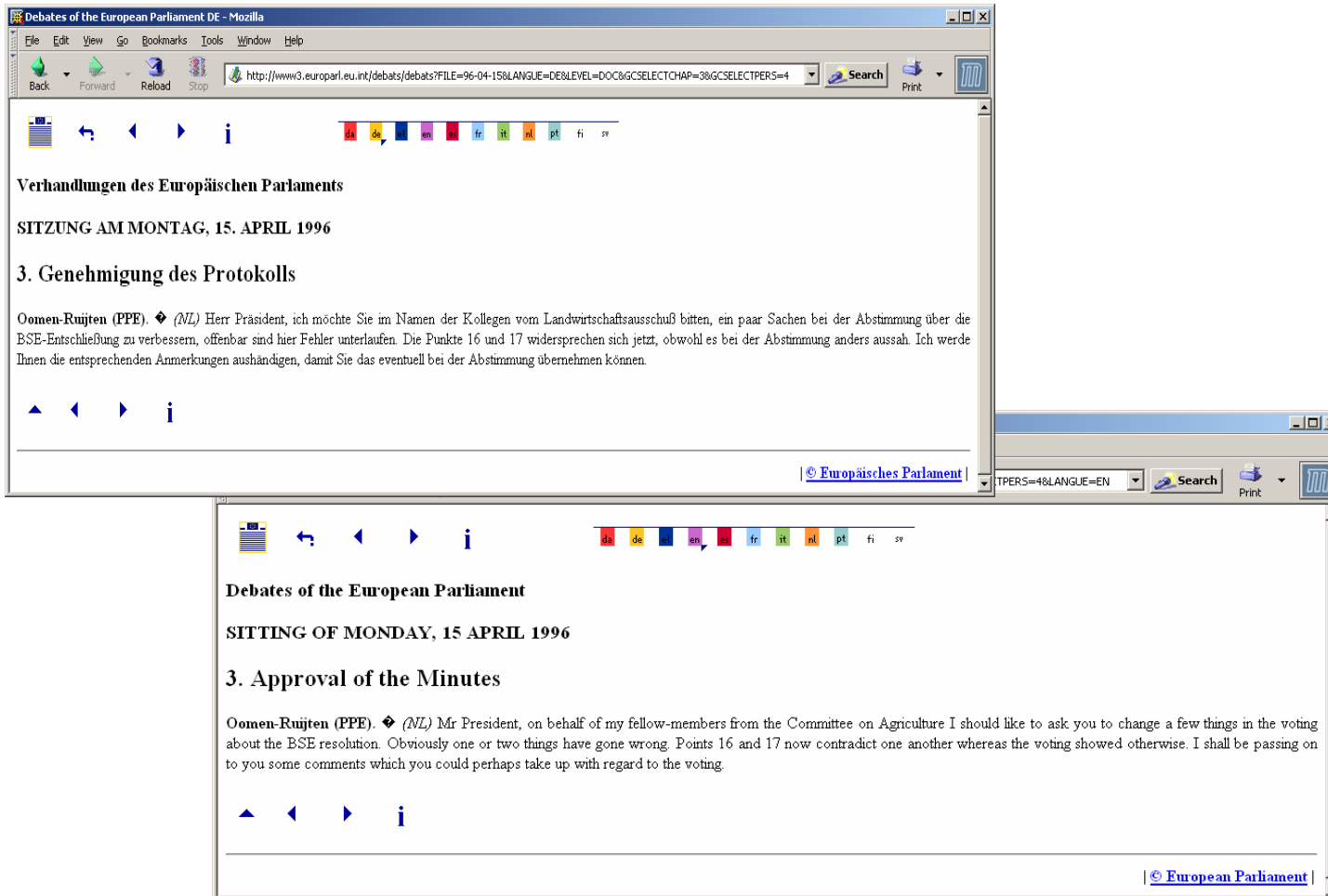
dependencies on context as in typical written text

In the meantime:

EU has been extended to 25 states with 20 official languages (more to come), official law has been translated to all these languages

(= “**Acquis Communautaire**” Corpus)

Parallel text: EUROPARL



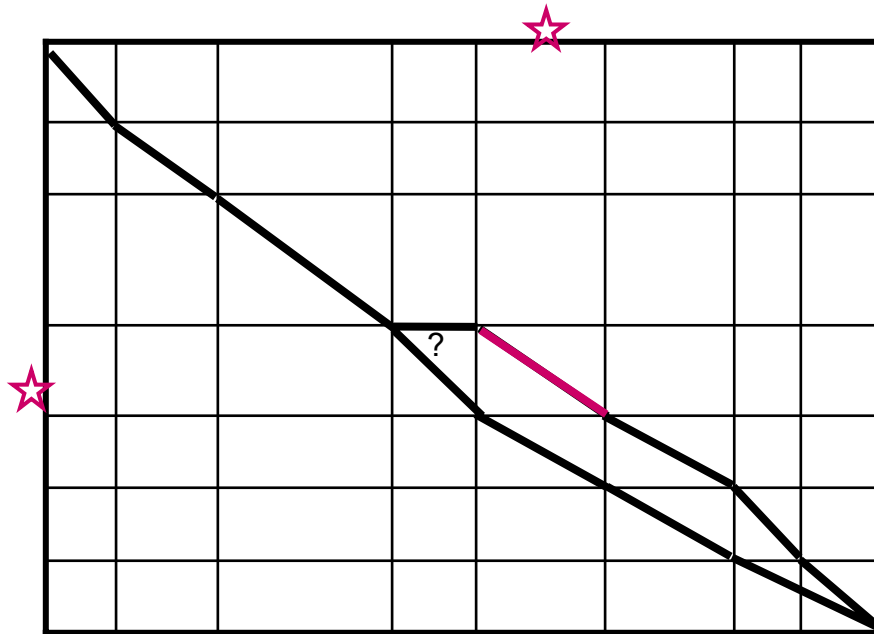
The image shows a side-by-side comparison of two web browser windows displaying the same content in different languages. The top window is in German, and the bottom window is in English. Both windows show the title 'Verhandlungen des Europäischen Parlaments' / 'Debates of the European Parliament' and the date 'SITZUNG AM MONTAG, 15. APRIL 1996' / 'SITTING OF MONDAY, 15 APRIL 1996'. The main heading is '3. Genehmigung des Protokolls' / '3. Approval of the Minutes'. The text in the German window is: 'Oomen-Ruijten (PPE). ♦ (NL) Herr Präsident, ich möchte Sie im Namen der Kollegen vom Landwirtschaftsausschuß bitten, ein paar Sachen bei der Abstimmung über die BSE-Entscheidung zu verbessern, offenbar sind hier Fehler unterlaufen. Die Punkte 16 und 17 widersprechen sich jetzt, obwohl es bei der Abstimmung anders aussah. Ich werde Ihnen die entsprechenden Anmerkungen aushändigen, damit Sie das eventuell bei der Abstimmung übernehmen können.' The text in the English window is: 'Oomen-Ruijten (PPE). ♦ (NL) Mr President, on behalf of my fellow-members from the Committee on Agriculture I should like to ask you to change a few things in the voting about the BSE resolution. Obviously one or two things have gone wrong. Points 16 and 17 now contradict one another whereas the voting showed otherwise. I shall be passing on to you some comments which you could perhaps take up with regard to the voting.'

- non-trivial to find texts in a form that facilitates processing (PDFs are quite inconvenient)
- millions of words distributed over 10000s of URLs
- some translations are missing
- the hard work of crawling the site has already been done for us (thanks to Philipp Koehn)
- results available from [**http://www.statmt.org/europarl/**](http://www.statmt.org/europarl/)

- Both can be tricky if you want to get all the details right
 - “That is not true!” he said... → 1 or 2 sentences?
 - doesn't → doesn + ' + t **vs.** does + n't ?
- Distinguishing end-of-sentence marks from sentence-internal punctuation requires recognition of abbreviations, which isn't quite trivial to do in 11 languages
- We may not always like Philipp Koehn's decisions, but they are good enough to get a system up and running

- Problem: During translation, sentences may have been split, merged, dropped or re-ordered.
- This does not look hard until you have enough data (Europarl: 11*1.3M sentences)
- If data is clean and some errors are acceptable: Simple length-based heuristic does the job
- Task can be seen as finding an optimal path through rectangular grid (see next slide)
- Europarl v.1: 10 sentence alignments XY $\Leftrightarrow$ EN
- Europarl v.2: sentences + generic alignment tool

- Can be solved by dynamic programming



- Complexity is $O(n*m)$
- Additional evidence (e.g. from invariant or cognate words) can be helpful

- Words may (dis-)appear during translation, they get reordered, words replace constructions ...
→ almost impossible to reach full agreement on valid correspondences
- Simple stochastic models will automatically get the typical cases right, but will miss the tricky (=interesting) cases
- For SMT, the typical cases are most important; we may have to live with 10% error rate

- Typical approach: use IBM models as implemented in GIZA++ system
- Apply it in both directions
- Take intersection of results (increasing precision at the cost of recall)
- Extend using various heuristics
- Partial word alignments for 4 language pairs
DE/ES/FI/FR ↔ EN available from
<http://www.statmt.org/wpt05/mt-shared-task/>

Sample from the DE↔EN alignment:

Die Punkte 16 und 17 widersprechen sich jetzt , obwohl es bei der Abstimmung anders aussah .
Points 16 and 17 now contradict one another whereas the voting showed otherwise .
0-9 1-0 2-1 3-2 4-3 5-5 6-5 7-4 9-8 10-9 11-8 12-9 13-10 14-12 15-6 15-7 15-11 15-12 16-13

Same sample represented graphically:

```

      Die
    .   Punkte
      *   16
    .   .   und
      *   .   17
    .   .   .   widersprechen
      *   .   .   sich
    .   .   .   .   jetzt
      *   .   .   .   /
    .   .   .   .   obwohl
  *   .   .   .   .   es
    .   .   .   .   *   bei
      *   *   .   .   .   der
    .   .   .   .   .   Abstimmung
  .   .   .   .   .   anders
    .   .   .   .   .   aussah
      .   .   .   .   .
    .   .   .   .   .   Points
      .   .   .   .   .   16
    .   .   .   .   *   and
      .   .   .   .   .   17
    .   .   .   .   *   now
      .   .   .   .   *   contradict
    .   .   .   .   *   one
      .   .   .   .   .   another
    .   .   .   .   .   whereas
      .   *   *   .   the
    .   *   .   voting
      .   .   showed
    .   *   otherwise
      .

```

Idea: collect pairs of substrings that are compatible with word alignment

Incorporating linguistic constraints at this level has (so far) deteriorated outcomes

→ more and coarser is better

Phrase table are annotated with scores that will be used during decoding

widersprechen ||| contradict ||| 0.5 0.174039 0.227273 0.119306 2.718
widersprechen , ||| to contradict ||| 0.333333 0.046708 0.2 0.0134216 2.718
Kommissar Bolkestein ausdrücklich widersprechen ||| expressly contradict Commissioner Bolkestein ||| 1
0.0417032 1 0.0147184 2.718
widersprechen ||| contravening ||| 0.333333 0.0320171 0.0113636 0.0032612 2.718
nicht widersprechen ||| not contradictory ||| 0.125 0.0291049 0.111111 0.017083 2.718
nicht widersprechen ||| does not contravene ||| 0.5 0.0288053 0.111111 0.000371669 2.718
widersprechen oder ||| contradictory or ||| 0.333333 0.0251621 1 0.0207105 2.718
widersprechen ||| run counter ||| 0.4 0.017062 0.0681818 0.00114863 2.718
widersprechen ||| disagree ||| 0.0106383 0.0167791 0.0113636 0.0714746 2.718
Wir widersprechen ||| We disagree ||| 0.0666667 0.00997179 1 0.0503599 2.718
teilweise widersprechen ||| partly contradictory ||| 1 0.00637625 1 0.00291665 2.718
widersprechen ||| inconsistent ||| 0.0169492 0.00598197 0.0113636 0.0032612 2.718
widersprechen uns ||| contradicts us ||| 1 0.00561145 1 0.00174914 2.718
nur dann widersprechen ||| only overrule ||| 1 0.00216227 1 0.000444817 2.718
auch der Konferenz der Präsidenten widersprechen ||| contradict both the Conference of Presidents ||| 1
0.001813 1 5.17342e-05 2.718
Herr Bolkestein widersprechen ||| Mr Bolkestein disagrees with ||| 1 0.00175593 1 0.00041956 2.718
könnte dem widersprechen ||| could gainsay that ||| 1 0.00174458 1 4.90747e-06 2.718
widersprechen muß ||| have to contradict ||| 0.333333 0.00163608 0.5 0.000911924 2.718
widersprechen , wird ||| contradictory , is ||| 1 0.00161673 1 0.00362608 2.718
Änderungsanträge widersprechen dem ||| amendments contravene the ||| 1 0.00160169 1 0.0101469 2.718
17 widersprechen sich jetzt ||| 17 now contradict ||| 1 0.00143452 1 0.0283876 2.718
und 17 widersprechen sich jetzt ||| and 17 now contradict ||| 1 0.00120543 1 0.0256701 2.718
widersprechen zu müssen ||| to have to contradict ||| 1 0.00111525 0.333333 0.00167714 2.718
Herrn Brinkhorst nicht widersprechen ||| not disagree with Mr Brinkhorst ||| 1 0.00103174 1 0.00613701 2.718
einander widersprechen ||| contradict ||| 0.025 0.00101814 1 0.0609116 2.718
sich nicht widersprechen ||| are not contradictory ||| 0.25 0.000998935 1 0.00137116 2.718
widersprechen ||| any case contrary ||| 1 0.000890016 0.0113636 4.16211e-07 2.718
16 und 17 widersprechen sich jetzt ||| 16 and 17 now contradict ||| 1 0.000830368 1 0.0236414 2.718
widersprechen ||| conflict with ||| 0.0465116 0.000750812 0.0454545 0.00236106 2.718
James Elles widersprechen ||| what James Elles said ||| 1 0.00071772 1 0.00011574 2.718
nicht widersprechen ||| not conflict with ||| 0.4 0.00060168 0.222222 0.00164904 2.718
Rassismus , Fremdenfeindlichkeit und Antisemitismus widersprechen ||| racism , xenophobia and antisemitism
are completely incompatible with ||| 1 0.00055052 1 1.87174e-08 2.718

Motivation:

Translations should satisfy 2 requirements:

- equivalence with source sentence $P(f|e)$
- well-formedness $P(e)$

So far, we have only dealt with equivalence

Well-formedness can be approximated via even simpler stochastic models, based on n-gram probabilities.

We know (since Chomsky '57...) that n-gram models cannot capture essential long-distance effects, but in practice, 3-grams seem to be good enough...

Toolkits for counting word co-occurrences and estimating sentence probabilities have been developed for speech recognition.

Some are freely available:

- SRILM (Stolcke)

- CMU/Cambridge (Clarkson&Rosenfeld)

Philipp Koehn's Pharaoh decoder uses the SRILM format,
Koehn even distributed an English LM trained on Europarl

Dilemma: More text of slightly different type may help or hurt,
one needs to try it out

The decoder...

takes source sentence f , phrase table to estimate $P(e|f)$, SLM to estimate $P(e)$

searches for target sentence e that maximizes $P(e) * P(f|e)$

uses beam-search approximation, as complete search for optimal solution is not feasible

has some additional bells and whistles (factored models) that will improve the quality

Motivation for hybrid MT (1)

In the early 90s, SMT and RBMT were seen in sharp contrast.

But advantages and disadvantages are complementary.

→ Search for integrated methods is now seen as natural extension for both approaches

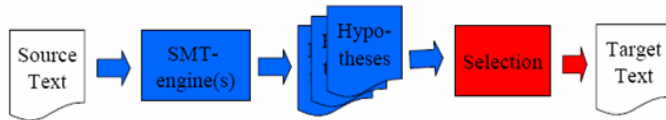
	RBMT	SMT
Syntax	+++	---
Structural Semantics	+	---
Lexical Semantics	-	+
Lexical Adaptivity	---	+

- Statistical and rule-based approaches address different types of knowledge:
 - Rule-based approaches focus on linguistic knowledge
 - Statistical approaches provide a holistic, integrated model that also incorporates (some) implicit knowledge of the world
- All available types of knowledge are urgently required, as the task is too difficult to ignore important aspects
- Research on a deep integration of statistical and linguistic approaches is required but this will take some time
- In the meantime, we can try to tinker with existing MT engines

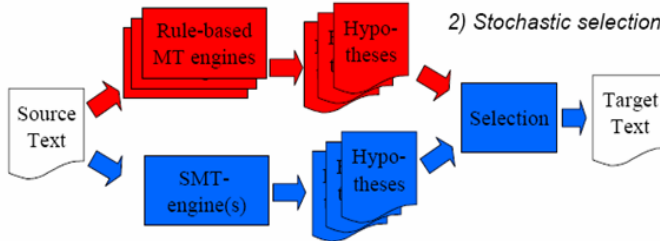
Some hybrid MT architectures

■ = SMT Module
■ = RBMT Module

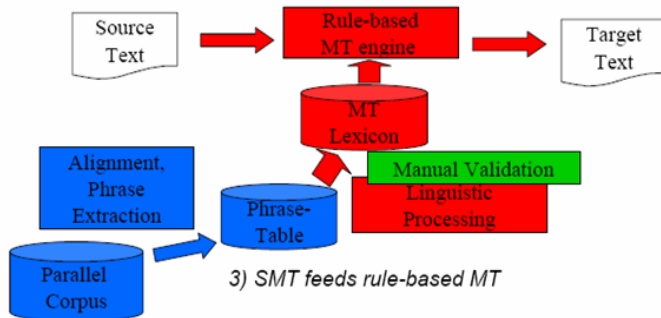
1) Syntactic selection



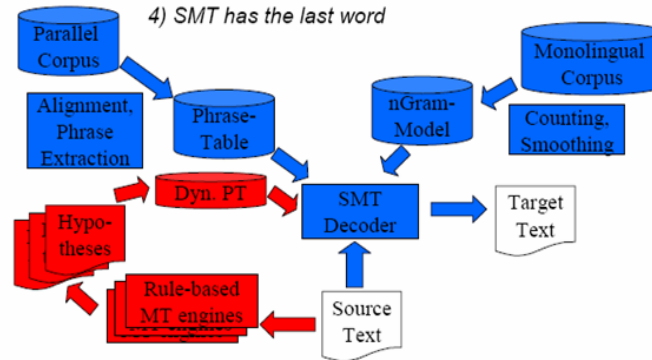
2) Stochastic selection



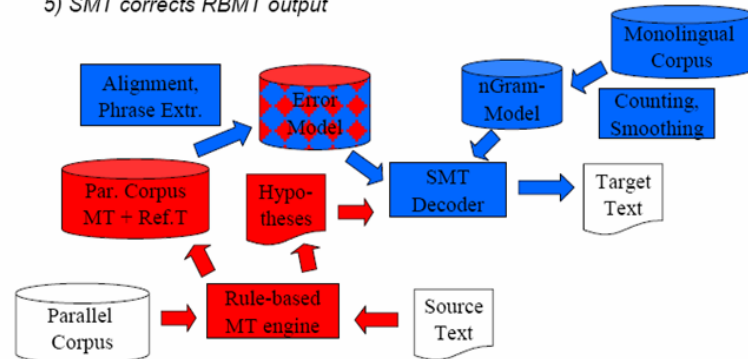
3) SMT feeds rule-based MT



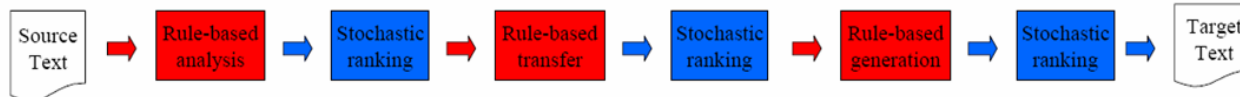
4) SMT has the last word



5) SMT corrects RBMT output



6) Rule-based transfer architecture interleaved with stochastic ranking



- Investigate two on-line MT engines by sending sentences through both of them and comparing the results
Use http://www.google.com/language_tools as a typical SMT system and compare with <http://babelfish.altavista.com/> (based on Systran)
- Try to find typical errors of both systems. Try variations of input sentences to find out whether the systems „understand“ what they translate