Machine Translation

May 21/23, 2013

UNIVERSITÄT DES SAARLANDES

Christian Federmann
Saarland University
cfedermann@coli.uni-saarland.de

Language Technology II
SS 2013

Machine Translation: Overview

- Relevance of MT, typical applications and requirements
- History of MT
- Basic approaches to MT
  - Rule-based
  - Example-based
  - **Statistical**
    - **word-based**
    - **tree-based**
  - **Hybrid, multi-engine**
- Evaluation techniques

Sources for Information

- MT in general, history:
  - http://www.MT-Archive.info: Electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools, regularly updated, contains over 3300 items
  - Hutchins, Somers: An introduction to machine translation. Academic Press, 1992, available under http://www.hutchinsweb.me.uk/IntroMT-TOC.htm
- MT systems:

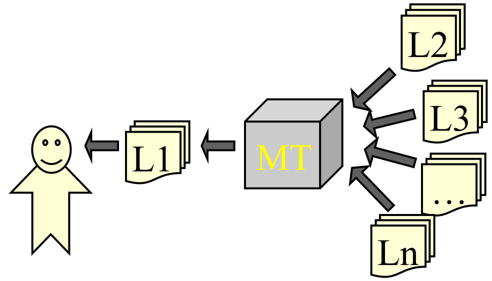
  Compendium of Translation Software, see http://www.hutchinsweb.me.uk/Compendium.htm
- Statistical Machine Translation:

  See www.statmt.org

  Book by Philipp Koehn is available in the coli-bib

Use cases and requirements for MT

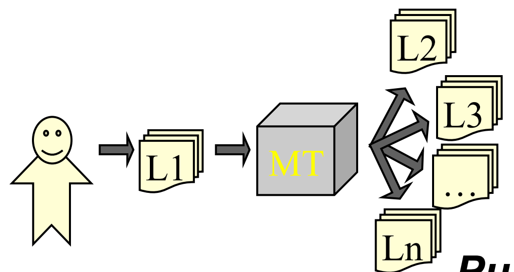
a) MT for assimilation
„inbound“



Robustness
Coverage

Daily throughput of online-MT-Systems > 500 M Words

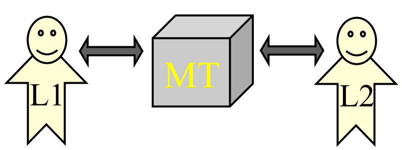
b) MT for dissemination
„outbound“



Textual quality

Publishable quality can only be authored by humans; Translation Memories & CAT-Tools mandatory for professional translators

c) MT for direct communication



Speech recognition, context dependence

Topic of many running and completed research projects (VerbMobil, TC Star, TransTac, …)
US-Military uses systems for spoken MT

On the Risks of Outbound MT

UNIVERSITÄT DES SAARLANDES

Some recent examples

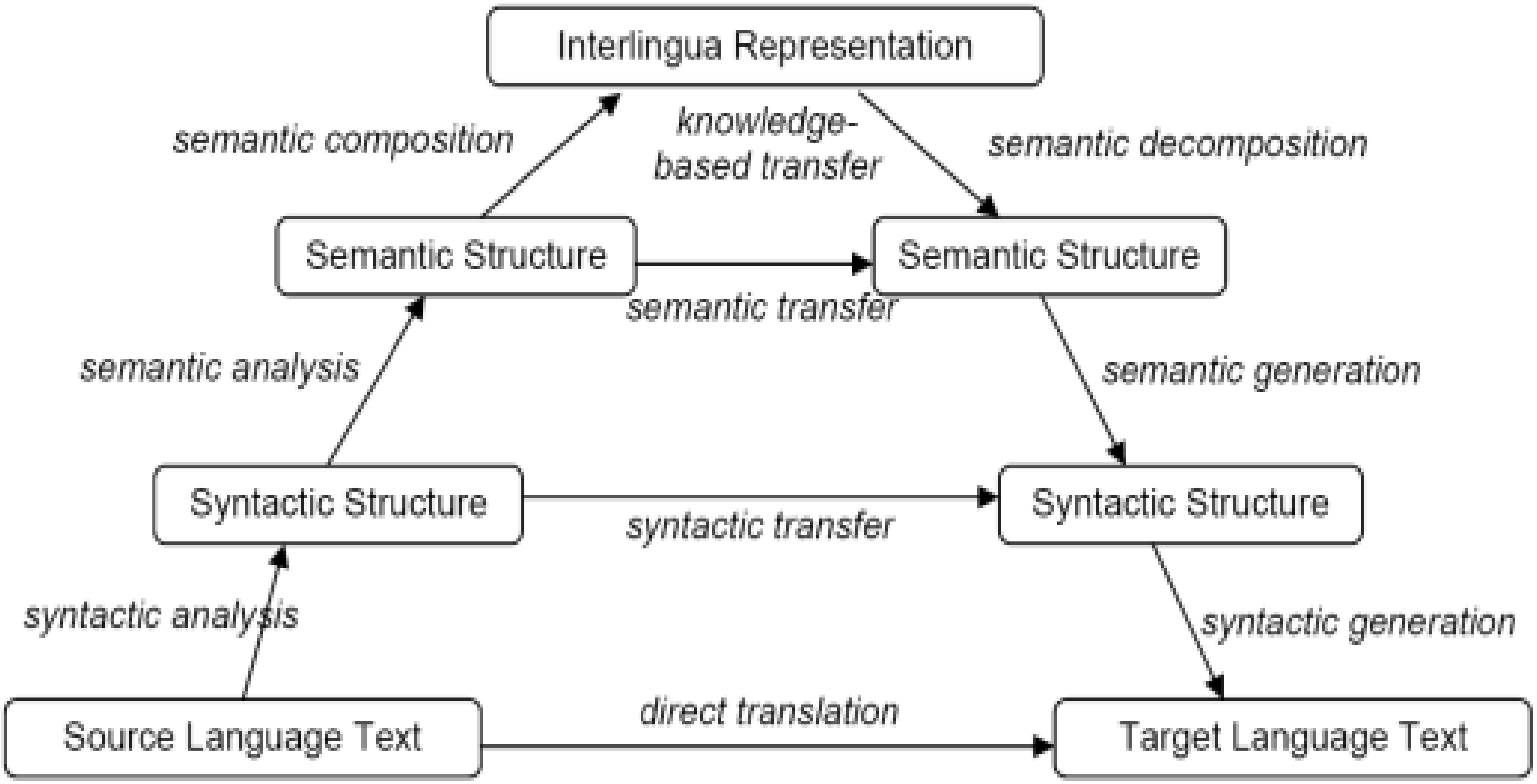
'I am not in the office at the moment. Please send any work to be translated'

Motivation for rule-based MT

- Good translation requires knowledge of linguistic rules
  - …for understanding the source text
  - …for generating well-formed target text
- Rule-based accounts for certain linguistic levels exist and should be used, especially for
  - Morphology
  - Syntax
- Writing one rule is better than finding hundreds of examples, as the rule will apply for new, unseen cases
- Following a set of rules can be more efficient than search for the most probable translation in a large statistical model

Possible (rule-based) MT architectures

The „Vauquois Triangle“



Motivation for statistical MT

UNIVERSITÄT DES SAARLANDES

- Good translation requires knowledge and decisions on many levels
  - syntactic disambiguation (POS, attachments)
  - semantic disambiguation (collocations, scope, word sense)
  - reference resolution
  - lexical choice in target language
  - application-specific terminology, register, connotations, good style …
- Rule-based models of all these levels are very expensive to build, maintain, and adapt to new domains
- Statistical approaches have been quite successful in many areas of NLP, once data has been annotated
- Learning from existing translation will focus on distinctions that matter (not on the linguist's favorite subject)
- Translation corpora are available in rapidly growing amounts
- SMT *can* integrate rule-based modules (morphologies, lexicons)
- SMT *can* use feed-back for on-line adaptation to domain and user preferences

History of SMT and Important Players I

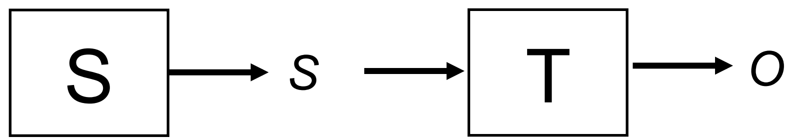
- 1949: Warren Weaver: *the translation problem can be largely solved by "statistical semantic studies"*
- 1950s..1970s: Predominance of rule-based approaches
- 1966: ALPAC report: general discouragement for MT (in the US)
- 1980s: example-based MT proposed in Japan (Nagao), statistical approaches to 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)

History of SMT and Important Players II

- 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
- 2009: Start of EuroMatrixPlus *"bringing MT to the user"*
- 2010: Start of many additional MT-related EU projects (Let's MT, ACCURAT, …)

Statistical Machine Translation

- Based on *„distorted channel"* paradigm
- Assume a signal that has to be transmitted through a channel that may add distortion/noise/etc.



- The source of the signal and the transmission channel can be characterized as probability distributions:
  - P(s): propability 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.

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

Applications of Distorted Channel Paradigm

- 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
- Error correction
  - S is the intended (correct) text
  - T is the modification by introducing typing, spelling and other errors
- OCR, …

Statistical Machine Translation

- How does that work in SMT?

$$\boxed{P(E)} \rightarrow E \rightarrow \boxed{P(F|E)} \rightarrow F$$

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

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

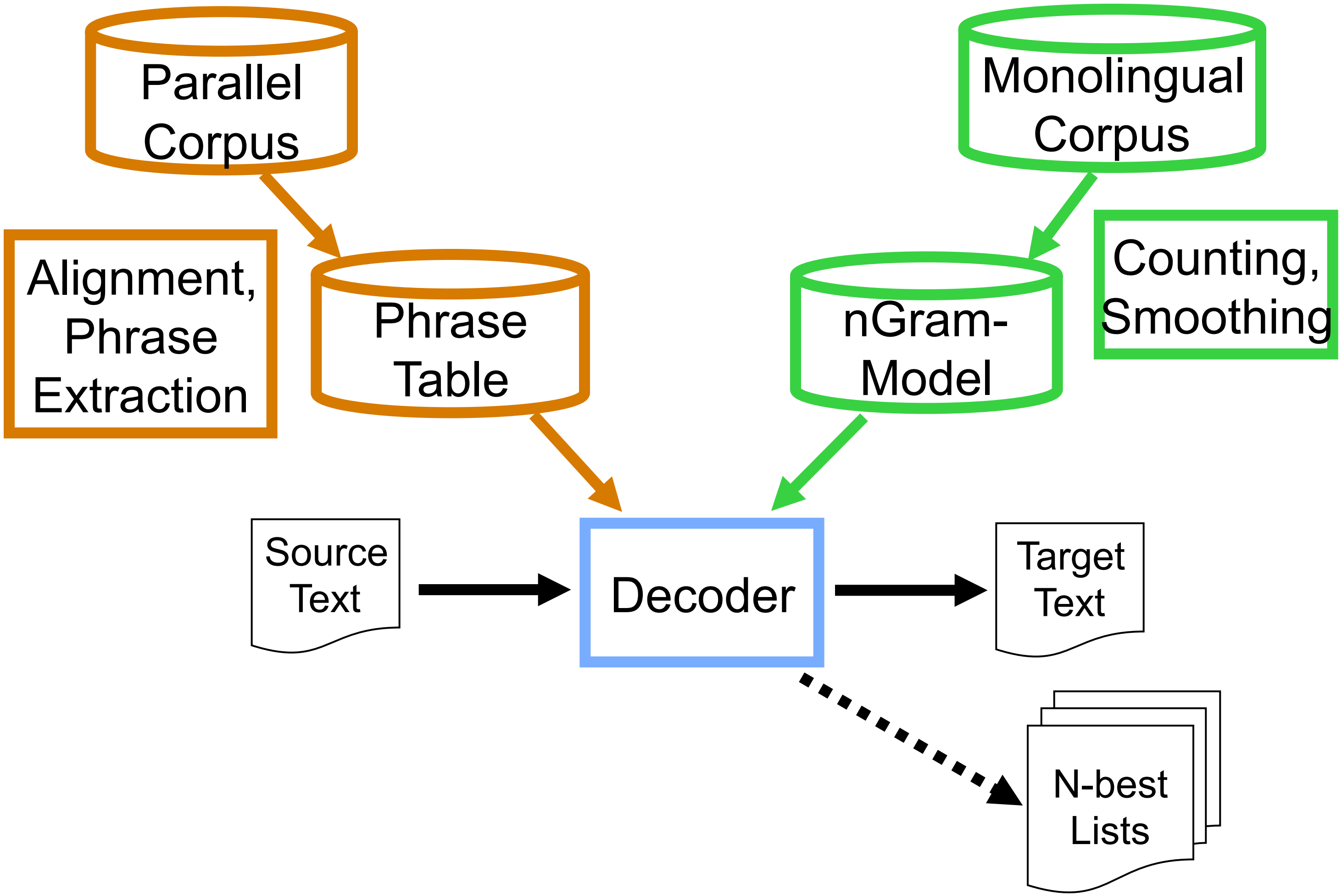
→ Three subproblems
Model of **P(E)**
Model of **P(F|E)**
Search for **E***

each has approximative solutions:
→ n-gram-Models $P(e_1 \ldots e_n) = \Pi P(e_i | e_{i-2}\ e_{i-1})$
→ Transfer of „phrases“ $P(F|E) = \Pi P(\mathbf{f}_i | \mathbf{e}_i) * P(d_i)$
→ Heuristic *(beam)* search

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

Statistical Machine Translation

schematic architecture



IBM Translation Models

- Brown et al. 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	.	.	.	.	.	.	.	.	.	.	####	.	.
?	.	.	.	.	.	.	.	.	.	.	.	.	*

IBM Translation Models

- 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.

IBM Translation Models

UNIVERSITÄT DES SAARLANDES

- 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

- 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

SMT: A Walkthrough

- Parallel text
  - Sentence segmentation and tokenization
  - Sentence alignment
  - Make sure you will have unseen test data
  - Word alignment
  - Phrasetable construction

- More text from target language
  - Stochastic (target) language model

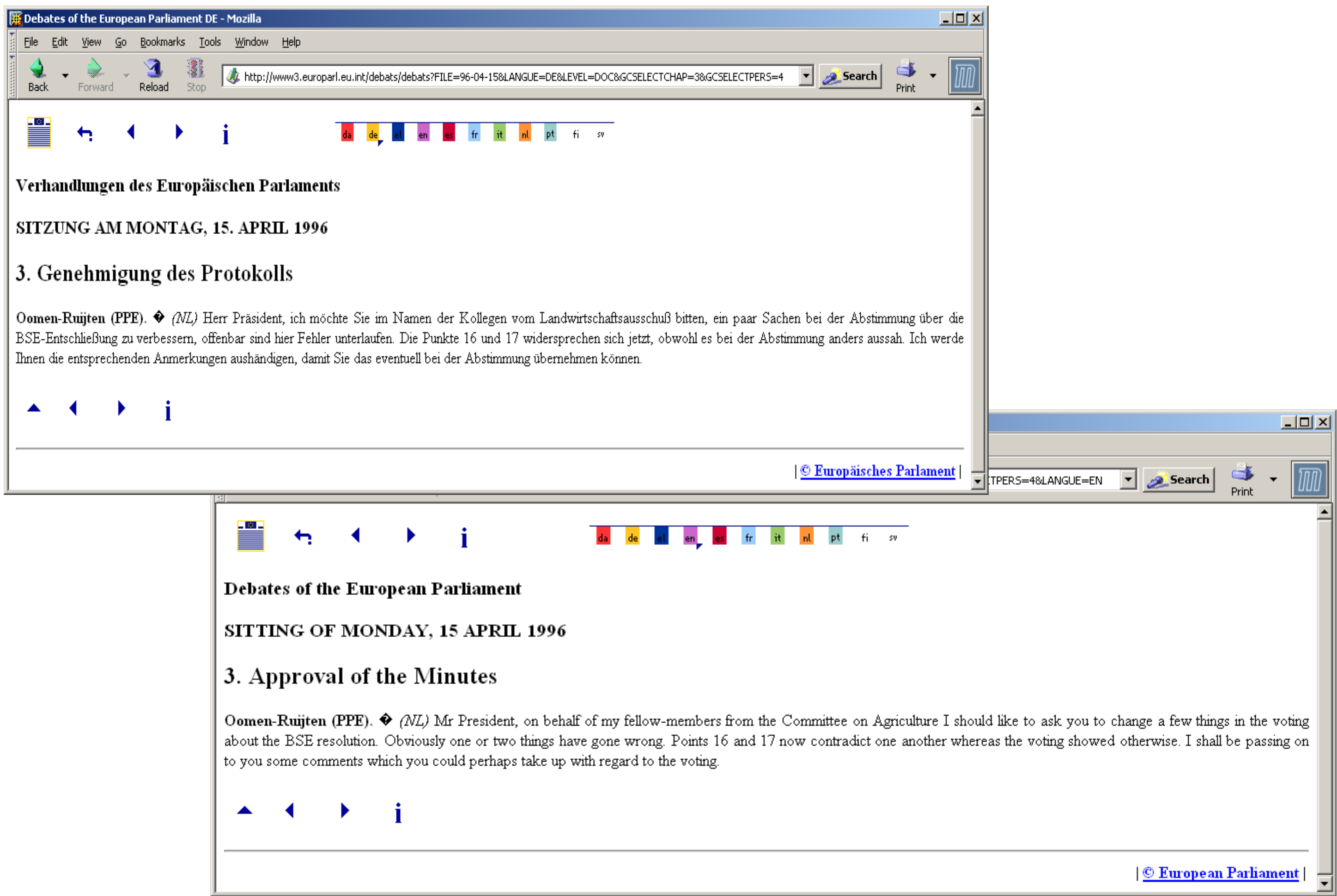
- Decoding
- Inspect/evaluate results

Parallel text

- **De-facto standard: EUROPARL corpus**
  - "Successor" of Canadian Hansards used by IBM
  - free, no legal constraints
  - current version includes 21 official EU languages
- **But:**
  - does not cover the most difficult/interesting languages (Chinese, Arabic, Japanese, Walpiri, Inuktitut, …)
  - not very technical
  - dependencies on context as in typical written text
- **In the meantime:**
  - EU has been extended to 27 states with 23 official languages
  - official law has been translated to all these languages

  ➔ "**Acquis Communautaire**" corpus

Parallel text: EUROPARL



Tokenization and sentence segmentation

- 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 are language-specific.

Sentence alignment

- Problem: During translation, sentences may have been split, merged, dropped or re-ordered.

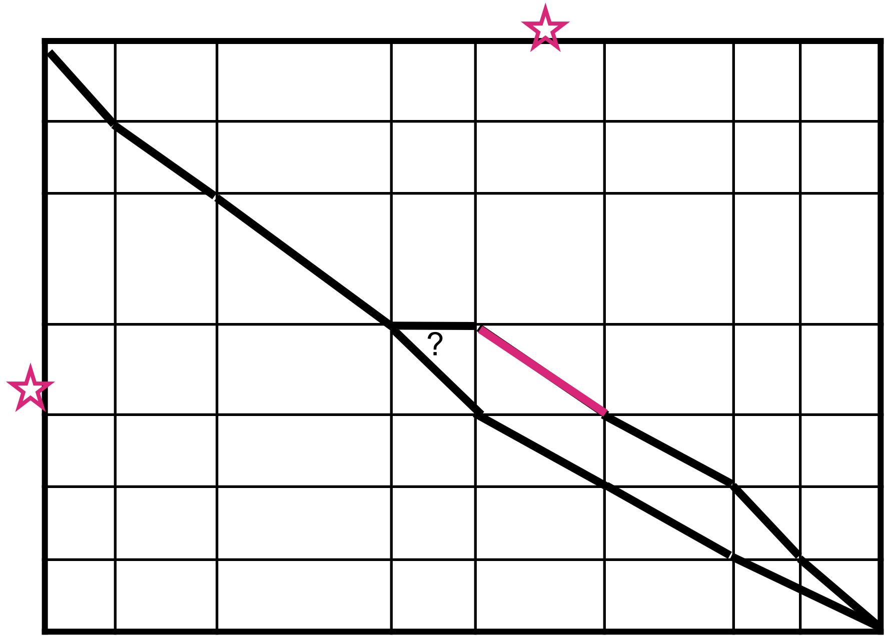
- If data is clean and some errors are acceptable: Simple length-based heuristic does the job

- Task can be seen as finding an optimal path through rectangular grid (see next slide)

- Europarl v.1: 10 sentence alignments XY ⇔ EN
- Europarl v.2ff: sentences + generic alignment tool

Sentence alignment

- Can be solved by dynamic programming



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

Word alignment

- The problem: We need to know alignments between texts and translations on word or phrase level

- This is more difficult as for sentences, as the order on both sides does not agree

- There is no a priory notion of similarity, possible correspondences need to be learned from data

Word alignment

UNIVERSITÄT DES SAARLANDES

- 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

Word alignment

- A typical solution:
  - Assume a probabilistic model for co-occurrences between words/phrases
  - Train parameters from data

- But we have a chicken-and-egg situation:
  - given alignments, we can learn the parameters
  - given parameters, we can estimate alignments
  - we don't know how to start

Expectation Maximization (EM)

- Similar situations are ubiquitous in learning stochastic models from raw data lacking annotation
  - There is a generic scheme for how to deal with this problem, called EM algorithm

- Basic idea:
  - Start with a simple model (e.g. a uniform probability distribution)
  - Estimate a probabilistic annotation
  - Train a model from this estimate
  - Iterate re-estimation until result is good enough

- Properties of EM:
  - Likelihood of model is guaranteed to increase in each iteration
  - EM hence converges towards a maximum likelihood estimate (MLE)
  - But this maximum is only local
  - (Even global) MLE need not be useful for unseen data, less iterations may give better models

IBM Model I

- Each word of the foreign sentence is generated/ explained by some English word
- There is no limitation on the number of foreign words a given English word may generate, these influences are seen as independent
- Word order is completely ignored (bag of word)
- These slightly unrealistic assumptions simplify the mathematical analysis tremendously: Given a model and a sentence pair (f,e), estimated counts for the events can be obtained in closed form.

IBM Model I

Joint Probability of alignment and translation:

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j}). \quad (5)$$

Probability of translation:

$$\Pr(\mathbf{f}|\mathbf{e}) = \frac{\epsilon}{(l+1)^m} \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}). \quad (6)$$

Can be reorganized into:

$$\sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(f_j|e_{a_j}) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i). \quad (15)$$

Counts for word-pair events can now be collected for foreign words, given bag of English words, but independent of foreign context

Simplified model for word alignment

- We will use a simplified version of IBM Model 1 (called Model 0), assuming that each word in a foreign language text f is the translation of (generated by) some word in the English version e

- Probability that the i-th foreign word f_i is generated, given an English sentence e, is modeled as:

$$P(f_i|e) = \sum_j P(f_i \mid e_j)$$

- Probability that the complete foreign sentence is generated (omitting some boring details):

$$P(f|e) = \prod_i P(f_i|e) = \prod_i \sum_j P(f_i \mid e_j)$$

EM algorithm for word alignment

- From a set of annotated data (i.e. sentence pairs with word alignments), we can obtain a new translation model:

$$P(f_i|e_j) = freq(f_i,e_j) / freq(e_j)$$

- From a model P, a foreign word f_i, and a sequence $e = e_1 \ldots e_n$ of possible “causes”, we can estimate frequencies as

$$freq(f_i|e_j) = P(f_i|e_j) / \sum_{k=1}^{n} P(f_i|e_k)$$

The training corpus and models

- Corpus:
  chien méchant ←→ dangerous dog
  petit chien ←→ small dog

- Initial model:
  - $p_0(f_i|e_j)$ = constant

- Update steps:
  - $P(f_i|e_j) = freq(f_i,e_j) / freq(e_j)$
  - $freq(f_i|e_j) = P(f_i|e_j) / \sum_{k=1}^{n} P(f_i|e_k)$

EM iteration 1

Local frequency estimates

$freq(f_i\|e_j)$	chien	méchant
dangerous	0.5	0.5
dog	0.5	0.5

$freq(f_i\|e_j)$	petit	chien
small	0.5	0.5
dog	0.5	0.5

Global frequencies and probabilities

$freq(f_i\|e_j)$	petit	chien	méchant
small	0.5	0.5	
dangerous		0.5	0.5
dog	0.5	1.0	0.5

$p(f_i\|e_j)$	petit	chien	méchant
small	0.5	0.5	
dangerous		0.5	0.5
dog	0.25	0.5	0.25

EM iteration 2

UNIVERSITÄT DES SAARLANDES

Probabilities from iteration 1

$p(f_i \mid e_j)$	petit	chien	méchant
small	0.5	0.5	
dangerous		0.5	0.5
dog	0.25	0.5	0.25

New frequency estimates

$freq(f_i \mid e_j)$	chien	méchant
dangerous	0.5	0.67
dog	0.5	0.33

$freq(f_i \mid e_j)$	petit	chien
small	0.67	0.5
dog	0.33	0.5

EM iteration 2

Local frequency estimates

$freq(f_i\|e_j)$	chien	méchant
dangerous	0.5	0.67
dog	0.5	0.33

$freq(f_i\|e_j)$	petit	chien
small	0.67	0.5
dog	0.33	0.5

Global frequencies and probabilities

$freq(f_i\|e_j)$	petit	chien	méchant
small	0.67	0.5	
dangerous		0.5	0.67
dog	0.33	1.0	0.33

$p(f_i\|e_j)$	petit	chien	méchant
small	0.57	0.43	
dangerous		0.43	0.57
dog	0.2	0.6	0.2

Word alignment

Sample from the DE⇔EN alignment:

Die$_0$ Punkte$_1$ 16$_2$ und$_3$ 17$_4$ widersprechen$_5$ sich$_6$ jetzt$_7$,$_8$ obwohl$_9$ es$_{10}$ bei$_{11}$ der$_{12}$ Abstimmung$_{13}$ anders$_{14}$ aussah$_{15}$.$_{16}$

Points$_0$ 16$_1$ and$_2$ 17$_3$ now$_4$ contradict$_5$ one$_6$ another$_7$ whereas$_8$ the$_9$ voting$_{10}$ showed$_{11}$ otherwise$_{12}$.$_{13}$

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

Word alignment

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
                                  .
```

Word alignment

UNIVERSITÄT DES SAARLANDES

- 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/`

Phrase-table construction

- Idea: collect pairs of substrings that are compatible with word alignment
- Phrasetable is annotated with scores that will be used during decoding
- Alternatively: in tree-based models we try to learn a grammar:
  - hierarchical: not based on any syntactic theory
  - syntax-based: needs annotated (=parsed) data

Phrase-table construction

```
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
```

Stochastic language model

- 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, 5-grams seem to be good enough.

Stochastic language model

- 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)
  - IRST-LM (FBK)

- Philipp Koehn's Moses decoder can make use of several different models; it comes with KenLM (Heafield)

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

Decoding

UNIVERSITÄT DES SAARLANDES

- The decoder…
  - uses source sentence *f* and phrase table to estimate *P(e|f)*
  - uses LM 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, tree-based) that will improve the quality