Christof Monz
Informatics Institute
University of Amsterdam

Applied Language Technology

Introduction to Statistical Machine Translation

Outline

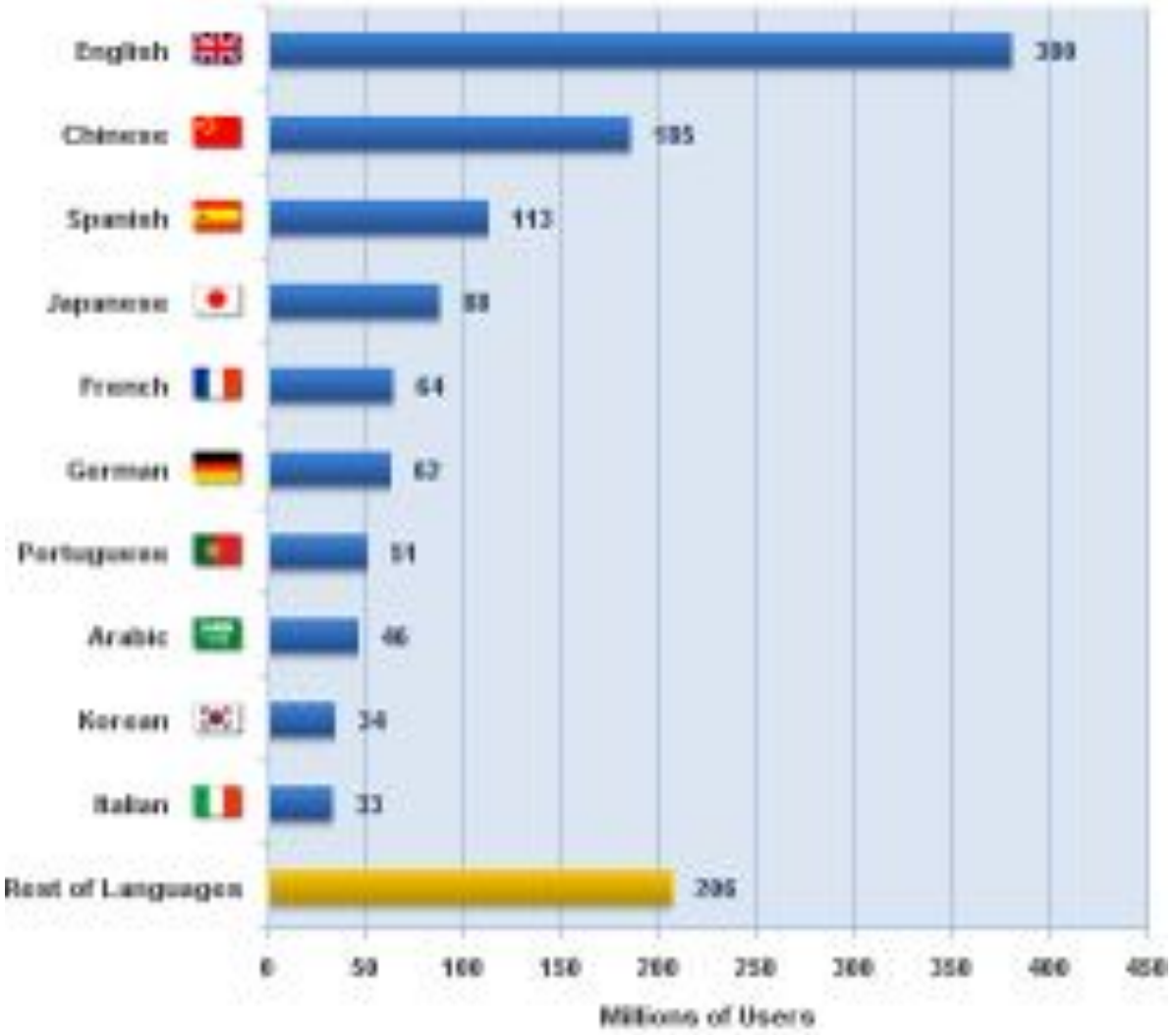
- What is machine translation about?
- What is the relevance of machine translation?
- Traditional approaches to machine translation
- The statistical turn
- The general architecture of statistical MT

What is Machine Translation?

- Simply put: The task of MT is to translate from one language into another
- Why? It's a globalizing world with more and more foreign language information being easily 'accessible' and relevant.
- This includes text-to-text, speech-to-speech, or any combination of these
- Doesn't seem to be too hard. We can translate between languages!

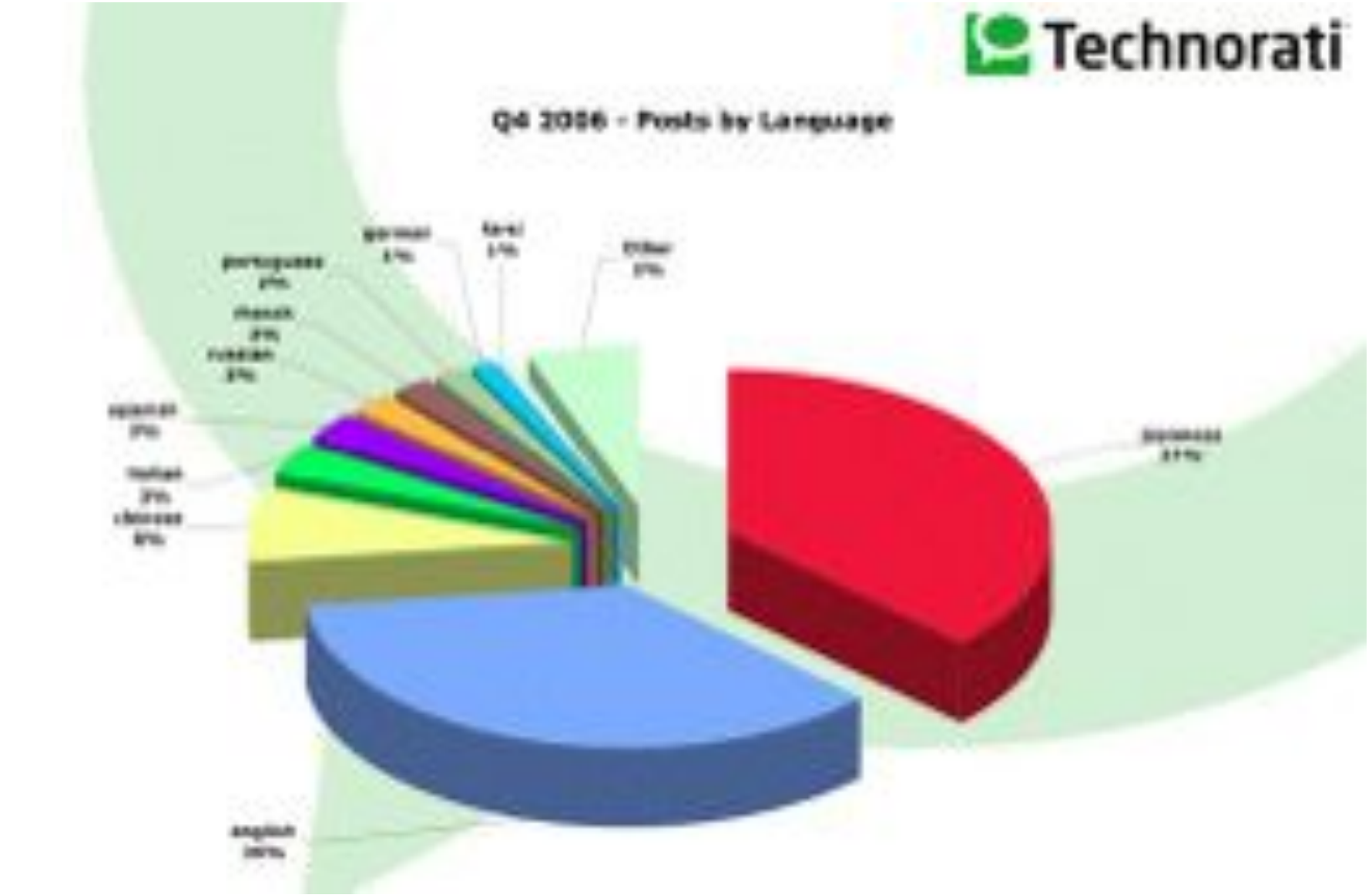
User Languages on the Web

Number of speakers (Nov 07)

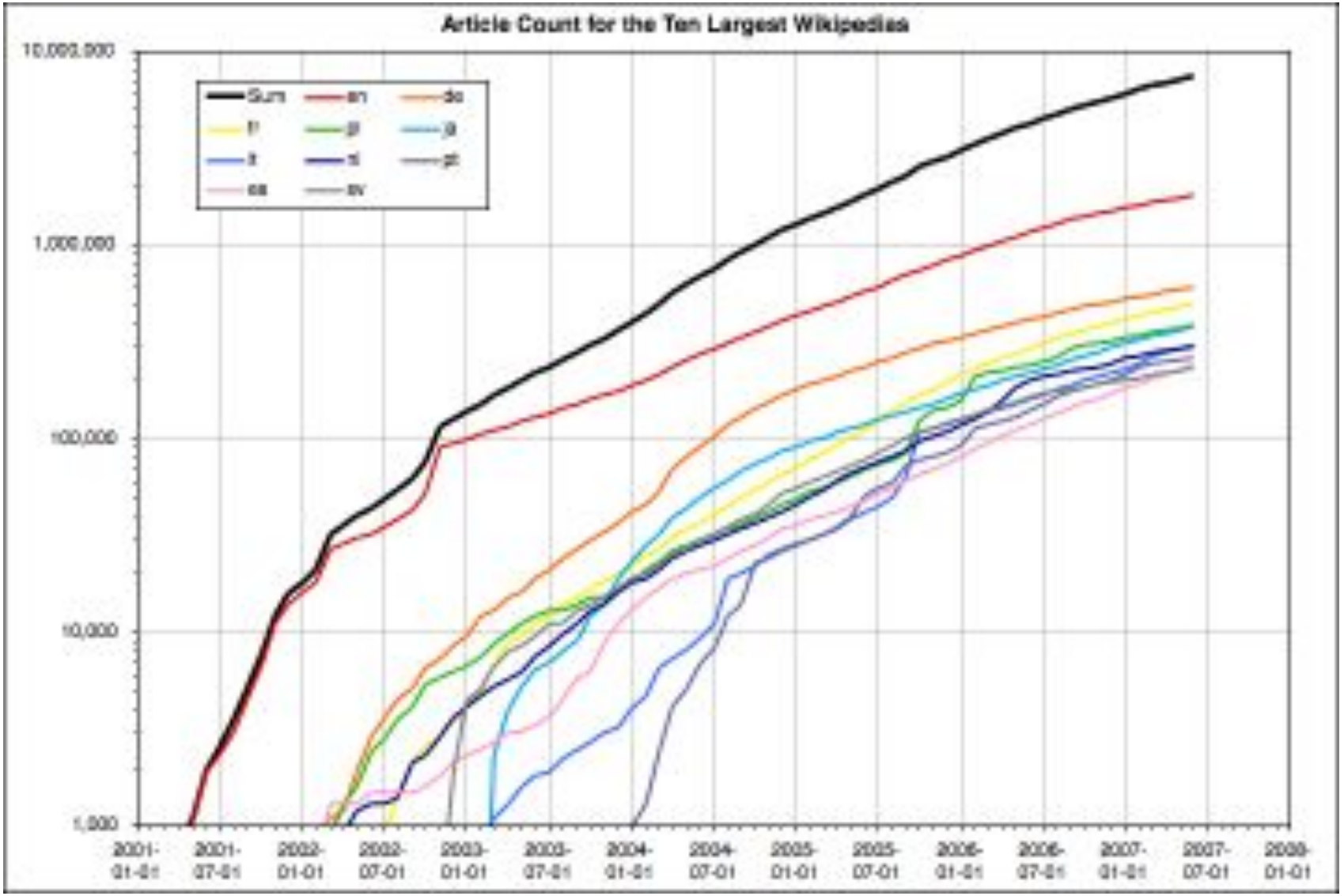


growth	(00–07)
167.3%	(en)
472.4%	(zh)
359.7%	(es)
85.9%	(ja)
422.7%	(fr)
123.5%	(de)
570.9%	(pt)
1,575.9%	(ar)
80.8%	(ko)
151.1%	(it)
534.8%	(rest)

Languages on Weblogs



Languages on Wikipedia



Languages and the Web

- Number of non-English Internet users grows at a faster pace
- Substantial amount of user-generated Web content is not in English
- English is losing its position as the lingua franca of the Web
- Language diversity can lead to a Balkanization of the Web

Breaking the Language-Barrier

- Expecting users to learn 10 foreign languages is unrealistic
- Need for automation
  - Tools that can translate between languages
  - Tools that can search data in foreign languages
  - Tools that can mine/analyze data in foreign languages
- Machine Translation (MT)
  - Infamous for its mistakes
  - Recent developments in statistical MT have led to great advances

Google and MT

- Google's mission is to 'organize the world's information and make it universally accessible and useful'
- Eric Schmidt (Google CEO): "*The most impressive products are those that use artificial intelligence that I cannot imagine are doable. Automatic language translation by computers...we will eventually do 100 languages.*"

Does it Work?

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

- **MT:** Baghdad 1-1 (AFP) - The official Iraqi news agency reported that the Chinese vice-president of the Revolutionary Command Council in Iraq, Izzat Ibrahim, met today in Baghdad, chairman of the Saudi Export Development Center, Abdel Rahman al-Zamil.

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

- **MT:** Baghdad 1-1 (AFP) - The official Iraqi news agency reported that the Chinese vice-president of the Revolutionary Command Council in Iraq, Izzat Ibrahim, met today in Baghdad, chairman of the Saudi Export Development Center, Abdel Rahman al-Zamil.

- **Human:** Baghdad 1-1 (AFP) - Iraq's official news agency reported that the Deputy Chairman of the Iraqi Revolutionary Command Council, Izzet Ibrahim, today met with Abdul Rahman al-Zamil, Managing Director of the Saudi Center for Export Development.

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

- **MT:** Baghdad 1-1 (AFP) - The official Iraqi news agency reported that the Chinese vice-president of the Revolutionary Command Council in Iraq, Izzat Ibrahim, met today in Baghdad, chairman of the Saudi Export Development Center, Abdel Rahman al-Zamil.

- **Human:** Baghdad 1-1 (AFP) - Iraq's official news agency reported that the Deputy Chairman of the Iraqi Revolutionary Command Council, Izzet Ibrahim, today met with Abdul Rahman al-Zamil, Managing Director of the Saudi Center for Export Development.

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

- **MT:** Baghdad 1-1 (AFP) - The official Iraqi news agency reported that the Chinese vice-president of the Revolutionary Command Council in Iraq, Izzat Ibrahim, met today in Baghdad, chairman of the Saudi Export Development Center, Abdel Rahman al-Zamil.

- **Human:** Baghdad 1-1 (AFP) - Iraq's official news agency reported that the Deputy Chairman of the Iraqi Revolutionary Command Council, Izzet Ibrahim, today met with Abdul Rahman al-Zamil, Managing Director of the Saudi Center for Export Development.

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

- **MT:** Baghdad 1-1 (AFP) - The official Iraqi news agency reported that the Chinese vice-president of the Revolutionary Command Council in Iraq, Izzat Ibrahim, met today in Baghdad, chairman of the Saudi Export Development Center, Abdel Rahman al-Zamil.

- **Human:** Baghdad 1-1 (AFP) - Iraq's official news agency reported that the Deputy Chairman of the Iraqi Revolutionary Command Council, Izzet Ibrahim, today met with Abdul Rahman al-Zamil, Managing Director of the Saudi Center for Export Development.

Does it Work?

- بغداد ١-١ (اف ب) ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد رئيس مجلس ادارة المركز السعودي لتطوير الصادرات عبد الرحمن الزامل.

- **MT:** Baghdad 1-1 (AFP) - The official Iraqi news agency reported that the Chinese vice-president of the Revolutionary Command Council in Iraq, Izzat Ibrahim, met today in Baghdad, chairman of the Saudi Export Development Center, Abdel Rahman al-Zamil.

- **Human:** Baghdad 1-1 (AFP) - Iraq's official news agency reported that the Deputy Chairman of the Iraqi Revolutionary Command Council, Izzet Ibrahim, today met with Abdul Rahman al-Zamil, Managing Director of the Saudi Center for Export Development.

How would you do it?

- Intuition:

How would you do it?

- Intuition:
  1. Analyze the source sentence
  2. Build an abstract language-independent representation
  3. Generate the target sentence from this representation

How would you do it?

- Intuition:
  1. Analyze the source sentence
  2. Build an abstract language-independent representation
  3. Generate the target sentence from this representation
- That's what is commonly referred to as the **interlingua-based** or **transfer-based** approach to MT
- One abstract interlingua that can represent the meaning of any sentence in any language
- Think of a mixture of logic and Esperanto

A bit of MT History

> It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

Warren Weaver, 1949

Weaver's 'Prophecy'

- Weaver's position emerged from war-driven (WWII and cold war) emphasis on cryptanalysis
- His ideas rely on 'cracking' the translation problem, i.e. a combination of statistical models and brute-force computing
- Does it actually work? (We come back to that later)

MT in the 50s and 60s

- Great advances in linguistics (Chomsky!), both syntactic and semantic theories emerged
- Several large-scale MT efforts were undertaken, but most worked either
  - well on very limited domains
  - or poorly on general domains
- Emphasis on Russian-English as most MT research was (and still is) funded by defense departments

The ALPAC Report

- The Automatic Language Processing Advisory Committee (ALPAC) report was commissioned by the US govt
- Outcomes
  - Very sceptical of the claims made by the MT community
  - There are more than enough human translators that can do it better and cheaper
- Consequence: Funding for MT came to a halt

The Ice Age of MT

- Research on MT continued during the 70s and 80s
  - mostly outside of the US
  - at a lower pace
  - focusing on commercial needs
- Most MT research focused on rule-based MT (continuing the trend from the 50s and 60s)
- Some of these systems are still under development (e.g., Systran)

Rule-Based MT

- Based on the idea of interlingua-based MT but not quite as strict
- The intermediate representations are somewhat language dependent (source and target)
- Translation rules (or transfer rules) are used to map between source and target structures
- Transfer rules are particularly suited for handling complex translation mappings

Rule-Based MT

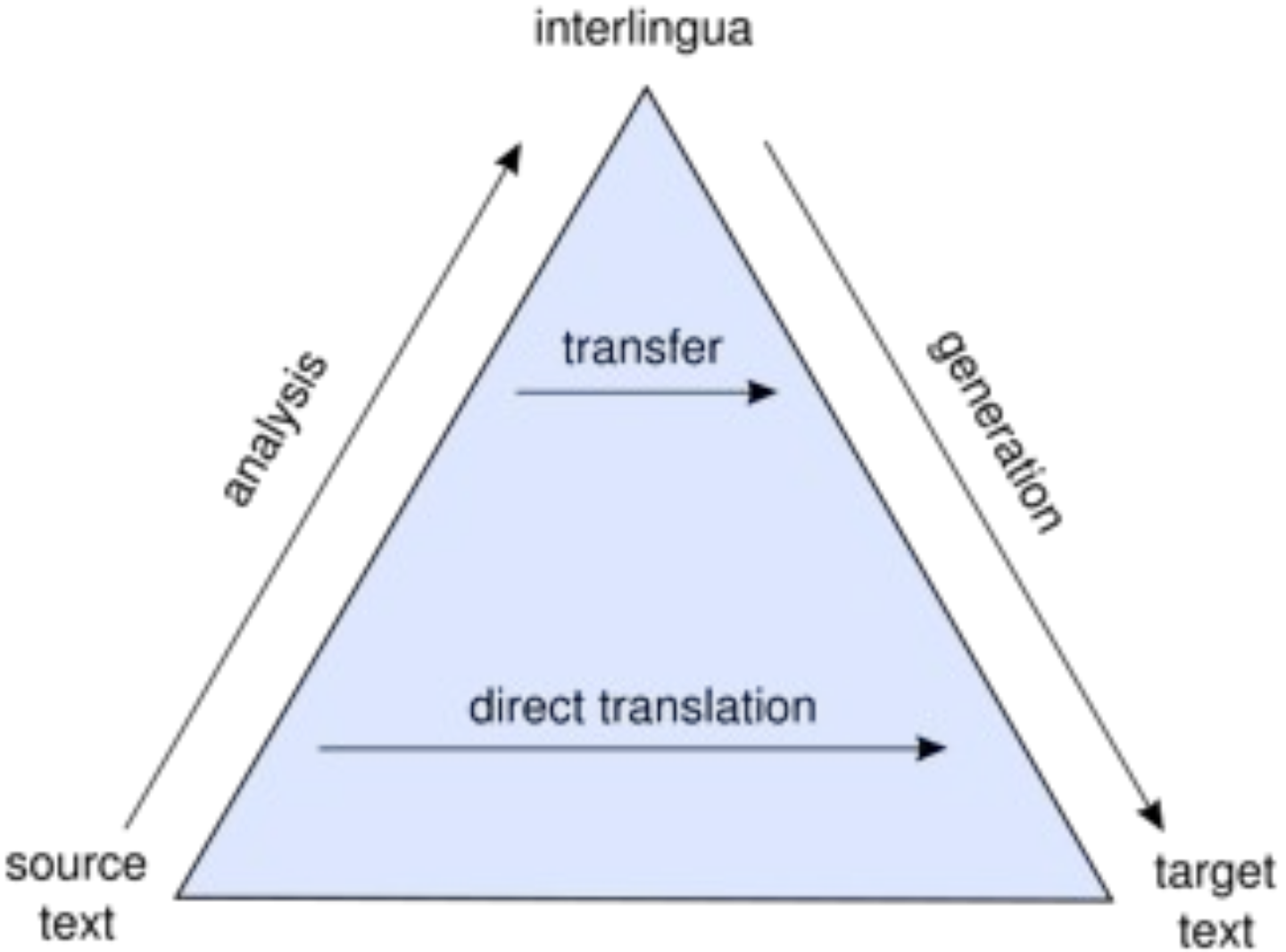
- Translation divergences (head swapping example)
  - Spanish: *X cruzar Y nadando* (X cross Y swimming)
  - English: *X swim across Y*
- Transfer rule (Dorr and Habash):

```
@TRANS_CORR
  @EN V1 [cat:verb manner:M]
    (ATTR Y [cat:prep path:P event:go] (II N))
  @SP V2 [cat:verb path:P event:go]
    (II N ATTR Z [manner:M])
```

Rule-Based MT

- Very accurate for certain phenomena
- Writing the transfer rules is very laborious
- Transfer rules tend to be language specific (at least wrt language families)
- Transfer-based MT requires ‘deep’ analysis such as (semantic) parsing
- What happens if multiple transfer rules can be applied to overlapping structures?

The MT Pyramid



The Statistical Turn

- 50 years after Weaver's comments, things look more promising for statistical MT
  - Computing power has increased dramatically
  - There's an abundance of data
- First attempt: IBM Models (early 90s)
- Goal: Find automatically word translations from training data (parallel corpora)
- There's no linguistics (no syntax, no semantics) involved

Parallel Corpus

⋮	⋮
李鹏会见新加坡前总统王鼎昌	Li Peng Meets With Former Singapore President Ong Teng Cheong
马来亚、新加坡、沙捞越、沙巴和文莱曾组成联邦,但最後分裂了。	Malaysia, Singapore, Sarawak, Sabah and Brunei once formed a federation, but it also fell apart in the end.
新加坡排行榜首,缅甸则排行榜尾 。	Singapore is at the head of the list, while Burma ranks last.
新加坡则在致力建造一个光纤网环绕的 “智能岛” 。	Singapore is also devoting itself to building a "intelligence island" embraced by a fiber-optical net.
⋮	⋮

Parallel Corpus

⋮	⋮
李鹏会见新加坡前总统王鼎昌	Li Peng Meets With Former Singapore President Ong Teng Cheong
马来亚、新加坡、沙捞越、沙巴和文莱曾组成联邦,但最後分裂了。	Malaysia, Singapore, Sarawak, Sabah and Brunei once formed a federation, but it also fell apart in the end.
新加坡排行榜首,缅甸则排行榜尾 。	Singapore is at the head of the list, while Burma ranks last.
新加坡则在致力建造一个光纤网环绕的 “智能岛” 。	Singapore is also devoting itself to building a "intelligence island" embraced by a fiber-optical net.
⋮	⋮

Architecture of Statistical MT

- Analyzing the parallel corpus yields a (probabilistic) lexicon
- There are still several things missing:
  - How do we select the 'best' translation for a word (in a given context)?
  - How do we map between different word-orderings? *Peter hat Anna gesehen* should not be translated as *Peter has Anna seen*
- Additional information (models) is needed to distinguish 'good' English from 'bad' English

Source Channel Model in SMT

- Given a sentence f in a foreign language, we want to find the English sentence e, such that

$$\underset{e}{\operatorname{argmax}}\ p(e|f)$$

Source Channel Model in SMT

- Given a sentence f in a foreign language, we want to find the English sentence e, such that

$$\underset{e}{\operatorname{argmax}}\ p(e|f)$$

- Normally re-formulated by applying Bayes' theorem:

$$\underset{e}{\operatorname{argmax}}\ \frac{p(f|e) \cdot p(e)}{p(f)}$$

Source Channel Model in SMT

- Given a sentence f in a foreign language, we want to find the English sentence e, such that

$$\underset{e}{\operatorname{argmax}}\ p(e|f)$$

- Normally re-formulated by applying Bayes' theorem:

$$\underset{e}{\operatorname{argmax}}\ \frac{p(f|e) \cdot p(e)}{\cancel{p(f)}}$$

Source Channel Model in SMT

- Given a sentence f in a foreign language, we want to find the English sentence e, such that

$$\underset{e}{\operatorname{argmax}}\ p(e|f)$$

- Normally re-formulated by applying Bayes' theorem:

$$\underset{e}{\operatorname{argmax}}\ p(f|e) \cdot p(e)$$

Source Channel Model in SMT

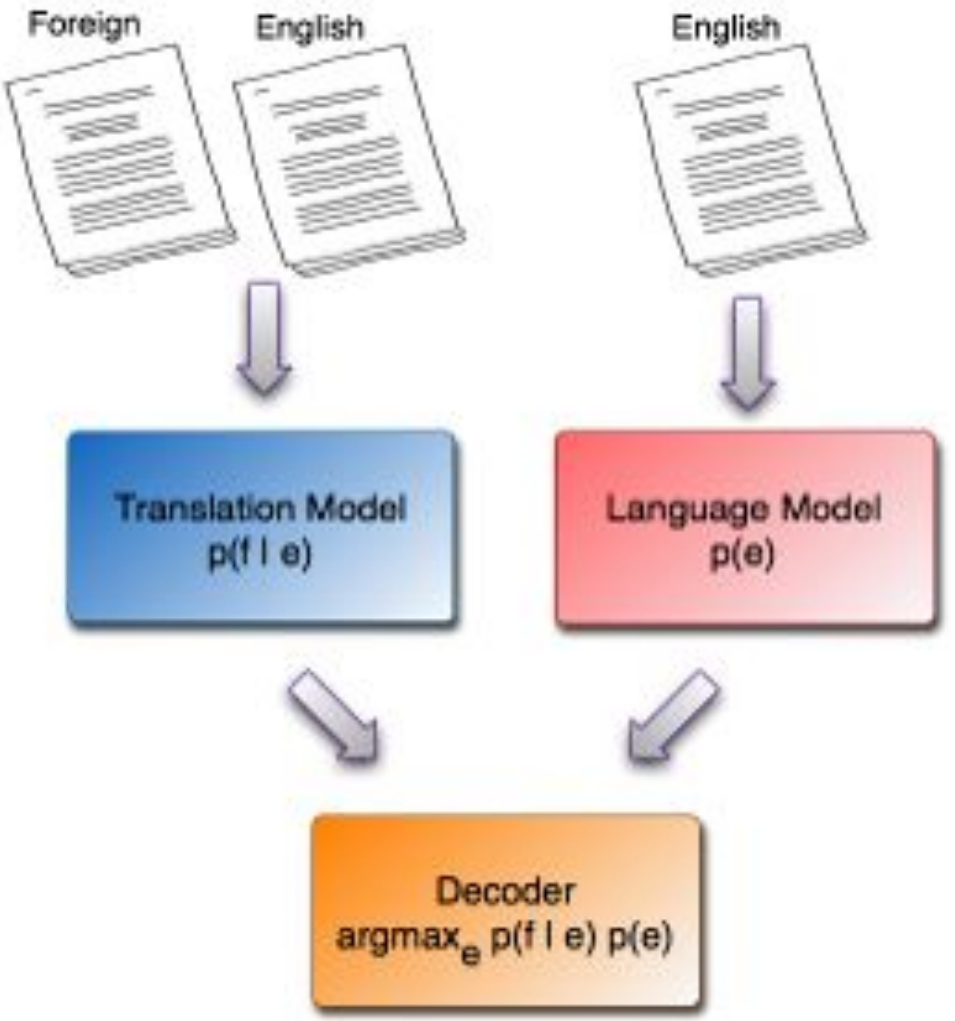
- Given a sentence f in a foreign language, we want to find the English sentence e, such that

$$\underset{e}{\operatorname{argmax}}\ p(e|f)$$

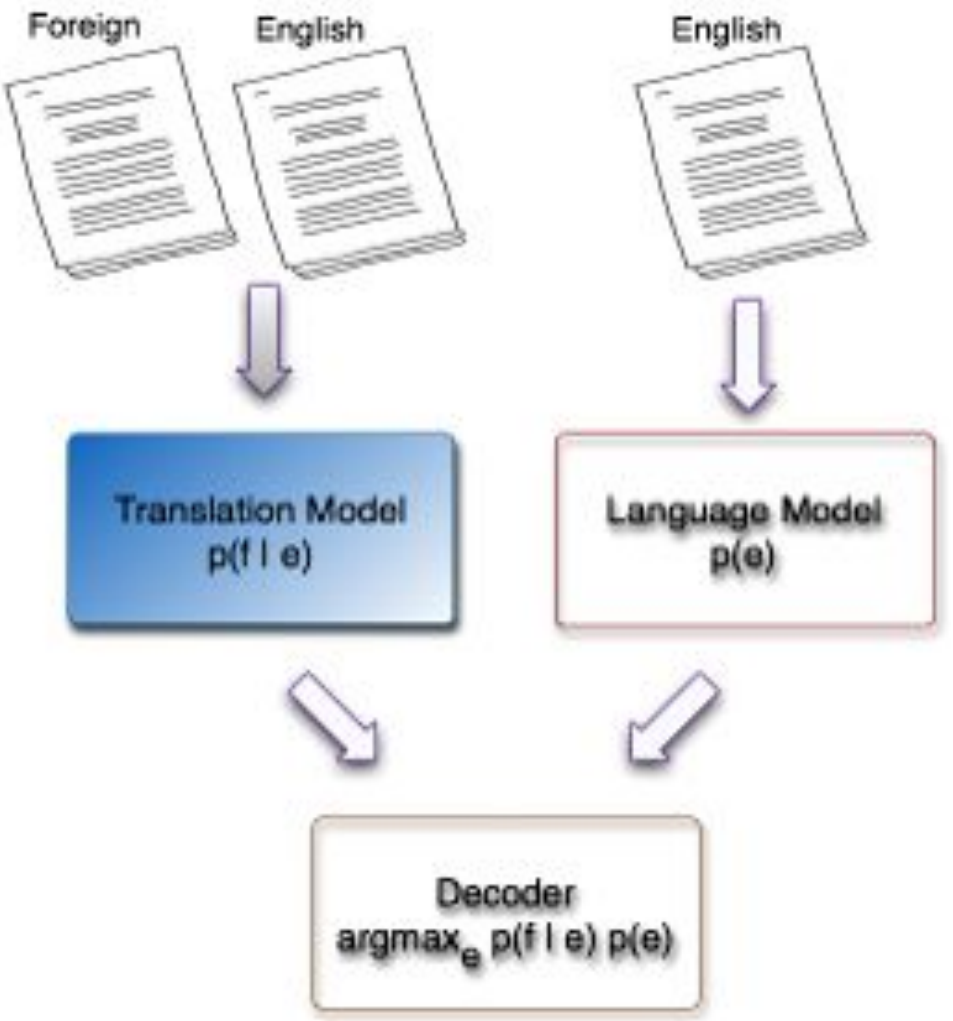
- Normally re-formulated by applying Bayes' theorem:

$$\underset{e}{\operatorname{argmax}}\ \underbrace{p(f|e)}_{\text{TransModel}} \cdot \underbrace{p(e)}_{\text{LangModel}}$$

General SMT Architecture



General SMT Architecture



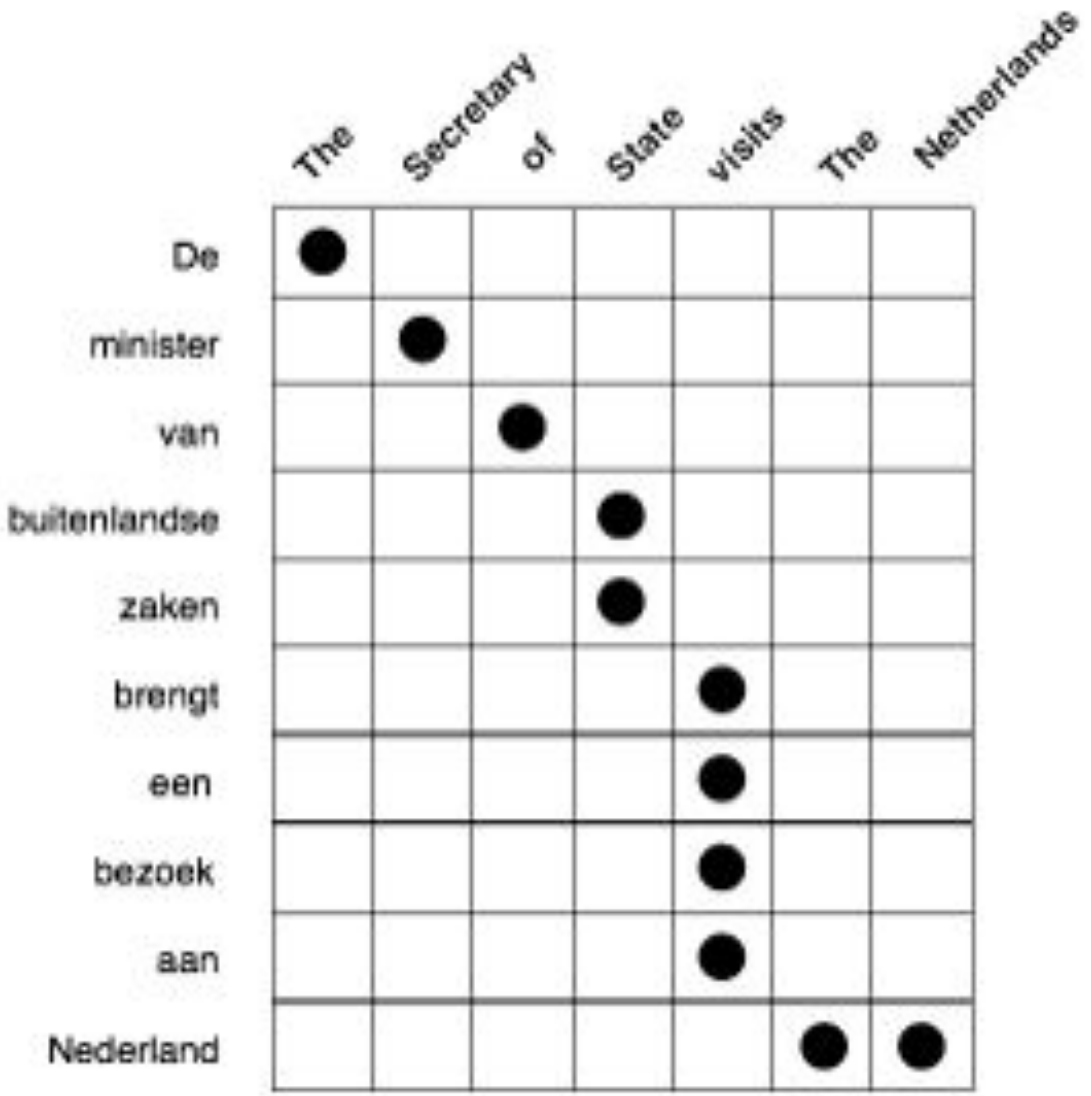
Translation Modeling

- The purpose is list possible translations for a foreign word or phrase
  - Machine-readable human-produced dictionaries
  - Statistically induced 'dictionaries'
- Human dictionaries
  - Clean
  - Limited scope (e.g., names)
  - No ranking/scoring of translations
- Statistical dictionaries
  - Noisy
  - Broader scope (depending on the training data)
  - Scores/probabilities associated with translation candidates (reliability?)

Translation Modeling

- Statistial translation modeling relies on co-occurrence counts
- Common co-occurrence metrics include
  - point-wise mutual information
  - log-likelihood ratios
  - IBM models
- In SMT: IBM Models (plus HMM)
- Co-occurence counts as alignment
  - Given a parallel sentence pair (f,e) find the most likely alignment links between the elements (words) of f and e
  - $\rightarrow$ class on word alignment

Translation Modeling

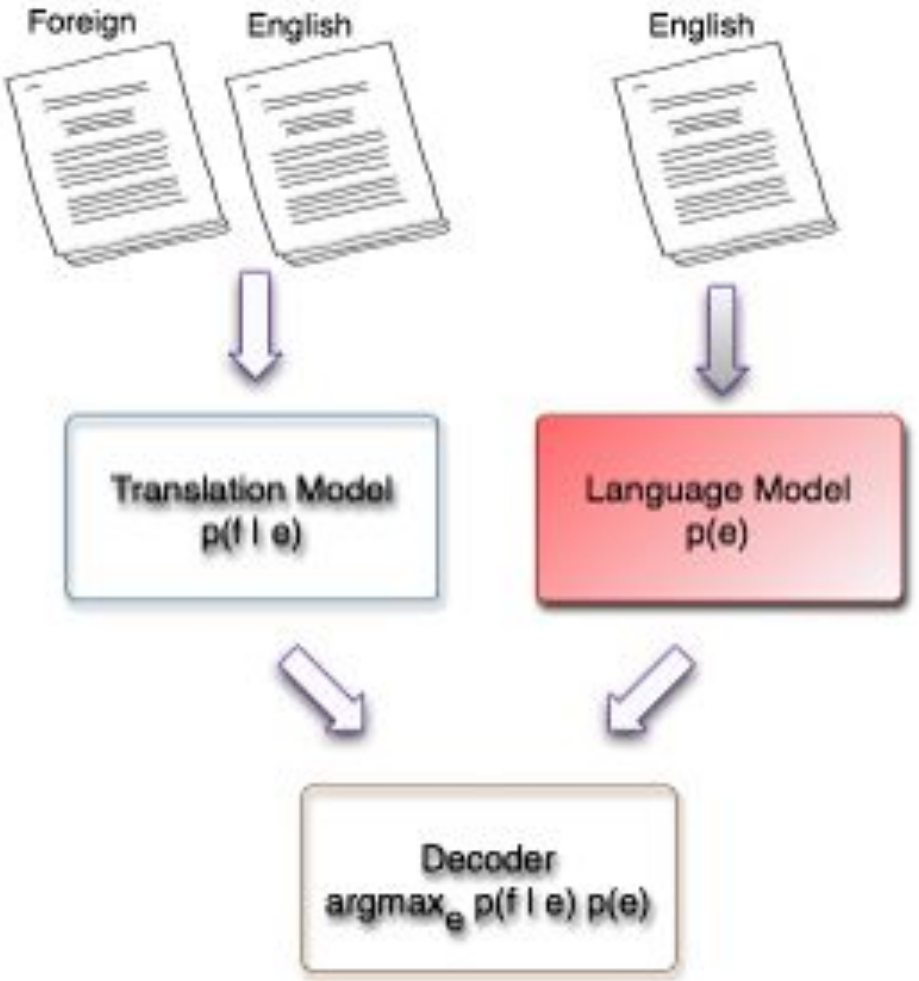


Translation Modeling

- From (aligned) words to phrases
- Current SMT systems use continuous or discontinuous phrases
  - Based on the word alignments
  - Phrases that are consistent with the word alignment are extracted
  - Compute $p(e|f)$ for all phrases e and f using maximum likelihood estimation, e.g.

    $p(\text{State} \mid \text{buitenlandse zaken}) = 0.5$
    $p(\text{foreign affairs} \mid \text{buitenlandse zaken}) = 0.3$
    $p(\text{foreign policy} \mid \text{buitenlandse zaken}) = 0.1$
    $p(\text{StateDepartment} \mid \text{buitenlandse zak.}) = 0.05$
    $p(\text{foreign} \mid \text{buitenlandse zaken}) = 0.05$

Language Models



Language Models

- The purpose of a language model is to estimate the probability of a string
- It is assumed that well-formed strings are more probable than ill-formed strings

Language Models

- The purpose of a language model is to estimate the probability of a string
- It is assumed that well-formed strings are more probable than ill-formed strings
- There are a number of language model approaches:
  - N-gram models (character- or word-based)
  - Structured language models
  - Syntax-based models

String probabilities

$$p(w_1 \ldots w_n) = p(w_1) \cdot \prod_{i=2}^{n} p(w_i | w_1 \ldots w_{i-1})$$

String probabilities

$$
\begin{aligned}
p(w_1 \ldots w_n) &= p(w_1) \cdot \prod_{i=2}^{n} p(w_i | w_1 \ldots w_{i-1}) \\
&\propto p(w_1) \cdot p(w_2 | w_1) \cdot \prod_{i=3}^{n} p(w_i | w_{i-2} w_{i-1})
\end{aligned}
$$

- *Markov assumption*: Shortening the history does not lead to decrease in performance

Tri-Gram Language Model

- Count the occurrences of all one, two, and three words appearing in sequence
- *<s> The Secretary of State visits The Netherlands </s>*

Tri-Gram Language Model

- Count the occurrences of all one, two, and three words appearing in sequence
- *<s> The Secretary of State visits The Netherlands </s>*
  - Uni-grams: Netherlands, of, <s>, </s>, Secretary, The (2×), visits

Tri-Gram Language Model

- Count the occurrences of all one, two, and three words appearing in sequence
- *<s> The Secretary of State visits The Netherlands </s>*
  - Uni-grams: Netherlands, of, <s>, </s>, Secretary, The (2×), visits
  - Bi-grams: Netherlands </s>, <s> The, State visits, Secretary of, The Netherlands, The Secretary, visits The

Tri-Gram Language Model

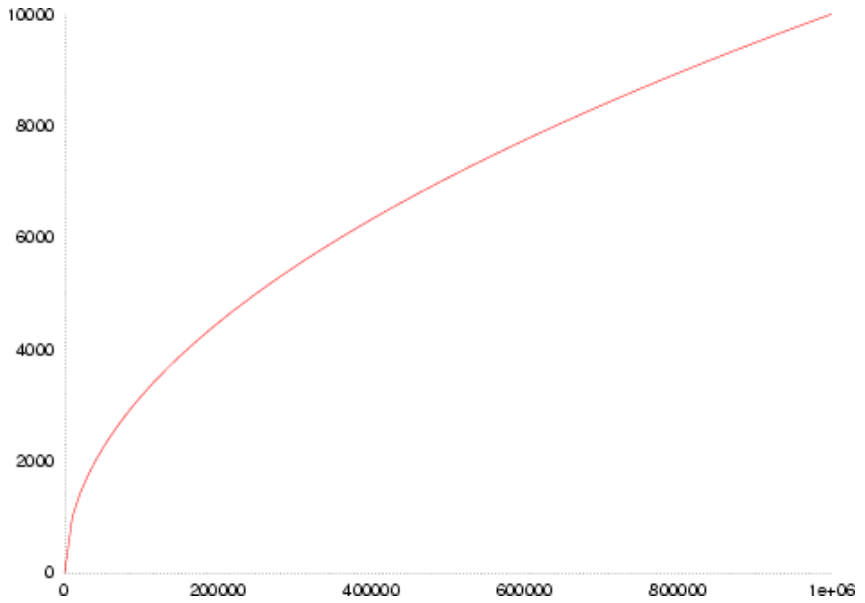
- Count the occurrences of all one, two, and three words appearing in sequence
- *<s> The Secretary of State visits The Netherlands </s>*
  - Uni-grams: Netherlands, of, <s>, </s>, Secretary, The (2×), visits
  - Bi-grams: Netherlands </s>, <s> The, State visits, Secretary of, The Netherlands, The Secretary, visits The
  - Tri-grams: <s> The Netherlands, State visits The, Secretary of State, The Netherlands </s>, The Secretary of, visits The Netherlands

Tri-Gram Language Model

- Count the occurrences of all one, two, and three words appearing in sequence
- *<s> The Secretary of State visits The Netherlands </s>*
  - Uni-grams: Netherlands, of, <s>, </s>, Secretary, The (2×), visits
  - Bi-grams: Netherlands </s>, <s> The, State visits, Secretary of, The Netherlands, The Secretary, visits The
  - Tri-grams: <s> The Netherlands, State visits The, Secretary of State, The Netherlands </s>, The Secretary of, visits The Netherlands
- $p(\text{The}|\text{<s>}) \cdot p(\text{Secretary}|\text{<s> The}) \cdot p(\text{of}|\text{The Secretary}) \cdot \ldots \cdot p(\text{Netherlands}|\text{visits The}) \cdot p(\text{</s>}|\text{The Netherlands})$

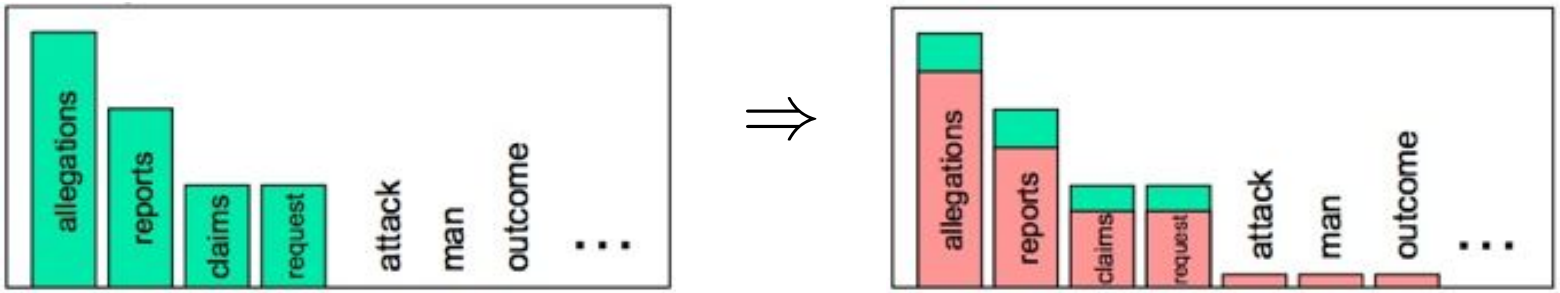
Data Sparseness

- The problem with n-grams is that they do not generalize (they are fully lexicalized)
- New words/n-grams appear all the time (Heap's law):



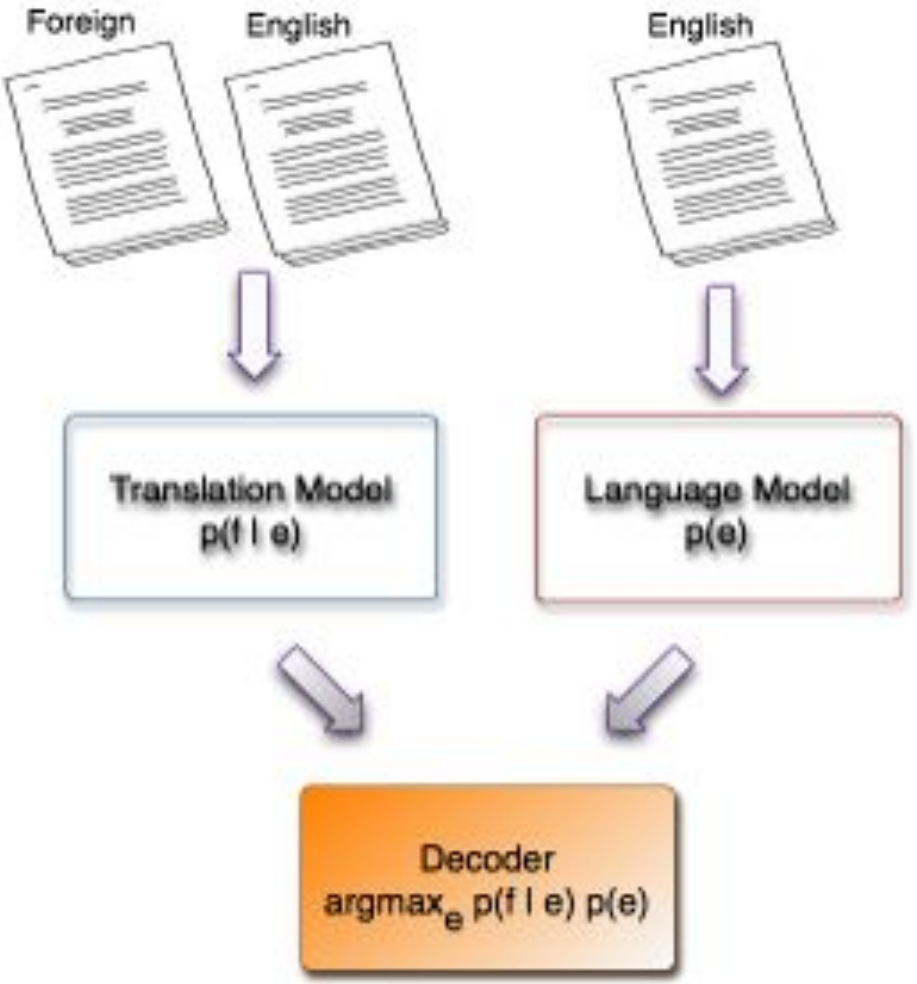
Smoothing

- If we haven't seen x (in the training data) it must be wrong, i.e. $p(x) = 0$
- As each training set is only a *sample*, this closed-world assumption can't hold
- Solution: set aside during training some of the available probability mass for *unseen* events



...the smoothed distribution becomes flatter

Decoding


Foreign
English
English
Translation Model
p(f | e)
Language Model
p(e)
Decoder
argmax_e p(f | e) p(e)

Phrase-Based MT Model

- The main parameters of the translation model are:
  - Phrase translation: $p(\bar{f}|\bar{e})$
  - Language model: $p(e)$
  - Re-ordering model: simple distance-based model
- Sentence f is segmented into I phrases: $\bar{f}_1^I = \bar{f}_1 \ldots \bar{f}^I$
- $p(\bar{e}_1^I|\bar{f}_1^I) = p(e) \prod_{i=1}^{I} p(\bar{f}_i|\bar{e}_i)\omega^{|\text{start}_i - \text{end}_{i-1} - 1|}$

Decoding

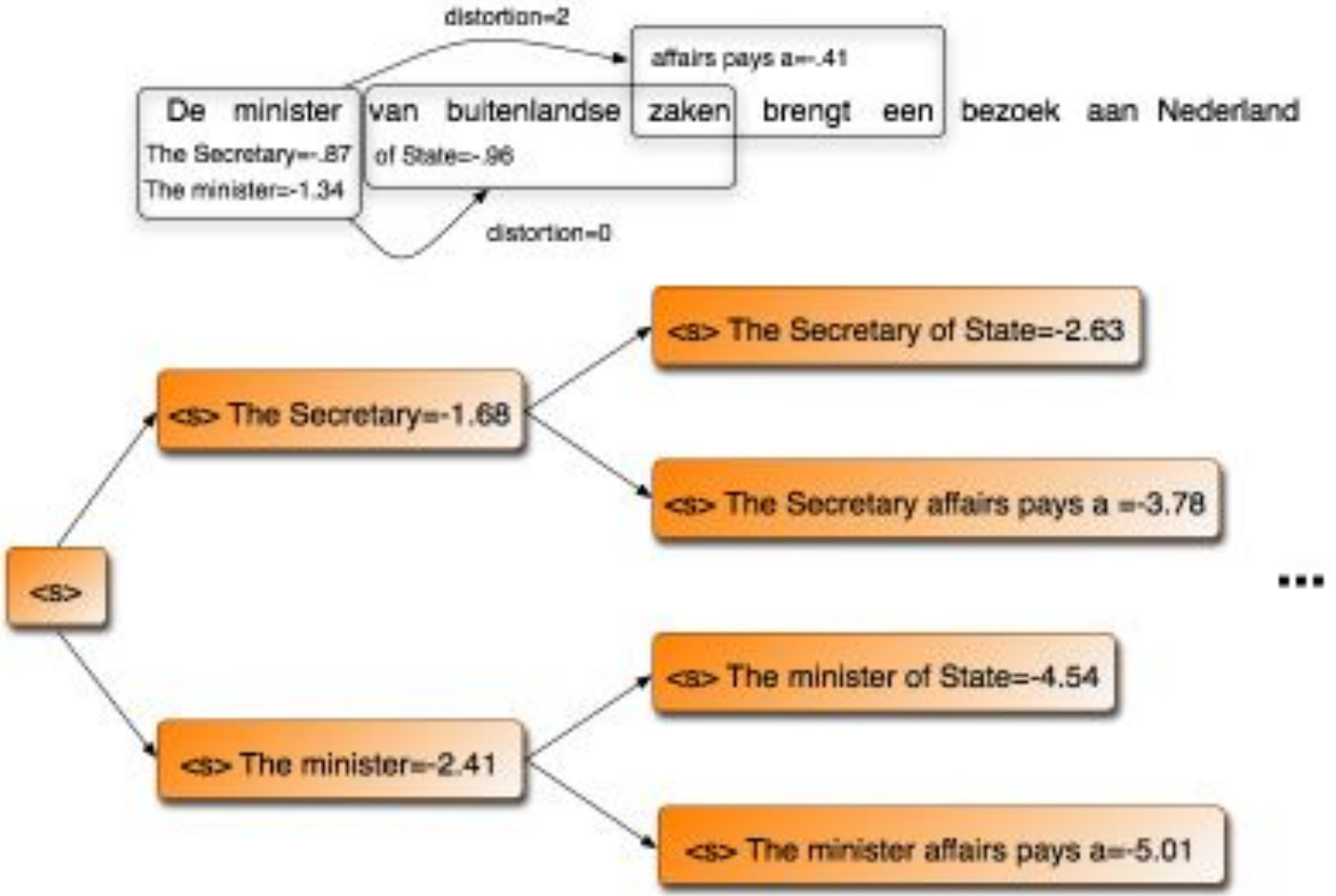
- The decoder searches for the best (i.e. most probable) translation according to our translation and language model
- Several standard search procedures are used:
  - A* search
  - Dynamic programming
- Since exploring the entire search space is not feasible, constraints are imposed that ignore search paths that are unlikely to result in good translations

Decoding in a Nutshell

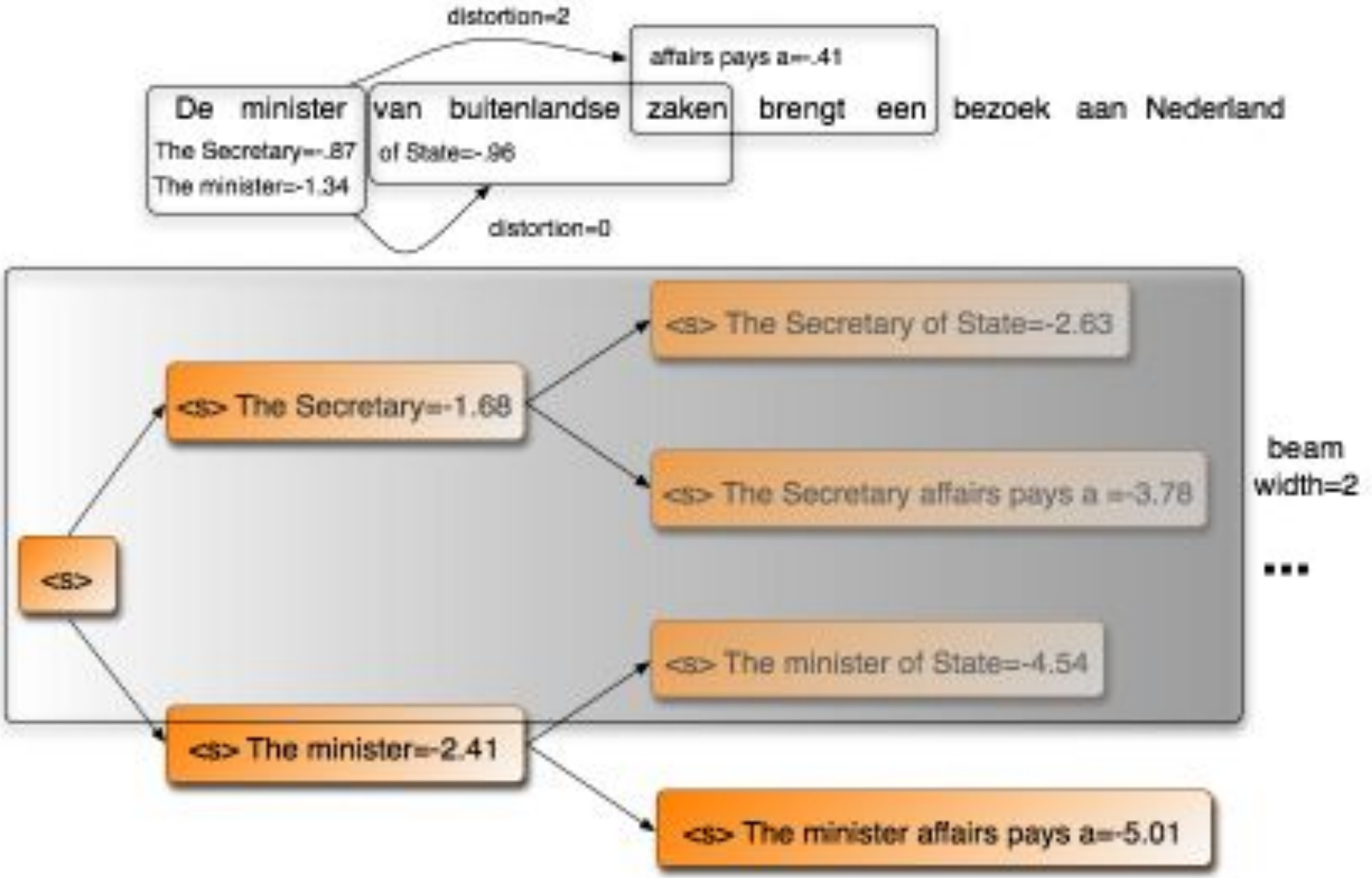
De minister van buitenlandse zaken brengt een bezoek aan Nederland

Decoding in a Nutshell

Decoding in a Nutshell



Decoding in a Nutshell



A* Search

- General search strategy using best-first graph search
- The cost of a hypothesis (state) $f(x) = g(x) + h(x)$ consists of two parts:
  - g(x): Costs that have been incurred so far to reach x
  - h(x): Estimate of the costs that will be incurred to reach the goal (future costs)
- A* is optimal if h never overestimates the actual costs of reaching the goal
- It is desirable that h can be computed efficiently

Evaluation of MT Quality

Evaluation of MT Quality

- Human evaluation of MT output rates it on two 5-point scales:
  - **Adequacy:** How much content of the original source sentence is captured by the translation?
  - **Fluency:** How readable (i.e. fluent) is the translation
- Although reliable, human evaluation is expensive and time consuming (inappropriate for rapid system development)
- Some inter-assessor disagreement, but that can be dealt with by using large test sets

Adequacy

- Example 1
  - *NL* In het café aan het Poolsterplein werd een talentenjacht gehouden waar veel publiek bij aanwezig was. Rond één uur stond een man op en begon plotseling te schieten op de bezoekers.
  - *EN* In the cafe at the North Square was a talent contest held where audience was present. Around an hour and was a man suddenly began to shoot at the visitors.
- Example 2
  - *NL* De drie gewonden zijn opgenomen in het ziekenhuis. De schutter zit vast.
  - *EN* The three injured in the hospital. The man is stuck.

Adequacy

- Example 1
  - *NL* In het café aan het Poolsterplein werd een talentenjacht gehouden waar veel publiek bij aanwezig was. Rond één uur stond een man op en begon plotseling te schieten op de bezoekers.
  - *EN* In the cafe at the North Square was a talent contest held where audience was present. Around an hour and was a man suddenly began to shoot at the visitors.
- Example 2
  - *NL* De drie gewonden zijn opgenomen in het ziekenhuis. De schutter zit vast.
  - *EN* The three injured in the hospital. The man is stuck.

Adequacy

- Example 1

  NL In het café aan het Poolsterplein werd een talentenjacht gehouden waar veel publiek bij aanwezig was. Rond één uur stond een man op en begon plotseling te schieten op de bezoekers.

  EN In the cafe at the North Square was a talent contest held where audience was present. Around an hour and was a man suddenly began to shoot at the visitors.

- Example 2

  NL De drie gewonden zijn opgenomen in het ziekenhuis. De schutter zit vast.

  EN The three injured in the hospital. The man is stuck.

Adequacy

- Example 1

  NL In het café aan het Poolsterplein werd een talentenjacht gehouden waar veel publiek bij aanwezig was. Rond één uur stond een man op en begon plotseling te schieten op de bezoekers.

  EN In the cafe at the North Square was a talent contest held where audience was present. Around an hour and was a man suddenly began to shoot at the visitors.

- Example 2

  NL De drie gewonden zijn opgenomen in het ziekenhuis. De schutter zit vast.

  EN The three injured in the hospital. The man is stuck.

Adequacy

- Example 1

  NL In het café aan het Poolsterplein werd een talentenjacht gehouden waar veel publiek bij aanwezig was. Rond één uur stond een man op en begon plotseling te schieten op de bezoekers.

  EN In the cafe at the North Square was a talent contest held where audience was present. Around an hour and was a man suddenly began to shoot at the visitors.

- Example 2

  NL De drie gewonden zijn opgenomen in het ziekenhuis. De schutter zit vast.

  EN The three injured in the hospital. The man is stuck.

Adequacy

- Example 1

  NL In het café aan het Poolsterplein werd een talentenjacht gehouden waar veel publiek bij aanwezig was. Rond één uur stond een man op en begon plotseling te schieten op de bezoekers.

  EN In the cafe at the North Square was a talent contest held where audience was present. Around an hour and was a man suddenly began to shoot at the visitors.

- Example 2

  NL De drie gewonden zijn opgenomen in het ziekenhuis. De schutter zit vast.

  EN The three injured in the hospital. The man is stuck.

Fluency

- Example 1
  - *NL* Elders in de stad werden hotels geblokkeerd om te voorkomen dat de delegaties naar de vergaderzaal vertrokken.
  - *EN* Elsewhere in the city hotels were blocked to prevent the delegations to the meeting left.
- Example 2
  - *NL* Hij wil het weghalen van de kernwapens uit Duitsland ter sprake brengen op een internationale conferentie over ontwapening.
  - *EN* He wants the removal of nuclear weapons from Germany to raise at an international conference on disarmament.

Fluency

- Example 1
  - *NL* Elders in de stad werden hotels geblokkeerd om te voorkomen dat de delegaties naar de vergaderzaal vertrokken.
  - *EN* Elsewhere in the city hotels were blocked to prevent the delegations to the meeting left.
- Example 2
  - *NL* Hij wil het weghalen van de kernwapens uit Duitsland ter sprake brengen op een internationale conferentie over ontwapening.
  - *EN* He wants the removal of nuclear weapons from Germany to raise at an international conference on disarmament.

Fluency

- Example 1
  - *NL* Elders in de stad werden hotels geblokkeerd om te voorkomen dat de delegaties naar de vergaderzaal vertrokken.
  - *EN* Elsewhere in the city hotels were blocked to prevent the delegations to the meeting left.
- Example 2
  - *NL* Hij wil het weghalen van de kernwapens uit Duitsland ter sprake brengen op een internationale conferentie over ontwapening.
  - *EN* He wants the removal of nuclear weapons from Germany to raise at an international conference on disarmament.

Parameter Estimation

- The costs of a (partial) translation are computed as weighted components:

$$p(e|f) = \sum_{i=1}^{k} \lambda_i h_i(e,f)$$

- where λ_i is the weight of a factor h_i (e.g., the language model or the distortion model)
- How to select the optimal value for each λ_i?
- Apply machine learning methods to estimate parameters

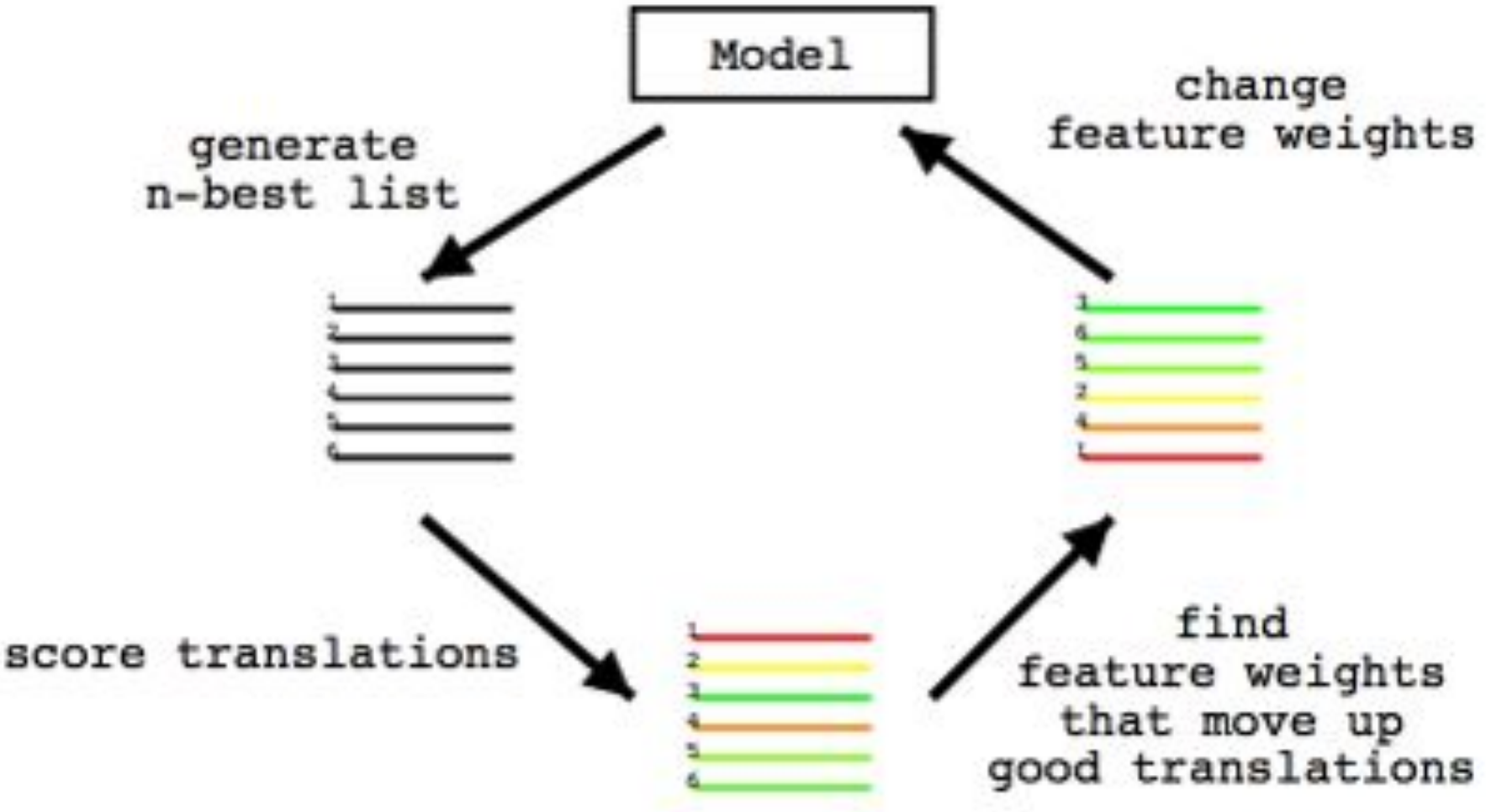
Parameter Estimation

- Data:
  - **Training data:** bitext to learn the translation model (and maybe a lexicalized distortion model); 50K+ sentence pairs
  - **Development data:** data that is not part of the bitext but is used to optimize the parameters iteratively; 1-2K sentence pairs
  - **Test data:** unseen data that is used to compare different methods; 1-2K sentence pairs
- Just like for the bitext, there must be known translations for the development and test data (and preferably multiple translations)

Parameter Estimation

- Additional challenge for MT, there is not one correct translation for a sentence, but dozens
- Use an automatic metric as the objective evaluation function (e.g., BLEU or WER)
- Re-score n-best lists such that better translations (according to the metric) are higher in the ranking than poorer translations

Parameter Estimation Cycle



SMT Pros and Cons

- Pros:
  - Simple general model
  - No manual rule-writing involved
  - Fast and scalable
  - Allows for ranking of translation candidates
  - Does not require any linguistic analysis of the input
- Cons:
  - Complex translation phenomena (e.g., translation divergences) are difficult to model
  - Relies on lexical parameters (limited generalization)
  - Constraints on the search space are often rather simplistic

Challenges in MT and Beyond

- MT quality is still far from perfect
- How to evaluate MT automatically and reliably
- Extending MT to new languages
  - Use of Internet resources
  - Learning translation models from comparable corpora
  - Translation into foreign languages
- Translation of speech
- Translation of different genres/domains
- Integrating MT into information retrieval
- Integrating MT into data mining applications

Current Trends in SMT

- More data = better models = better quality (true?)
  - Bigger LMs do help (Google experience)
  - Bigger parallel corpora sometimes help
- Careful use of syntactic information
  - ISI uses a syntax-based MT system (syntax on the target side)
  - BBN use dependency-based language models
  - ISI's formal grammar MT approach
- Syntax-based re-ordering of the source as pre-processing (Edinburgh)

Internet Applications of MT

- Web site translation
- Cross-language search
- Cross-language Instant Messaging
- Computer aided-translation of news content
- Plagiarism detection of copyright violations (through translation)
- Translations of movie subtitles
- ...

Recap

- The role of MT in today's information society
- MT developments since the 50s
  (interlingua-based $\rightarrow$ rule-based $\rightarrow$ statistical)
- The components of SMT systems:
  - Translation model (form a parallel corpus)
  - Language model
  - Decoder
- Automated evaluation of MT