Statistical Machine Translation

-introduction-

Artem Sokolov
Computerlinguistik
Universität Heidelberg
Sommersemester 2015

material from A. Fraser, P. Koehn, S. Vogel, B. Dorr, C. Monz, S. Riezler, C. Botet, H. Thompson

- Vorlesung – Artem Sokolov
  - Thursdays, 11:15-12:45
  - INF 325 / SR 3
- holiday on Thursdays
  - 14.05 Himmelfahrt
  - 04.06 Fronleichnam
- Sprechstunde
  - Thursdays, 14:00-15:00
  - email beforehand to `sokolov@cl...`
  - business trip on 1.06

Organizational II

- Übung – Sariya Karimova
  - ➡ Tuesdays, 14:15-15:45
  - ➡ INF 346 / SR 10
- no sessions after lectures that fall on holidays
  - ➡ 19.05 first Tuesday after Himmelfahrt
  - ➡ 09.06 first Tuesday after Fronleichnam
- Sprechstunde
  - ➡ Wednesdays, 14:00-15:00

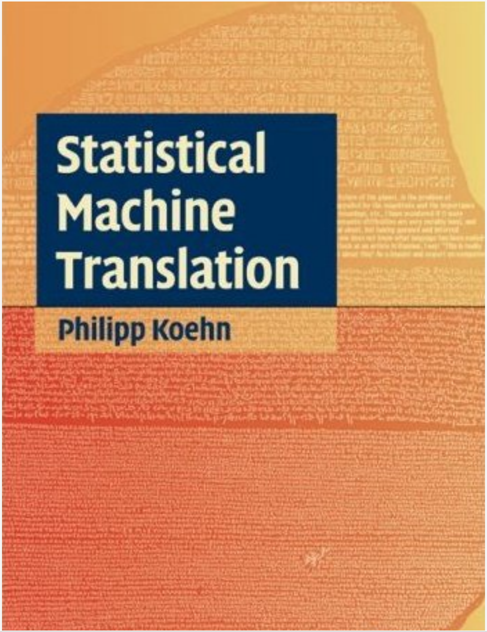
- attendance of lectures and practice sessions
- developed SMT system
- **homework**
- **exam** 23.07

You will learn:

- basics of learning to translate from corpus data
- basics of internals of mainstream SMT systems
- mathematical details necessary
- analyze the bottlenecks of SMT

1. **Softwareprojekt**,
   Tuesdays, 14:15-17:45
   (partial overlap with SMT Übung should be no problem)
2. **Hauptseminar** "Learning and Search in Structured Prediction",
   Tuesdays, 11:15-12:45

Book



questions?

1 Organization

2 Machine Translation

- dreams about automating translation at least since ..th century

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / 'code talkers':
  - ➡ WW1, 1918, Choctaw
  - ➡ WW2, 1942-1945, Navajo, Basque
  - ➡ Balkans, 1990s, Welsh

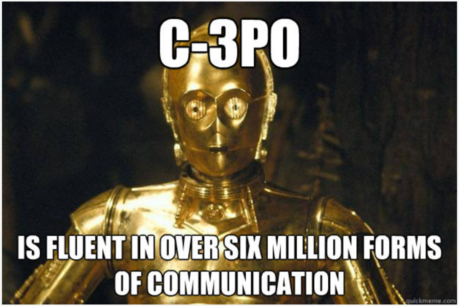
- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / ‘code talkers’:
  - ➡ WW1, 1918, Choctaw
  - ➡ WW2, 1942-1945, Navajo, Basque
  - ➡ Balkans, 1990s, Welsh
- serious projects conceived after 1947
  - ➡ Warren Weaver, “Translation” memorandum
    - **language as a code**
      “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”
    - language & invariants (interlingua)
    - meaning & context (window context to disambiguate)
      EN: ‘fast’ → DE: ‘schnell’, ‘rasch’ oder ‘bewegungslos’, ‘fest’
    - language & logic
      (translation as formal “proof” from source “assumptions”)
  - ➡ controlled language (tech. manual, internal docs of corporations)

- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / ‘code talkers’:
  - ➡ WW1, 1918, Choctaw
  - ➡ WW2, 1942-1945, Navajo, Basque
  - ➡ Balkans, 1990s, Welsh
- serious projects conceived after 1947
  - ➡ Warren Weaver, “Translation” memorandum
    - **language as a code**
      “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”
    - language & invariants (interlingua)
    - meaning & context (window context to disambiguate)
      EN: ‘fast’ $\rightarrow$ DE: ‘schnell’, ‘rasch’ oder ‘bewegungslos’, ‘fest’
    - language & logic
      (translation as formal “proof” from source “assumptions”)
  - ➡ controlled language (tech. manual, internal docs of corporations)
- first system in 1954 (Georgetown experiment)

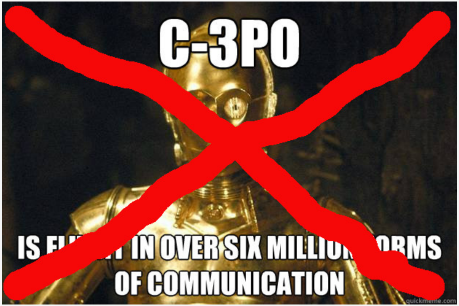
- dreams about automating translation at least since ..th century
- some amateur attempts since 1930s
- war anecdotes / 'code talkers':
  - ➡ WW1, 1918, Choctaw
  - ➡ WW2, 1942-1945, Navajo, Basque
  - ➡ Balkans, 1990s, Welsh
- serious projects conceived after 1947
  - ➡ Warren Weaver, "Translation" memorandum
    - **language as a code**
      "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
    - language & invariants (interlingua)
    - meaning & context (window context to disambiguate)
      EN: 'fast' → DE: 'schnell', 'rasch' oder 'bewegungslos', 'fest'
    - language & logic
      (translation as formal "proof" from source "assumptions")
  - ➡ controlled language (tech. manual, internal docs of corporations)
- first system in 1954 (Georgetown experiment)
- IBM model 1980s

1. commercial
   - governments invest in MT languages used by countries that pose economic/military threats
   - online translation is VERY popular
     (the most used of Google's special projects)
   - EU spends more than $1 billion on translation costs each year
   - (semi-)automated translation leads to huge savings for businesses
     - Systran, Unbabel (internships!), Duolingo, Safaba, Fliplingo, ...

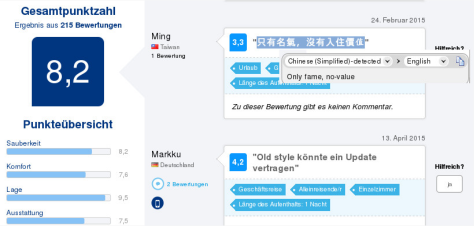
1. commercial
   - governments invest in MT languages used by countries that pose economic/military threats
   - online translation is VERY popular
     (the most used of Google's special projects)
   - EU spends more than $1 billion on translation costs each year
   - (semi-)automated translation leads to huge savings for businesses
     - Systran, Unbabel (internships!), Duolingo, Safaba, Fliplingo, ...
2. academic
   - (probably) the most challenging problem in NLP
   - requires knowledge from many NLP sub-areas
     (semantics, parsing, morphology, stat. modeling)
   - enables resource transfer from one language to another over an established link between them


C-3PO
IS FLUENT IN OVER SIX MILLION FORMS
OF COMMUNICATION
quickmeme.com

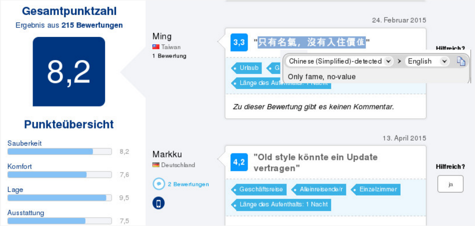
the goal is not to build C-3PO!



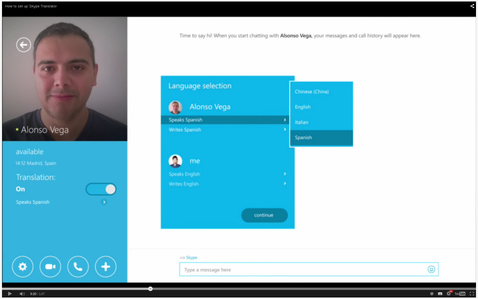
- gisting
  - ➡ get core message (news digests, hotel reviews)



- gisting and grounding
  - ➡ get core message (news digests, hotel reviews)
  - ➡ enable action (shopping, booking)



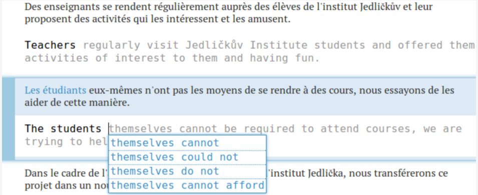
- gisting and grounding
  - ➡ get core message (news digests, hotel reviews)
  - ➡ enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)



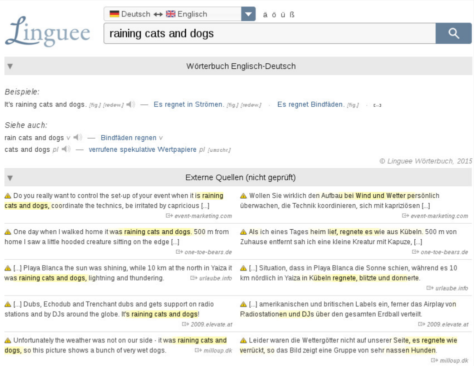
- gisting and grounding
  - ➡ get core message (news digests, hotel reviews)
  - ➡ enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)

- gisting and grounding
  - ➡ get core message (news digests, hotel reviews)
  - ➡ enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)

- gisting and grounding
  - ➡ get core message (news digests, hotel reviews)
  - ➡ enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)
- support of professional translations
  - ➡ rough translation, then post-editing

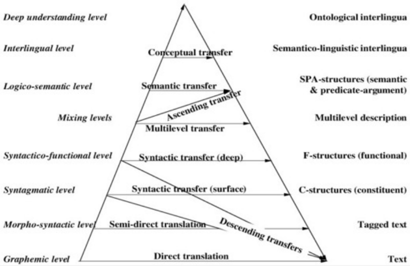


- gisting and grounding
  - ➡ get core message (news digests, hotel reviews)
  - ➡ enable action (shopping, booking)
- integration with speech (ambiguity propagation, real-time)
- MT on portable devices (tourists, medical workers, soldiers, augmented reality)
- support of professional translations
  - ➡ rough translation, then post-editing
  - ➡ translation memory

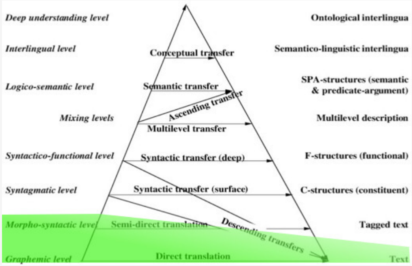


- MT task: generate medium- or high-quality translations of documents
- **all** current MT systems work only at sentence level!
- independent translation of sentences is already a very difficult problem
- important discourse phenomena are ignored:
  Example: How to translate English ‘it’ to German (feminine/masculine/neutral) if object referred to was in previous sentences?

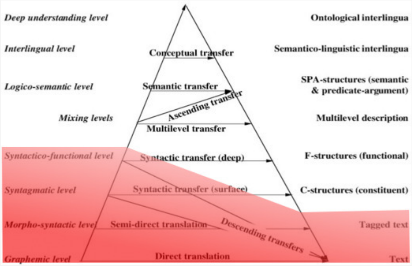
- grammar-based / rule-based
  - ➡ interlingua
  - ➡ transfer
- direct
  - ➡ statistical
  - ➡ example-based



- grammar-based / rule-based
  - ➡ interlingua
  - ➡ transfer
- direct
  - ➡ statistical
  - ➡ example-based



- grammar-based / rule-based
  - ➡ interlingua
  - ➡ transfer
- direct
  - ➡ statistical
  - ➡ example-based



- using statistical models
  - create many alternatives, hypotheses
  - give a score to each hypothesis
  - select the best $\rightarrow$ search

- using statistical models
  - ➡ create many alternatives, hypotheses
  - ➡ give a score to each hypothesis
  - ➡ select the best $\rightarrow$ search
- advantages
  - ➡ avoids hard decisions
  - ➡ speed can be traded with quality, no all-or-nothing
  - ➡ works better in the presence of unexpected/disfluent input
  - ➡ learns from real world, abundant data
  - ➡ high model and methods reusability

- using statistical models
  - ➡ create many alternatives, hypotheses
  - ➡ give a score to each hypothesis
  - ➡ select the best $\rightarrow$ search
- advantages
  - ➡ avoids hard decisions
  - ➡ speed can be traded with quality, no all-or-nothing
  - ➡ works better in the presence of unexpected/disfluent input
  - ➡ learns from real world, abundant data
  - ➡ high model and methods reusability
- disadvantages
  - ➡ difficulties handling structurally rich models, mathematically and computationally
  - ➡ need more data to train the model with increasing number of parameters
  - ➡ not easily interpretable, difficult to distill rules by observing the system

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)
2. **sentence alignment**: in each document pair find those sentences which are translations of one another
   - ➡ results in sentence pairs (sentence and its translation)

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)
2. **sentence alignment**: in each document pair find those sentences which are translations of one another
   - ➡ results in sentence pairs (sentence and its translation)
3. **word alignment**: in each sentence pair annotate those words which are translations of one another
   - ➡ results in aligned word-phrases

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)
2. **sentence alignment**: in each document pair find those sentences which are translations of one another
   - ➡ results in sentence pairs (sentence and its translation)
3. **word alignment**: in each sentence pair annotate those words which are translations of one another
   - ➡ results in aligned word-phrases
4. estimate a **statistical model** from the word-aligned sentence pairs
   - ➡ results in translation model parameters

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)
2. **sentence alignment**: in each document pair find those sentences which are translations of one another
   - ➡ results in sentence pairs (sentence and its translation)
3. **word alignment**: in each sentence pair annotate those words which are translations of one another
   - ➡ results in aligned word-phrases
4. estimate a **statistical model** from the word-aligned sentence pairs
   - ➡ results in translation model parameters

Language Modeling:

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)
2. **sentence alignment**: in each document pair find those sentences which are translations of one another
   - ➡ results in sentence pairs (sentence and its translation)
3. **word alignment**: in each sentence pair annotate those words which are translations of one another
   - ➡ results in aligned word-phrases
4. estimate a **statistical model** from the word-aligned sentence pairs
   - ➡ results in translation model parameters

Language Modeling:

5. large **monolingual corpus**
   - ➡ texts in target language

Training:

1. large **parallel corpus**
   - ➡ consists of document pairs (document and its translation)
2. **sentence alignment**: in each document pair find those sentences which are translations of one another
   - ➡ results in sentence pairs (sentence and its translation)
3. **word alignment**: in each sentence pair annotate those words which are translations of one another
   - ➡ results in aligned word-phrases
4. estimate a **statistical model** from the word-aligned sentence pairs
   - ➡ results in translation model parameters

Language Modeling:

5. large **monolingual corpus**
   - ➡ texts in target language
6. estimate a **statistical model** from examples of well-formed language
   - ➡ results in language model: how likely a word will follow a given history

Tuning:

6. define how important is every model for translation quality
   - ➡ results in a complete model

Tuning:

6. define how important is every model for translation quality
   - ➡ results in a complete model

Testing:

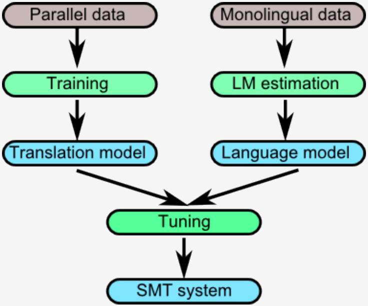
- given new text to translate, apply model to get most likely translation

Tuning:

6. define how important is every model for translation quality
   - ➡ results in a complete model

Testing:

- given new text to translate, apply model to get most likely translation



Traditional focus was on high-resourced languages:

- high demand $\Rightarrow$ data collection efforts
- available data $\Rightarrow$ spawns research
- quality systems $\Rightarrow$ proliferation, new markets $\Rightarrow$ more demand

Traditional focus was on high-resourced languages:

- high demand $\Rightarrow$ data collection efforts
- available data $\Rightarrow$ spawns research
- quality systems $\Rightarrow$ proliferation, new markets $\Rightarrow$ more demand

No clear-cut definition in number of words:

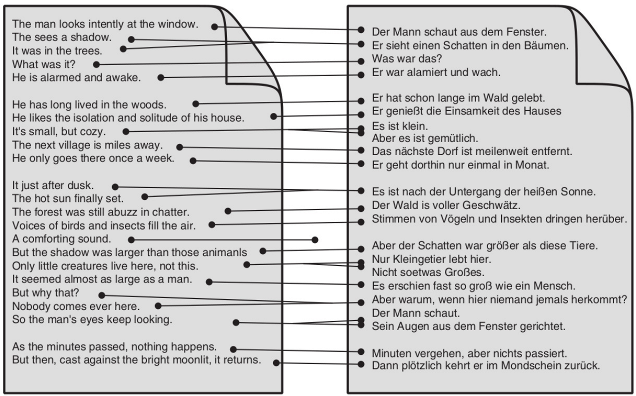
- > 200M high-resourced — French, Chinese, Arabic
- ~ 50M medium-resourced — German, Portuguese, Italian
- < 5M under-resourced — Tatar, Uzbek, Estonian
- < 100K close to none — Chechen, Udmurt, *Silbo*, *Klingon* :)
- heavily depends on a language pair and direction:
  for example: ZH-EN is well-resourced, FR-ZH is much less so

english	german
Diverging opinions about planned tax reform	Unterschiedliche Meinungen zur geplanten Steuerreform
The discussion around the envisaged major tax reform continues .	Die Diskussion um die vorgesehene grosse Steuerreform dauert an .
The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 .	Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafuer aus , wesentliche Teile der fuer 1999 geplanten Reform vorzuziehen .

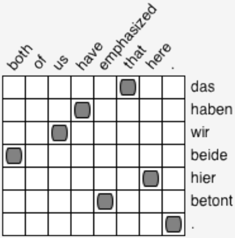
english	**german**
Diverging opinions about planned tax reform	Unterschiedliche Meinungen zur geplanten Steuerreform
The discussion around the envisaged major tax reform continues .	Die Diskussion um die vorgesehene grosse Steuerreform dauert an .
The FDP economics expert , Graf Lambsdorff , today came out in favor of advancing the enactment of significant parts of the overhaul , currently planned for 1999 .	Der FDP - Wirtschaftsexperte Graf Lambsdorff sprach sich heute dafuer aus , wesentliche Teile der fuer 1999 geplanten Reform vorzuziehen .

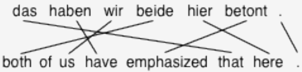
- note some pre-processing (tokenization, normalization)

- if document D_e is translation of document D_f, how to find the translation for each sentence?
- the n-th sentence in D_e is not necessarily the translation of the n-th sentence in D_f
- in addition to 1:1 alignments, there are also 1:0, 0:1, 1:n, and n:1
- in EuroParl proceedings, $\sim 90\%$ of the sentence alignments are 1:1

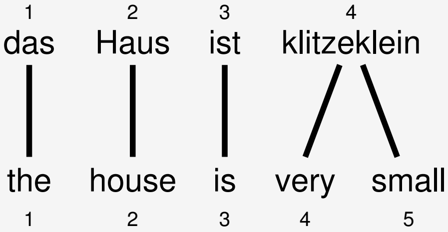


- given sentences that are translation of one another, how to know which words are mutual translations?





- given sentences that are translation of one another, how to know which words are mutual translations?



Goal:

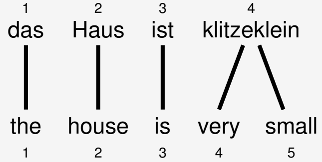
- get a score function $p(e|f)$ – goodness of translation e given foreign input f
  1. $p(\text{'die Waschmaschine läuft'}, \text{'the washing machine is running'}) = 0.95$
  2. $p(\text{'die Waschmaschine läuft'}, \text{'the car drove'}) = 0.03$
- convenient to think of p as probability
- models to some extent natural language's uncertainty and ambiguity
- translation: $\arg\max_e p(e|f)$

What kind of function can $p(e|f)$ be?:

- one naïve way to determine $p(e|f)$:
  1. count how many times f was translated by e_1 or e_2 in the training data
  2. set $p(e_1|f) = \frac{\#\{f \to e_1\}}{\#\{f \to ?\}}$
  3. set $p(e_2|f) = \frac{\#\{f \to e_2\}}{\#\{f \to ?\}}$
  - only works of we saw exactly the f and e_1,e_2 in our training data
  - we can't generalize to unseen sentences

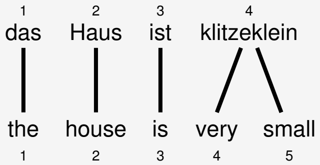
➡ **solution – decompose input and output into parts**

4. Translation Model - Maximum Likelihood Estimation



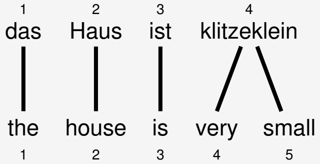
- generate a word alignment for each sentence pair
- count the number of times every source word was linked to every target word:
  1. $\#\{\text{das} \rightarrow \text{the}\} = 1$
  2. $\#\{\text{Haus} \rightarrow \text{house}\} = 1$
  3. $\#\{\text{ist} \rightarrow \text{is}\} = 1$
  4. $\#\{\text{klitzeklein} \rightarrow \text{very}\} = 1$
  5. $\#\{\text{klitzeklein} \rightarrow \text{small}\} = 1$

4. Translation Model - Maximum Likelihood Estimation



- generate a word alignment for each sentence pair
- count the number of times every source word was linked to every target word:
  1. $\#\{\text{das} \rightarrow \text{the}\} = 1.0$
  2. $\#\{\text{Haus} \rightarrow \text{house}\} = 1.0$
  3. $\#\{\text{ist} \rightarrow \text{is}\} = 1.0$
  4. $\#\{\text{klitzeklein} \rightarrow \text{very}\} = 0.5$
  5. $\#\{\text{klitzeklein} \rightarrow \text{small}\} = 0.5$
- divide by the number of occurrences of the source word

4. Translation Model - Maximum Likelihood Estimation



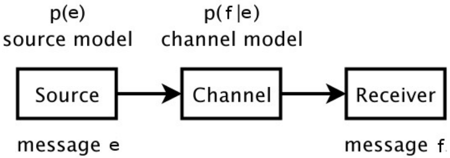
- generate a word alignment for each sentence pair
- count the number of times every source word was linked to every target word:
  1. $\#\{\text{das} \rightarrow \text{the}\} =$
  2. $\#\{\text{Haus} \rightarrow \text{house}\} =$
  3. $\#\{\text{ist} \rightarrow \text{is}\} =$
  4. $\#\{\text{klitzeklein} \rightarrow \text{very}\} =$
  5. $\#\{\text{klitzeklein} \rightarrow \text{small}\} =$
- divide by the number of occurrences of the source word
- this is our word/phrase translation probability $p(w_e|w_f)$

- decomposing can introduce output disfluencies
- need to somehow improve fluency in translations
- learn what is "fluent" from examples of well-formed language
- results in language model: how likely a word will follow a given history
  - ➡ $p(\text{Haus}|\text{Das kleine}) > p(\text{Haus}|\text{Die kleine})$

Translating is usually referred to as **decoding** (W. Weaver, 1947)

Noisy Channel Model

$$\arg\max_{e} p(e|f)$$



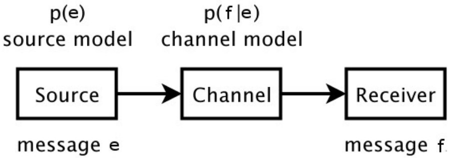
SMT was born from automatic speech recognition:

- $p(e)$ = language model
- $p(f|e)$ = acoustic model

Translating is usually referred to as **decoding** (W. Weaver, 1947)

Noisy Channel Model

$$\arg\max_e p(e|f) = \arg\max_e p(f|e)p(e)$$



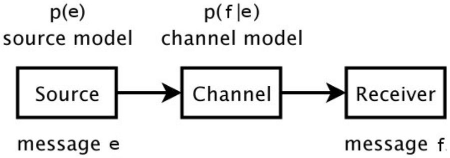
SMT was born from automatic speech recognition:

- $p(e)$ = language model
- $p(f|e)$ = acoustic model

Translating is usually referred to as **decoding** (W. Weaver, 1947)

Noisy Channel Model

$$\arg\max_e p(e|f) = \arg\max_e \underbrace{p(f|e)}_{\text{transl. model}} \underbrace{p(e)}_{\text{lang. model}}$$



SMT was born from automatic speech recognition:

- $p(e)$ = language model
- $p(f|e)$ = acoustic model
- however, SMT must deal with word reordering!

Injecting Domain Knowledge

$$\arg\max_e p(e|f) = \arg\max_e p(f|e)p(e)$$

Injecting Domain Knowledge

$$\arg\max_{e} \log p(e|f) = \arg\max_{e}\ \log p(f|e) +\ \log p(e)$$

- move to log-space

Injecting Domain Knowledge

$$\arg\max_e \log p(e|f) = \arg\max_e \alpha \log p(f|e) + \beta \log p(e)$$

- move to log-space
- models may have different importance (weight)

Injecting Domain Knowledge

$$\arg\max_e \log p(e|f) = \arg\max_e \alpha \log p(f|e) + \beta \log p(e) + \gamma f(\cdot)$$

- move to log-space
- models may have different importance (weight)
- we may want to add more models

Injecting Domain Knowledge

$$\arg\max_e \log p(e|f) = \arg\max_e \alpha f_1(\cdot) + \beta f_2(\cdot) + \gamma f_3(\cdot)$$

- move to log-space
- models may have different importance (weight)
- we may want to add more models
- they even need not to be log-probabilities (features)

Generalization

$$\arg\max_e \log p(e|f) = \arg\max \sum_{i=1}^{n} w_i f_i(e,f)$$

- move to log-space
- models may have different importance (weight)
- we may want to add more models
- they even need not to be log-probabilities (features)
- maximize score function – a weighted linear combination of features

Generalization

$$\arg\max_e \log p(e|f) = \arg\max \sum_{i=1}^{n} w_i f_i(e, f)$$

1. find such w_i that maximize translation quality
2. many methods exist and still an active research area

- how to know if your SMT system works well?
- run it on a large number of unseen sentences and evaluate the quality
- but what is ‘quality’?
  - ➡ can evaluate MT at corpus, document, sentence or word level..
  - ➡ in the MT the unit of translation is the sentence
- human evaluation of MT quality is difficult (expensive)
- need an abstract measure of usefulness of the output
  - ➡ evaluation metric: assigns a score to a hypothesized translation
  - ➡ automatic evaluation metrics rely on comparison with selected human translations

- **WER** (word error rate)
  - ➡ edit distance to reference translation (insertion, deletion, substitution)
  - ➡ captures fluency well, adequacy not so well
  - ➡ rigid: gives no credit for translating 'Frau' instead of 'Fräulein'
- **TER** (translation error rate)
  - ➡ edit distance to reference translation (+ **block moves**)
  - ➡ captures reordering freedom better, very good correlation with humans
  - ➡ common problems: synonyms,
- **BLEU** (most popular)
  - ➡ counts matching n-grams
  - ➡ captures fluency, rewards long and fluent matches
  - ➡ penalizes the noisy channel model's tendency to produce short outputs
  - ➡ well-correlates with humans, very intuitive, easier then TER for learning
  - ➡ cons: no credit for synonyms, for legitimate but slightly reordered outputs
- **METEOR**
  - ➡ combines synonyms, stemming, WordNet synsets
  - ➡ most "human like"
  - ➡ attempts to capture language flexibility
  - ➡ cons: language dependent (stemmer, WordNet)

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning
3. learning from weak feedback (‘under-paid turkers’)

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning
3. learning from weak feedback (‘under-paid turkers’)
4. learning in non-cooperative environment

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning
3. learning from weak feedback ('under-paid turkers')
4. learning in non-cooperative environment
5. learning from non-parallel data

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning
3. learning from weak feedback ('under-paid turkers')
4. learning in non-cooperative environment
5. learning from non-parallel data
6. SMT with neural networks

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning
3. learning from weak feedback ('under-paid turkers')
4. learning in non-cooperative environment
5. learning from non-parallel data
6. SMT with neural networks
7. include over-sentential context

Our groups research directions

1. cross-lingual information retrieval (e.g., patents)
2. grounded learning
3. learning from weak feedback ('under-paid turkers')
4. learning in non-cooperative environment
5. learning from non-parallel data
6. SMT with neural networks
7. include over-sentential context

SMT projects from this term's SWP (today, 16:15, INF327 SR2):

- quasi-parallel corpus creation
- kernel-SMT without alignments
- SMT on character levels
- neural networks for bilingual word representations
- user feedback based SMT learning

1. Word-Based Models
2. Phrase-Based SMT
3. Decoding
4. Language Models
5. Evaluation
6. Tree-Based SMT

see you the day after tomorrow at 11:15, INF 327 SR3