Machine Translation
Lecture I – Organization

Sebastian Pado

Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

April 20, 2021

About me

- Sebastian Pado
  - Studied computational linguistics, computer science (Saarland University, Edinburgh)
  - PhD in computational linguistics (Saarland)
  - Postdoc in Stanford
  - Professor for computational linguistics in Heidelberg, Stuttgart
  - Main research: Semantics, Multilinguality, DH/CSS

Organization

- We‘re still doing this in a mixed offline/online setting
- Lectures on Tuesdays 9:45
  - As video lectures, available from the ILIAS page
  - 20 min blocks, most reused from 2020 (..sorry!)
- Tutorials on Thursdays 15:45
  - As video conferences at unistuttgart.webex.com/meet/sebastian.pado
  - Opportunity to ask questions about video lectures
  - Discuss problem sets (details below)

Material

- ILIAS course page
  - Announcements, materials, forum, chat room, …
- I realized not everybody has access at this moment
  - Material (initially) replicated at https://nlpado.de/~sebastian/mt21

Prerequisites, Module

- The course assumes that you have a foundational knowledge of CL (in partiular Language Models) and probability theory
  - „Methods of CL“ should be fine
  - Otherwise, have a look at Manning/Schütze textbook
- Standard setup (2017 PO of CL MSc):
  - Module „Machine Translation“
  - Contact me if you are in another situation and need advice

Exam, Problem Sets

- Exam at end of semester
  - Date/form to be determined
  - Questions similar to problem sets
  - How exactly we are going to do the exam will depend on the developments until then
- The course comes with accompanying problem sets
  - About 8—10 over the course of the semester
  - You must submit problem sets to be admitted to exam (guideline: 75% submissions “in good faith“)
    - Encourages ongoing engagement with topics

Problem Sets

- I encourage submission in small groups (up to 3 people)
  - Expected to be stable over the course of the semester
  - I still **strongly recommend** that everyone works out the solutions first
- Goal: More interaction in smaller groups
  - Less pressure on larger online meetings
- This is what I currently plan:
  - Problem set becomes available on Tue together with a lecture
  - Solutions to be submitted the following Tuesday
  - Thursday‘s tutorial session allots time both for questions on the (current) lecture and for detailed discussions on your solutions

Apologies in advance

- Online teaching is not a perfect substitute
- Particular problem: lack of **backchannel**
  - Please contact me **proactively** in case of problems, suggestions, …

Machine Translation
Lecture I – Course Overview

Sebastian Pado

Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

April 20, 2021

slide credit: Alexander Fraser, Daniel Quernheim

About the course - Content

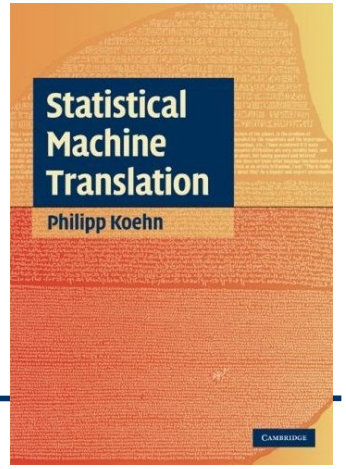
1. Overview and History of Machine Translation
2. Statistical Machine Translation
   - Dominant for 10 years (2005-2015)
   - Many methods of interest beyond MT
3. Neural Machine Translation
   - The new king of the hill

Outline

- Lecture 1: Introduction, Overview
- Lecture 2, 3: Word Alignment
- Lecture 4: Phrase-based Translation
- Lecture 5: Decoding
- Lecture 6: Tuning
- Lecture 7: Evaluation
- Lecture 8: Hierarchical, Syntax-based SMT
- Lecture 9: Neural MT, Basics
- Lecture 10: Neural MT, Established Methods
- Lecture 11: Neural MT, Current Research
- Lecture 12: Language-Specific Processing

Remarks

- Some lectures will have quite a bit of math
- Others not so much
  - I will separate into **main ideas** (easily accessible) and **details** (CS/math background helpful)
  - I expect you to get into **both**
  - This might require you to **brush up your math**:
- Lectures 3--8 roughly follow Koehn‘s „Statistical Machine Translation“
  - No good reading for NMT yet



Lecture 1 – Introduction

- **Machine translation**
- Statistical machine translation in a nutshell

How it started

- Machine translation was one of the first applications envisioned for computers
- Warren Weaver (1949): “I have a text in front of me which is written in **Russian** but I am going to pretend that it is really written in English and that it has been **coded** in some strange symbols. I will now proceed to **decode**.”
  - Why choice of boldfaced words?
- First demonstrated by IBM in 1954 with a basic word-for-word translation system

Ups and Downs

- 1966: Pierce Report leads to US funding cut
  - "...[it] was slower, less accurate and twice as expensive as human translation, there [was] no immediate or predictable prospect of useful machine translation."
- Many MT enthusiasts become sceptical
  - Bar-Hillel: fully automatic high quality translation (FAHQT) is unattainable "not only in the near future but altogether"
- Renewed interested since 1990 – why?

Renewed interest since 1990

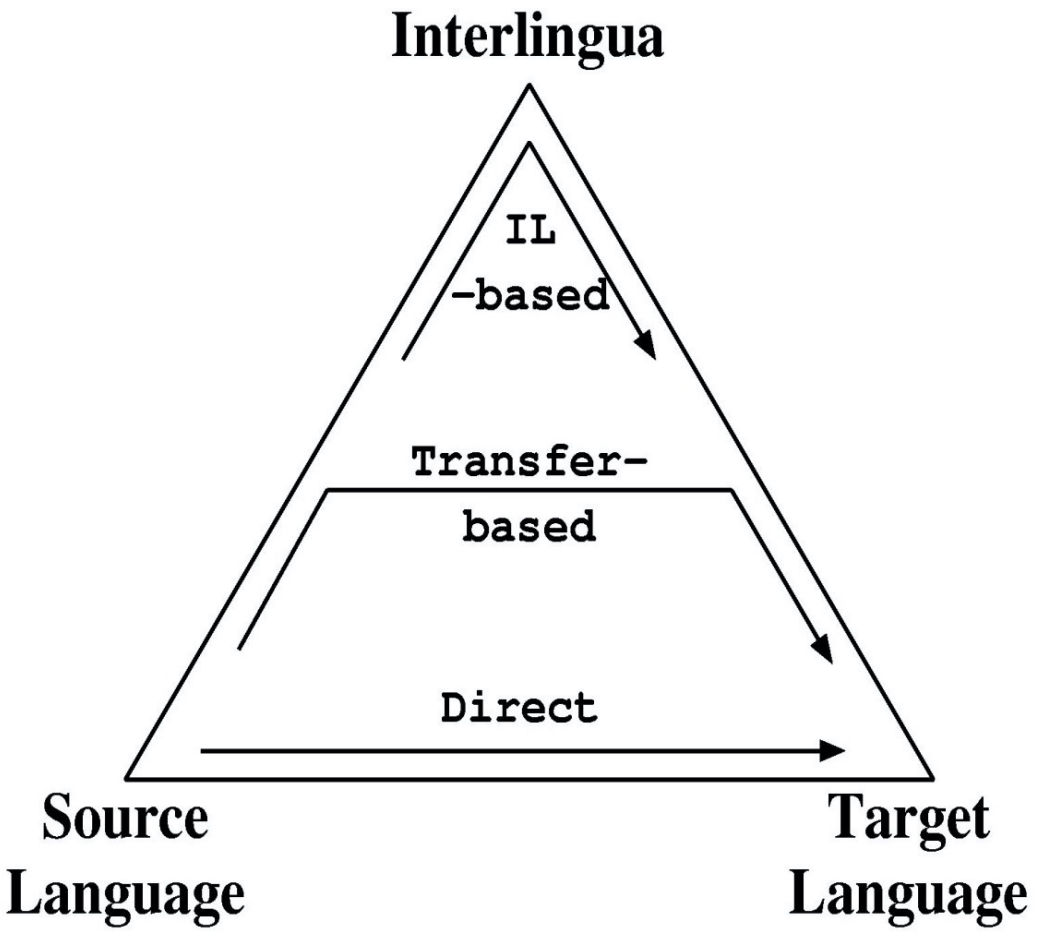
- Correlated with paradigm shift in NLP
  - Availability of „big data“ through the internet
- from **rule-based approaches**
  - hand-crafted, linguistically motivated
- towards **corpus-based approaches**
  - data-oriented, statistical → **SMT, NMT**
- MT is a difficult engineering problem
  - New approaches must mature
    - SMT needed 20 years to become better than rule-based approaches (Systran) – hand-over around 2000
    - NMT needed at least 5 years – hand-over recently (2015-2017)

Goals of Machine Translation

- Another factor of the renaissance: More realistic and nuanced expectations
- MT does not perfectly for each language pair
  - from *gisting*
  - to fully automatic high-quality translation (still a vision!)
- Quality/resource trade-off

Approaches: The classical view

- Grammar-based
  - Interlingua-based
  - Transfer-based
- Direct
  - Example-based
  - Statistical / neural



Modified from Vogel

The SMT view (2000-2015)

- SMT and grammar-based MT are not really opposites
- Two dichotomies
  - Probabilities vs. everything is equally likely (in between: heuristics)
  - Rich (deep) structure vs. no or only flat structure
- Both dimensions are continuous
- Examples
  - EBMT: flat structure and heuristics
  - SMT: flat structure and probabilities
  - XFER: deep(er) structure and heuristics
- Vision of SMT: develop **structurally rich probabilistic models**

	No Probs	Probs
Flat Structure	EBMT	phrase-based SMT
Deep Structure	Transfer, Interlingua	hierarchical SMT

Some examples of translation rules

- Example: *ship* → *Schiff* (noun) vs. *ship* → *verschicken* (verb)
- Rule-based translation (EBMT)

  `(ship, Schiff) near harbor`
- Word-based statistical translation

  `(ship, Schiff,.65); (ship,verschicken,.35)`
- Phrase-based statistical translation

  `(do you ship, verschicken Sie, 0.9);`
  `(the ship, das Schiff, 0.6)`
- Transfer-based translation (XFER)

  `(ship, Schiff) in NP`
- Hierarchical statistical translation

  `([`$_{\text{NP}}$ `X ship], [`$_{\text{NP}}$ `X Schiff], 0.8)`

The rise of SMT: Arabic – English 03-05

Description of the Iraqi President George Bush American elections– which will follow in the current month of the thirty–that they constitute a historic moment, recognizing that the organization of elections in current circumstances difficult issue.

It was considered bush in the press that the pronouncements of the possible organization of elections in most regions of the Iraqi punctually wish that the turnout where high. He added that "Iraqi 14 governorates ال18 appear in the relative calm."

v.2.0 – October 2003

A description of the American president George W. Bush elections– Iraq, which will take place on the thirtieth session of the month-- as a historic moment, acknowledging that the organization of elections in the current difficult circumstances.

Bush said in press statements that it is possible to organize elections in most regions of Iraq to the deadline and I wish that the turnout are high. He added that "14 governorates of Iraq's 18 appeared in relative calm".

v.2.4 – October 2004

Iraqi troops had become a target always Iraqi gunmen (French)

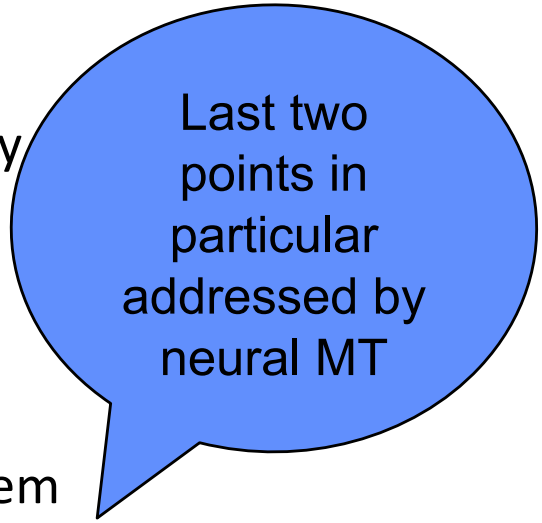
US President George W. Bush described Iraq elections-which will take place on the 30th of this month-- as a historic moment, acknowledging that the elections in the current situation is difficult.

Bush said in a press statement that it be possible to organize elections in most regions of Iraq in time and hoped that the rate of participation in the high. He added that "Iraqi 14 of the provinces of 18 appears to be relatively calm."

v.3.0 - February 2005

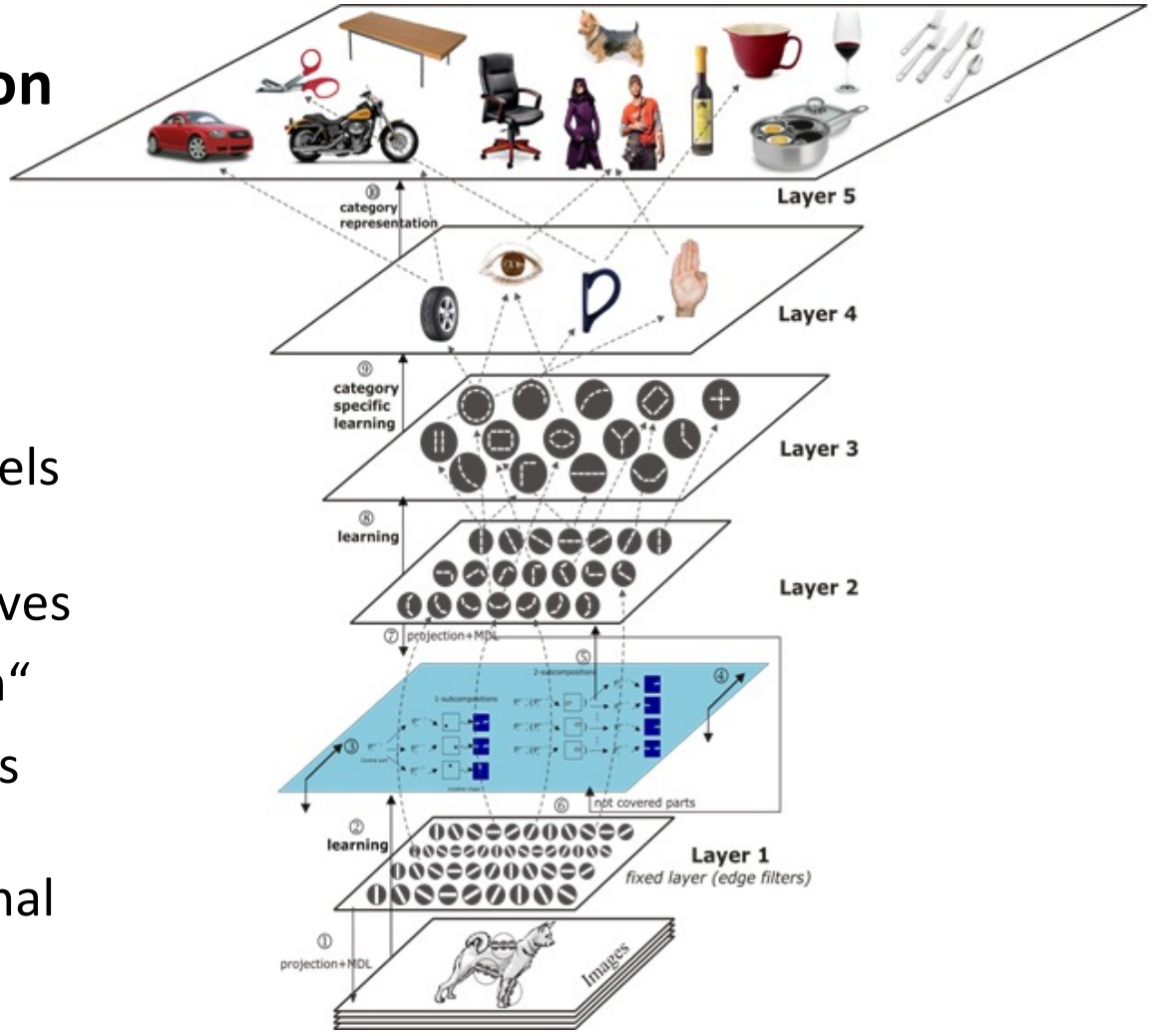
Statistical Approach

- What does it mean to use a statistical approach?
  - Create many alternatives, called hypotheses
  - Assign a **probability** to each hypothesis
  - Select the best hypothesis → search
- Advantages
  - Avoid hard decisions: Speed can be traded with quality
  - Works better in the presence of unexpected input
  - Uses data to train the model
- Disadvantages
  - Difficult to understand decision process made by system
  - **Difficulties handling structurally rich models, mathematically and computationally**
  - **Bottleneck: linguistic analysis (reliability and informativeness)**



The neural view (since 2015)

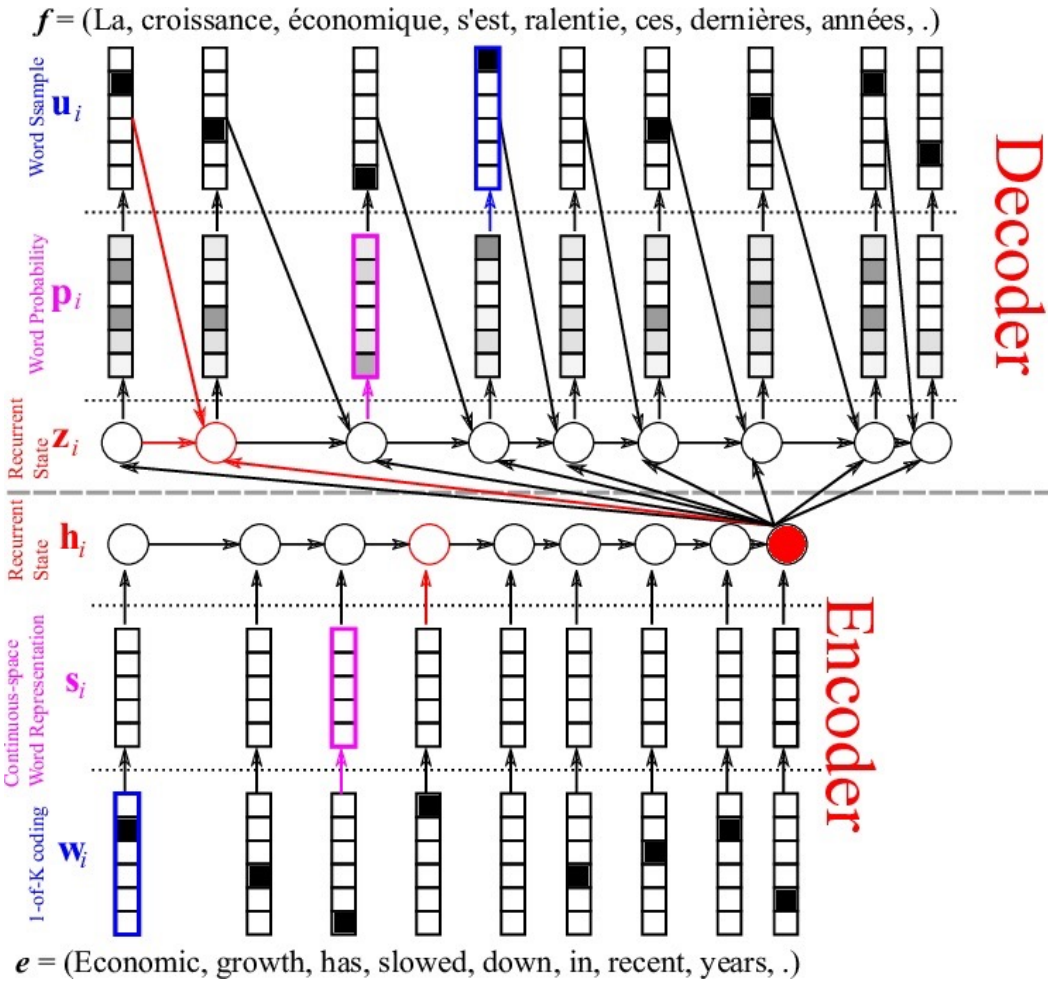
- Main appeal of neural networks: **representation learning**
  - No need to define / develop features that describe a problem
  - Given enough data, models learn informative representations themselves
  - „Incremental abstraction“
  - In NLP: word embeddings as models of meaning („distributed/distributional semantics“)



https://www.vicos.si/Research/LearnedHierarchyOfParts

The neural view (since 2015)

- Lift to neural machine translation: **Encoder-Decoder Architecture**
  - Start from source language word embeddings
  - Compose into sentence meaning vector (encoder)
  - „Unpack“ into word embeddings in the target language (decoder)
- Advantages
  - „end to end learning“
  - approximation to interlingua



Cho et al (EMNLP 2014)

NMT 2015

- Break-even between NMT and phrase-based SMT
  - Still dependent on language pair
  - Clearest benefits for „difficult“ tagets languages
  - Example: EN → DE (PBSY, HPB, SPB: phrase-based) [Bentivogli et al. EMNLP 2016]

system	BLEU	HTER	mTER
PBSY	25.3	28.0	21.8
HPB	24.6	29.9	23.4
SPB	25.8	29.0	22.7
NMT	31.1*	21.1*	16.2*

Commercial interest

- Both “organizational” and “end” users
- U.S. has invested in machine translation (MT) for intelligence purposes
- EU spends more than $1 billion on translation costs each year

- MT is popular on the web—it is the most used of Google’s applications
- Globalization (eCommerce, social media, etc.)

Academic interest

- A very challenging problem in NLP research
- Requires knowledge from many NLP sub-areas, e.g., lexical semantics, syntactic parsing, morphological analysis, statistical modeling,…
- Being able to establish links between two languages allows for transferring resources from one language to another

Machine Translation
Lecture I – SMT in a Nutshell

Sebastian Pado, Enrica Troiano

Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart

April 21, 2019

slide credit: Alexander Fraser, Daniel Quernheim

Lecture 1: Outline

- Machine translation
- **Statistical machine translation in a nutshell**

Building an SMT system in a nutshell

- Recall: SMT is a search problem
- Core: Construct a **scoring function** for a target sentence *e* given a source sentence *f*
  - Formalize as conditional probability p(e|f)
  - p(the washing machine is running | die Waschmaschine läuft) = high number
  - p(the car drove | die Waschmaschine läuft) = low number
- Any ideas?

Where to get a scoring function from?

- Most obvious source: **Parallel corpora**
  - Sets of documents together with their translations
- Constructing a scoring function p(e|f) from a parallel corpus:
  - Count how often sentence *e* occurs
  - Count how often sentence *f* occurs as translation of *e*
- Why is this impractical?
  - Can only translate sentences that were seen during training: Not very useful!

Back to the basics

- We need to somehow **decompose** those sentences
- First approximation (of many): Sentence *e* is a good translation of sentence *f* if the words in *e* are good translations of the words in *f*
  - Intermediate goal: Quantify translation probabilities p(e|f) at the *word level*
    - Word / lexical translation probabilities t(e|f)

How to Build an SMT System

- Start with a large **parallel corpus**
- **Sentence alignment**: in each document pair find sentence pairs which are translations of one another
  - Results in sentence pairs (sentence and its translation)
- **Word alignment**: in each sentence pair find word pairs which are translations of one another
  - Results in word-aligned sentence pairs
- Compute **word translation probabilities**
- Learn an MT model
  - Part 1: Define scoring function for hypotheses
  - Part 2: Define search procedure for hypotheses (decoding)

Parallel corpus

- Example from DE-News (8/1/1996)

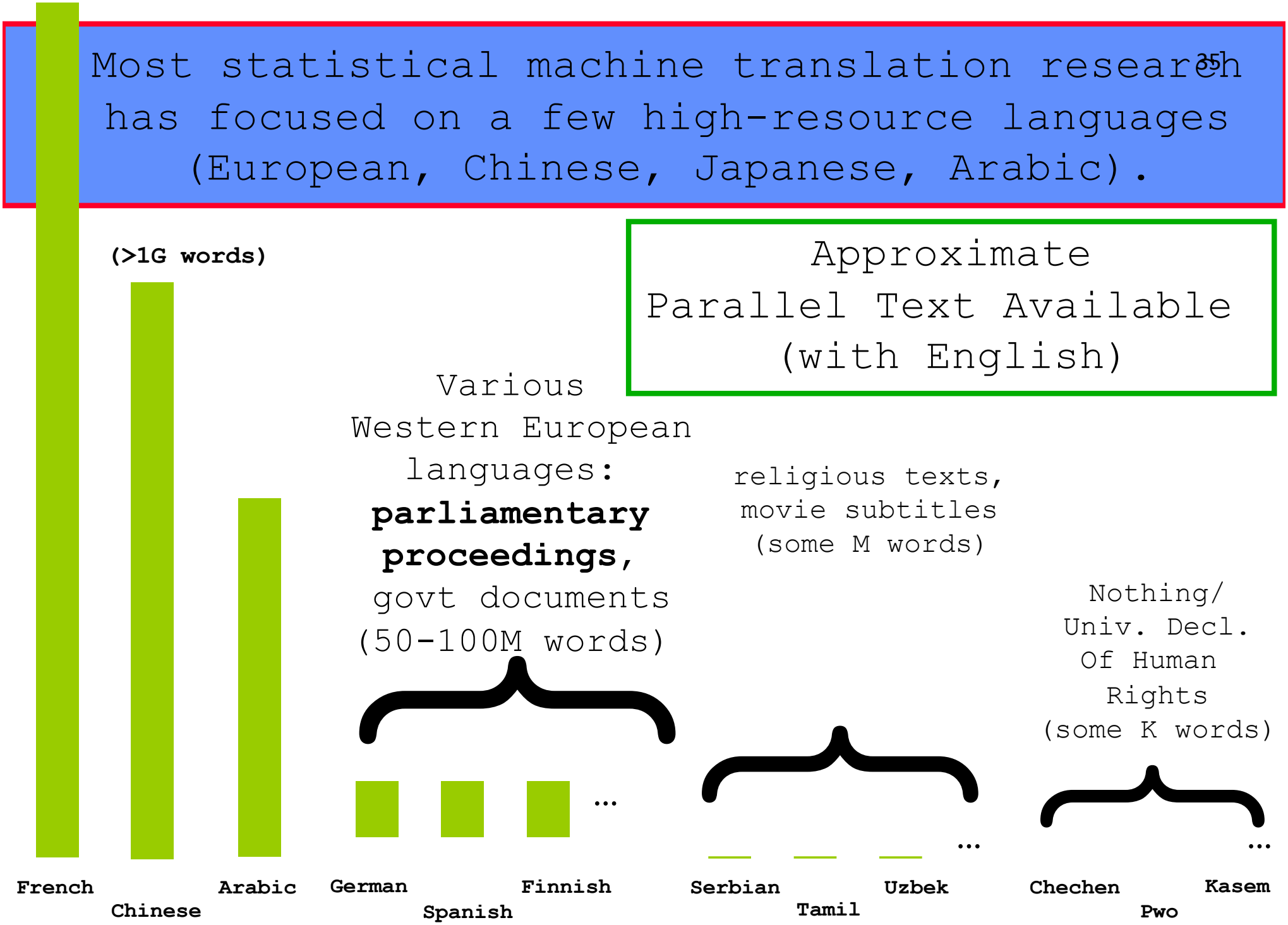
English	German
Diverging opinions about planned tax reform	Unterschiedliche Meinungen zur geplanten Steuerreform
The discussion around the envisa- ged major tax reform continues .	Die Diskussion um die vorgesehene grosse Steuerreform dauert an .
The FDP economics expert , Graf Lambsdorff , today came 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 .

- Are there many parallel corpora for all languages?

Modified from Dorr, Monz

Most statistical machine translation research has focused on a few high-resource languages (European, Chinese, Japanese, Arabic).

Approximate Parallel Text Available (with English)



Modified from Schafer&Smith

Sentence alignment

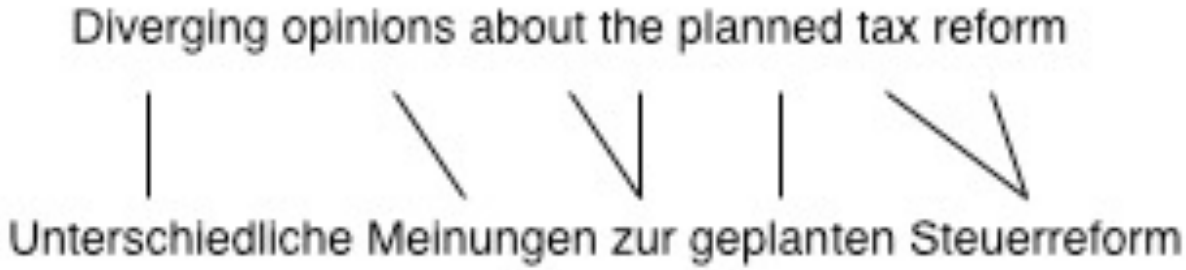
- If document D_e is translation of document D_f how do we find the translation for each sentence?
- The *n*-th sentence in D_e is not necessarily the translation of the *n*-th sentence in document D_f
- In addition to 1:1 alignments, there are also 1:0, 0:1, 1:n, and n:1 alignments
- In European Parliament proceedings, approximately 90% of the sentence alignments are 1:1
  - Little or much compared to other domains?

Sentence alignment

- There are many sentence alignment algorithms:
  - Align (Gale & Church): Considers sentence character length (shorter sentences tend to have shorter translations then longer sentences). Works well.
  - Char-align: (Church): Considers shared character sequences. Works fine for similar languages or technical domains.
  - K-Vec (Fung & Church): Learns a translation lexicon based on parallel texts from distributions of foreign-English word pairs.
  - Cognates (Melamed): Use positions of cognates (including punctuation)
  - Length + Lexicon (Moore; Braune and Fraser): Two passes, high accuracy
- Sentence alignment is a mostly solved problem

Word alignments

- Given a parallel sentence pair we can link (align) words or phrases that are translations of each other:



- How can we define word translation probabilities?
  - Let f(x) be the frequency of word x in a parallel corpus
  - Let f(x,y) be the corpus frequency of an alignment between x and y
  - Then t(y|x) = f(x,y)/f(x) is the word translation probability

Now let's put everything together

- Recall: We need to define a scoring function for target **sentence** e given source sentence f
  - Formalize as conditional probability p(e|f)
  - p(the washing machine is running | die Waschmaschine läuft) = high number
  - p(the car drove | die Waschmaschine läuft) = low number
- Simplest option to define p: As a multiplication of word translation probabilities
  - "Sentence *e* is a good translation of sentence *f* if the words in *e* are good translations of the words in *f*"

Word-based parametrization

- Suppose we translate:
  - "die" to "the"
  - "Waschmaschine" to "washing machine"
  - "läuft" to "is running"
  - (and suppose we ignore word order and many other aspects)
- Then:

  p(the washing machine is running | die Waschmaschine läuft)

  = t(the | die) t(washing | Waschmaschine) t(machine | Waschmaschine) t(is | läuft) t(running | läuft)
- Compare to:

  p(the car is running | die Waschmaschine läuft)

  = t(the | die) t(car| Waschmaschine) t(is | läuft) t(running | läuft)
- System will make correct choice if **t(car | Waschmaschine) low**

Building an SMT system in a nutshell

- Step 2: Implement a **search algorithm** which, given a source language sentence, finds the target language sentence which maximizes f (”**Decoding**”)
  - To translate a new, unseen sentence, call the search algorithm
- Seems similar to Automatic Speech Recognition (ASR) – given a speech signal, find the best transcription
  - Claim: But MT decoding is substantially harder. **Why?**
    - Answer: Translation is not monotone (left-to-right)
    - Much more to say about this!

Literature

- Koehn: Statistical Machine Translation. Chapter 1
- Sentence Alignment:
  - `http://www.statmt.org/survey/Topic/SentenceAlignment`

Final announcement

- First problem set should be online now too
  - Form teams and get going!
  - Submit until next Tuesday
- First tutorial session on Apr 29th, 15:45
  - unistuttgart.webex.com/meet/sebastian.pado
- Watch the ILIAS forum for updates!