Statistical and Neural Machine Translation

Part I - Introduction

Alexander Fraser
CIS, LMU München

2020 04 22 Machine Translation

SMT and NMT

- MT = machine translation
- SMT = statistical machine translation
  - Models built using simple statistics
  - Critical knowledge source: parallel corpora
  - Dominant approach until 2015
- NMT = neural machine translation
  - Models built using deep learning
  - Critical knowledge source: parallel corpora
  - Cutting edge

Lecture 1 – Introduction + Eval

- Machine translation
- Data driven machine translation
  - Parallel corpora
  - Sentence alignment
- Overview of statistical machine translation
- Evaluation of machine translation

A brief history

- 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. All I need to do is strip off the code in order to retrieve the information contained in the text."
- First demonstrated by IBM in 1954 with a basic word-for-word translation system

Interest in machine translation

- Commercial interest:
  - U.S. has invested in machine translation (MT) for intelligence purposes
  - MT is popular on the web—it is the most used of Google's special features
  - EU spends more than $1 billion on translation costs each year.
  - (Semi-)automated translation could lead to huge savings

Interest in machine translation

- Academic interest:
  - One of the most challenging problems in NLP research
  - Thought to require 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

- Goals of machine translation (MT) are varied, everything from *gisting* to rough draft
- Largest known application of MT: Microsoft knowledge base
  - Documents (web pages) that would not otherwise be translated at all

Language Weaver Arabic to English

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 appear in the relative calm 18ال governorates".

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

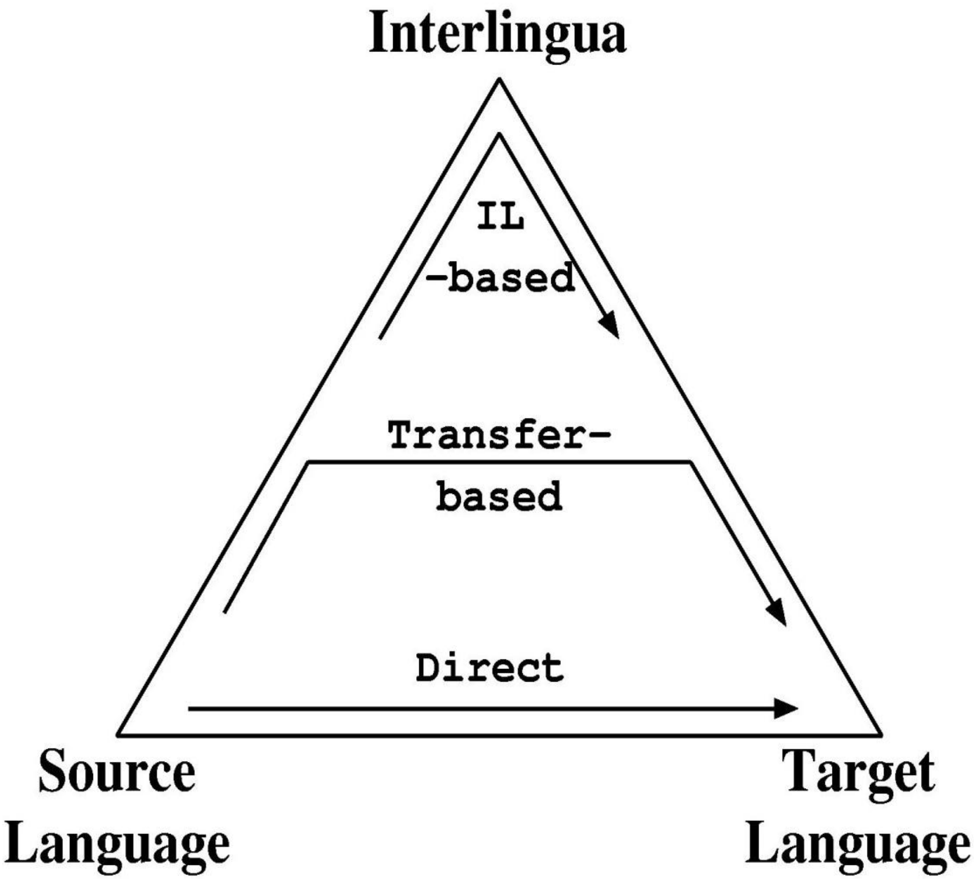
Alex Fraser
CIS, Uni München

Document versus sentence

- MT problem: generate high quality translations of **documents**
- However, all current MT systems work only at **sentence** level!
- Translation of independent sentences is a difficult problem that is worth solving
- But remember that important discourse phenomena are ignored!
  - Example: How to translate English *it* to French (choice of feminine vs masculine *it*) or German (feminine/masculine/neuter *it*) if object referred to is in another sentence?

Machine Translation Approaches

- Grammar-based
  - Interlingua-based
  - Transfer-based
- Direct
  - Example-based
  - Statistical
  - Neural



Modified from Vogel

Statistical versus Grammar-Based

- Often statistical and grammar-based MT are seen as alternatives, even opposing approaches – wrong !!!
- Dichotomies are:
  - Use probabilities – everything is equally likely (in between: heuristics)
  - Rich (deep) structure – 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

	No Probs	Probs
Flat Structure	EBMT	SMT
Deep Structure	XFER, Interlingua	Holy Grail

- Goal: structurally rich probabilistic models

Statistical Approach

- Using statistical models
  - Create many alternatives, called hypotheses
  - Give a score to each hypothesis
  - Select the best -> search
- Advantages versus rule-based
  - Avoid hard decisions
  - Speed can be traded with quality, no all-or-nothing
  - Works better than rule-based in the presence of unexpected input
- Disadvantages
  - Need data to train the model parameters
  - Difficulties handling structurally rich models, mathematically and computationally
  - Fairly difficult to understand decision process made by system

Neural Approach

- Predict one word at a time
  - Use structurally rich models!
  - Structure is learned by neural net, does not look like linguistic structure (e.g., no syntactic parse trees)
  - This *is* a statistical approach (but we've given it a new name)
- Advantages: same as previous statistical work
  - Additionally: much better generalization through learned rich structure!
  - Features are learned, rather than specified in advance
- Disadvantages
  - Like SMT, need data to train (learn) the model parameters
  - Heavy computing at training time, specialized hardware (GP-GPUs)
  - Nearly impossible to understand decision process made by system

Outline

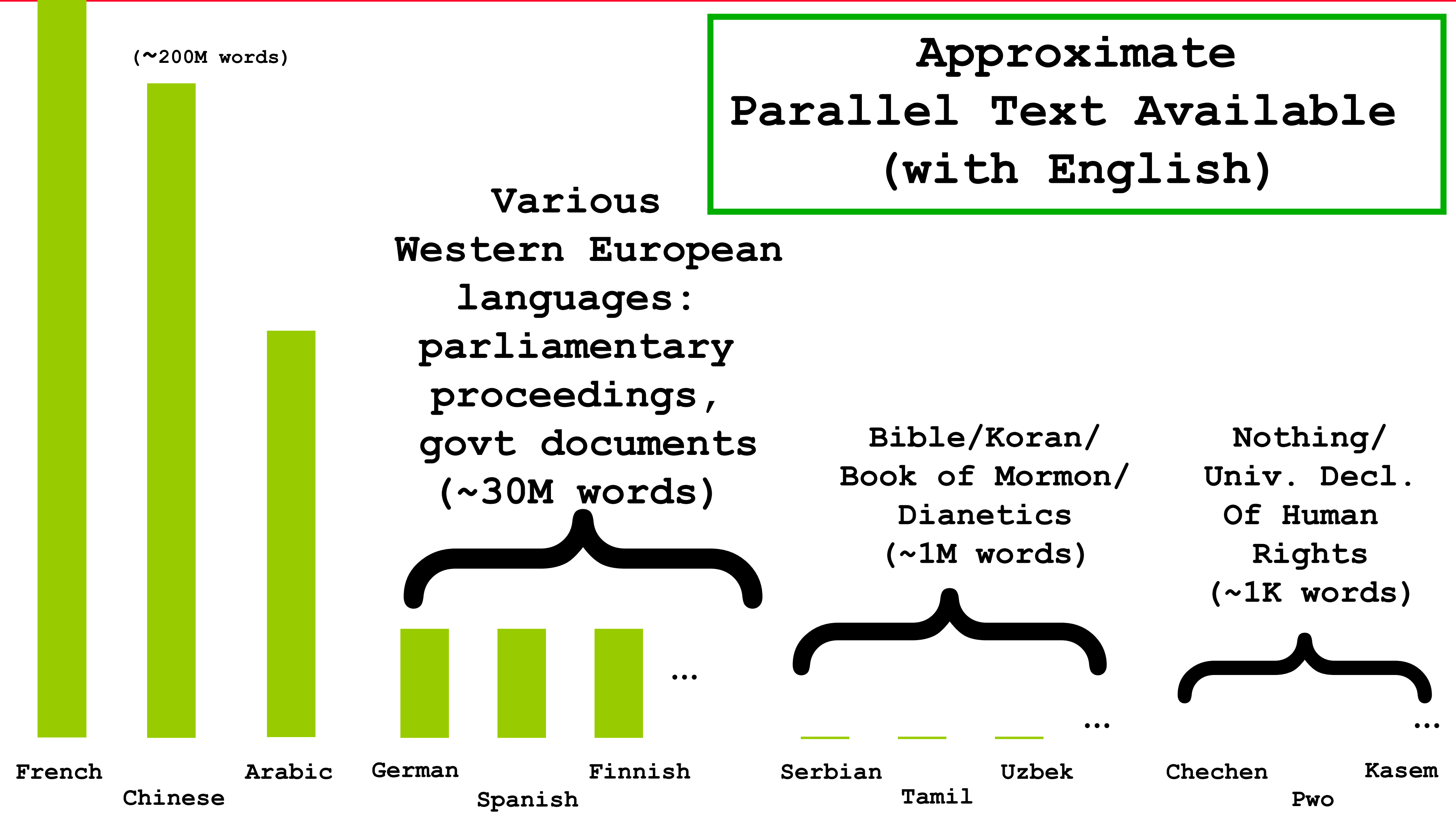
- *Machine translation*
- Data-driven machine translation
  - Parallel corpora
  - Sentence alignment
- Overview of statistical machine translation
- Evaluation of machine translation

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 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 .

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



Modified from Schafer&Smith

How to Build an SMT System

- Start with a large parallel corpus
  - Consists of document pairs (document and its translation)
- Sentence alignment: in each document pair automatically find those sentences which are translations of one another
  - Results in sentence pairs (sentence and its translation)
- Word alignment: in each sentence pair automatically annotate those words which are translations of one another
  - Results in word-aligned sentence pairs
- Automatically estimate a statistical model from the word-aligned sentence pairs
  - Results in model parameters
- Given new text to translate, apply model to get most probable translation

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

Modified from Dorr, Monz

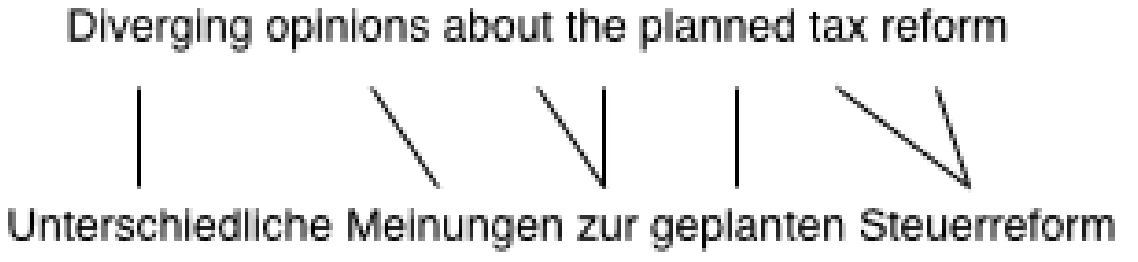
Sentence alignment

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

Modified from Dorr, Monz

Word alignments

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



Modified from Dorr, Monz

How to Build an SMT System

- Construct a function g which, given a sentence in the source language and a hypothesized translation into the target language, assigns a goodness score
  - g(die Waschmaschine läuft , the washing machine is running) = high number
  - g(die Waschmaschine läuft , the car drove) = low number

Using the SMT System

- Implement a search algorithm which, given a source language sentence, finds the target language sentence which maximizes g
- To use our SMT system to translate a new, unseen sentence, call the search algorithm
  - Returns its determination of the best target language sentence
- To see if your SMT system works well, do this for a large number of **unseen** sentences and evaluate the results

SMT modeling

- We wish to build a machine translation system which given a Foreign sentence “f” produces its English translation “e”
  - We build a model of P(e | f), the probability of the sentence “e” given the sentence “f”
  - To translate a Foreign text “f”, choose the English text “e” which maximizes P(e | f)

Noisy Channel: Decomposing P(e|f)

$$\underset{e}{\text{argmax}}\; P(e \mid f) = \underset{e}{\text{argmax}}\; P(f \mid e)\; P(e)$$

- P(e) is referred to as the “language model”
  - P (e) can be modeled using standard models (N-grams, etc)
  - Parameters of P (e) can be estimated using large amounts of monolingual text (English)
- P(f | e) is referred to as the “translation model”

SMT Terminology

- Parameterized Model: the form of the function g which is used to determine the goodness of a translation

  g(die Waschmaschine läuft, the washing machine is running)
  = P(e | f)

  P(the washing machine is running|die Waschmaschine läuft)=

SMT Terminology

- Parameterized Model: the form of the function g which is used to determine the goodness of a translation

g(die Waschmaschine läuft, the washing machine is running) = P(e | f)

P(the washing machine is running|die Waschmaschine läuft)=

What??

Unless we have seen exactly the input sentence in our training data, we can't GENERALIZE.

So we will decompose this translation into parts, so that we can generalize to new sentences.

SMT Terminology

- Parameterized Model: the form of the function g which is used to determine the goodness of a translation

  g(die Waschmaschine läuft, the washing machine is running)
  = P(e | f)

  P(the washing machine is running|die Waschmaschine läuft)=

  Suppose we translate:

  "die" to "the"

  "Waschmaschine" to "washing machine"

  "läuft" to "is running"

  (and further suppose we don't worry about word order…)

SMT Terminology

- Parameterized Model: the form of the function g which is used to determine the goodness of a translation

  g(die Waschmaschine läuft, the washing machine is running)
  = P(e | f)

  P(the washing machine is running|die Waschmaschine läuft)=

  n(1 | die) t(the | die)

  n(2 | Waschmaschine) t(washing | Waschmaschine)
  t(machine | Waschmaschine)

  n(2 | läuft) t(is | läuft) t(running | läuft)

  l(the | START) l(washing | the) l(machine | washing) l(is | machine)
  l(running | is)

SMT Terminology

- Parameters: lookup tables used in function g

P(the washing machine is running|die Waschmaschine läuft) =

n(1 | die) t(the | die)

n(2 | Waschmaschine) t(washing | Waschmaschine)

t(machine | Waschmaschine)

n(2 | läuft) t(is | läuft) t(running | läuft)

l(the | START) l(washing | the) l(machine | washing) l(is | machine) l(running | is)

0.1 x 0.1
x 0.5 x 0.4
x 0.3
x 0.1 x 0.1 x 0.1
x 0.0000001

SMT Terminology

- Parameters: lookup tables used in function g

P(the washing machine is running|die Waschmaschine läuft) =

n(1 | die) t(the | die)

n(2 | Waschmaschine) t(washing | Waschmaschine)

t(machine | Waschmaschine)

n(2 | läuft) t(is | läuft) t(running | läuft)

l(the | START) l(washing | the) l(machine | washing) l(is | machine) l(running | is)

0.1 x 0.1
x 0.5 x 0.4
x 0.3
x 0.1 x 0.1 x 0.1
x 0.0000001

Change "washing machine" to "car"
0.1 x 0.1
x 0.1 x 0.0001 n(1 | Waschmaschine)
t(car | Waschmaschine)
x 0.1 x 0.1 x 0.1
x also different

SMT Terminology

- Training: automatically building the lookup tables used in g, using parallel sentences
- One way to determine t(the|die)
  - Generate a word alignment for each sentence pair
  - Look through the word-aligned sentence pairs
  - Count the number of times „die“ is linked to „the“
  - Divide by the number of times „die“ is linked to any word
  - If this is 10% of the time, we set t(the|die) = 0.1

SMT Last Words

- Translating is usually referred to as decoding (Warren Weaver)
- SMT was invented by automatic speech recognition (ASR) researchers. In ASR:
  - P(e) = language model
  - P(f|e) = acoustic model
  - However, SMT must deal with word reordering!

Outline

- *Machine translation*
- *Data-driven machine translation*
  - *Parallel corpora*
  - *Sentence alignment*
  - *Overview of statistical machine translation*
- **Evaluation of machine translation**

Alex Fraser
CIS, Uni München

Evaluation driven development

- Lessons learned from automatic speech recognition (ASR)
  - Reduce evaluation to a single number
    - For ASR we simply compare the hypothesized output from the recognizer with a transcript
    - Calculate similarity score of hypothesized output to transcript
    - Try to modify the recognizer to maximize similarity
  - Shared tasks – everyone uses same data
    - May the best model win!
- These lessons widely adopted in NLP and Information Retrieval

Evaluation of machine translation

- We can evaluate machine translation at corpus, document, sentence or word level
  - Remember that in MT the unit of translation is the sentence
- Human evaluation of machine translation quality is difficult
- We are trying to get at the abstract usefulness of the output for different tasks
  - Everything from gisting to rough draft translation

Sentence Adequacy/Fluency

- Consider German/English translation
- **Adequacy**: is the meaning of the German sentence conveyed by the English?
- **Fluency**: is the sentence grammatical English?
- These are rated on a scale of 1 to 5

Modified from Dorr, Monz

Human Evaluation

INPUT: Ich bin müde. **(OR INPUT: Je suis fatigué.)**

	Adequacy	Fluency
Tired is I.	5	2
Cookies taste good!	1	5
I am tired.	5	5

Modified from Schafer, Smith

Automatic evaluation

- Evaluation metric: method for assigning a numeric score to a hypothesized translation
- Automatic evaluation metrics often rely on comparison with previously completed human translations

Word Error Rate (WER)

- WER: edit distance to reference translation (insertion, deletion, substitution)
- Captures fluency well
- Captures adequacy less well
- Too rigid in matching

Hypothesis = „he saw a man and a woman“

Reference = „he saw a woman and a man“

WER gives no credit for „woman“ or „man“ !

Position-Independent Word Error Rate (PER)

- PER: captures lack of overlap in *bag of words*
- Captures adequacy at single word (unigram) level
- Does not capture fluency
- Too flexible in matching

Hypothesis 1 = „he saw a man"

Hypothesis 2 = „a man saw he"

Reference = „he saw a man"

Hypothesis 1 and Hypothesis 2 get same PER score!

BLEU

- Combine WER and PER
  - Trade off between rigid matching of WER and flexible matching of PER
- BLEU compares the 1,2,3,4-gram overlap with one or more reference translations
  - BLEU penalizes generating short strings with the brevity penalty, precision for short strings is very high
  - References are usually 1 or 4 translations (done by humans!)
- BLEU correlates well with average of fluency and adequacy at corpus level
  - But not at sentence level!

BLEU discussion

- BLEU works well for comparing two similar MT systems
  - Particularly: SMT system built on fixed training data vs. Improved SMT system built on same training data
  - Other metrics such as METEOR extend these ideas and work even better – ongoing research!
- BLEU does not work well for comparing dissimilar MT systems
- There is no good automatic metric at sentence level
- There is no automatic metric that returns a meaningful measure of **absolute** quality

Language Weaver Arabic to English

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

Alex Fraser
CIS, Uni München

- Questions?

- Thank you for your attention!