INF5820/INF9820

LANGUAGE TECHNOLOGICAL APPLICATIONS

Jan Tore Lønning, Lecture 2, 31 Aug. 2016
jtl@ifi.uio.no

Machine Translation, lecture 2

- **Why is (machine) translation hard?**
  - Typological differences
  - Translational differences
- Evaluation in MT
  - Human evaluation of MT Quality
  - Automatic evaluation in Language Technology
  - Word precision and recall
  - BLEU

Why (machine) translation is hard.

Why can’t we just use a dictionary?

Because:

- Languages are constructed differently (typology)
- Translation is not one-to-one

Language typology: morphology

- Number of morphemes per word
  - Isolating: 1,
    - Chinese, Vietnamese
  - Synthetic: >1
  - Polysenthetic: >>1
- Morphemfusion:
  - Agglutanitive
    - putting morphemes after each other
    - Japanese, Turkish, Finnish, Sami
  - Fusion
    - Russian

Washakotya'tawitsherahetkvhta'se
"He made the thing that one puts on one's body ugly for her"
"He ruined her dress"
(Mohawk, polysynthetic, Src: Wikipedia)

(3.1) uygarlaştıramadıklarımızdanmışsınızcasına

uygar	*+laş*	*+tır*	*+ama*	*+dık*	*+lar*	*+ımız*	*+dan*	*+mış*	*+sınız*	*+casına*
civilized	+BEC	+CAUS	+NABL	+PART	+PL	+P1PL	+ABL	+PAST	+2PL	+AsIf

"(behaving) as if you are among those whom we could not civilize"

Turkish, agglutanitive, polysynthetic J&M, Ch. 3

Language typology: Syntax

- Word order:
  - Subject-Verb-Object, SVO
  - SOV
  - VSO
- Prepositions vs postpositions
- Modifiers before or after:
  - Red wine vs. vin rouge
- Verb-framed vs. satelite-framed
  - Marking of direction
  - Marking of manner

Jorge swam across the river.
Jorge cruzó a nado el río.

Language typology: Markers

- One language may contain a marker which is lacking – or very different – in another language:
  - Tense
  - Aspect:
    - She smiles vs she is smiling
  - Case
  - Definiteness

Translational discrepancies

- Translation is not only about typological differences
- Even between typologically similar languages, the translation is not always one-to-one

Ambiguity!

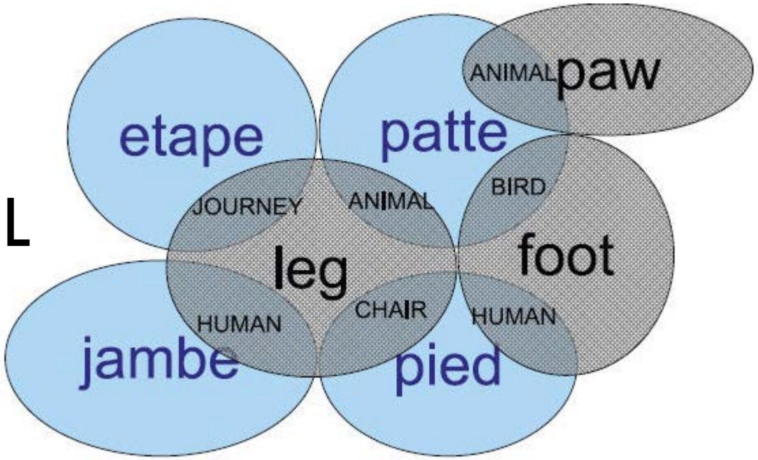
Lexical ambiguities in SL

Word form	Norw: ”dekket”				
POS	Noun			Verb	Adjective
Base form	”dekk”		”dekke”		
Homonymy	”dekk på båt”	”dekk på bil”			
Polysemy					
Gloss	”deck”	”tire”			

More examples		
	Norw	English
Verb/noun	løp, løper, bygg, bygget	fish, run, runs, ring
Homonymy	bygg (Noun), ball	bank, ball, bass
Polysemy	hode	head, bass (music)

Lexical choice in transfer

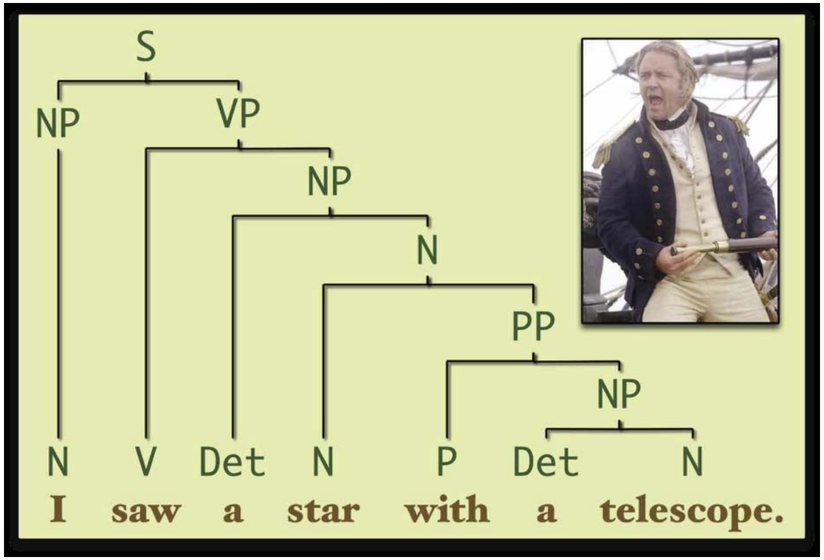
- The TL may make more distinctions than SL
  - No: *tak*, Eng: *ceiling/roof*
  - Eng: *grandmother*,
    No: *farmor/mormor*
- Context dependent choice in TL
  - *Strong tea, powerful government*
  - *Dekke på bordet* → *set the table*
  - *Dekke bordet* → *set/cover the table*

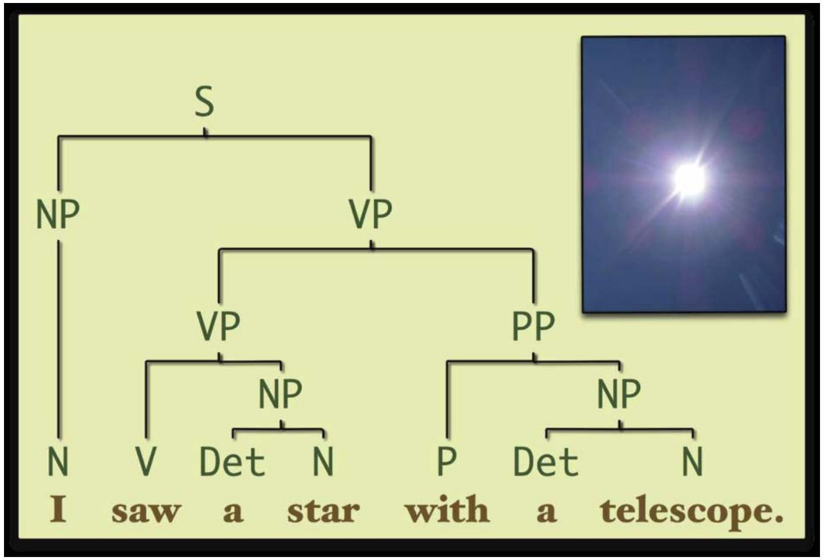


- Languages may draw different distinctions
  - *Morgen* – *morning*, *legg* – *leg*

Syntactic ambiguities in SL

- Global ambiguities





- Local ambiguities:
  - De kontrollerte bilene → They controlled the cars
  - De kontrollerte bilene er i orden → The controlled cars are OK

Structural mismatch

- Thematic divergence/argument switching
  - E: I like Mary.
  - S: Mary me gusta.
- Head switching:
  - E: Kim likes to swim.
  - G: Kim schwimmt gern.

- More divergence:
  - N: Han heter Paul.
  - E: His name is Paul.
  - F: Il s'appell Paul.
- Idiomatic expressions

Beyond sentence meaning

- Tracking the referent,
  No: den/det   han/hun
- Metaphors, idioms

- Changre,
- Rhime, rythm
- Deliberate ambiguity, humor
- …

Machine Translation, lecture 2

- Why is (machine) translation hard?
  - Typological differences
  - Translational differences
- Evaluation in MT
  - **Human evaluation of MT Quality**
  - Automatic evaluation in Language Technology
  - Word precision and recall
  - BLEU

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

NIST evaluation task 2001, from Koehn: SMT

Translation quality – Human eval.

- Given output of MT system + either
  1. Source text + reference translation (bilingual evaluator)
  2. Source text only (bilingual evaluator)
  3. Reference translation only (monolingual evaluator)
  4. Nothing (output only) (only fluency)
- Rate the translations (one sentence a time)
- Across several dimensions, typically
  - Adequacy: Does the output convey the same as the original/reference translation?
  - Fluency: Is this good target language?
  - and maybe several other dimensions

Judge Sentence

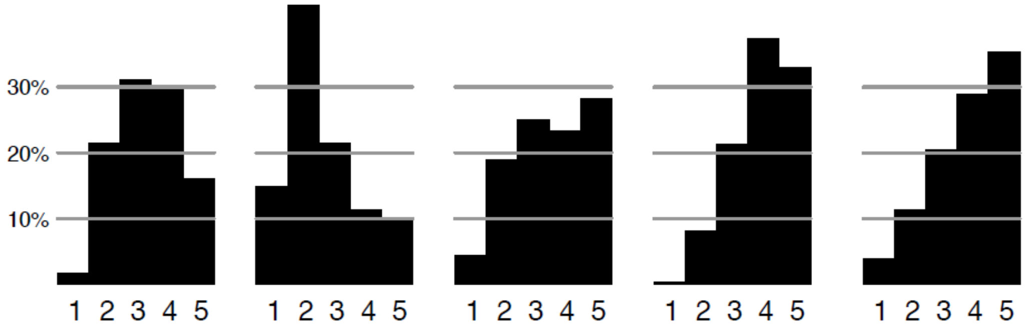
You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

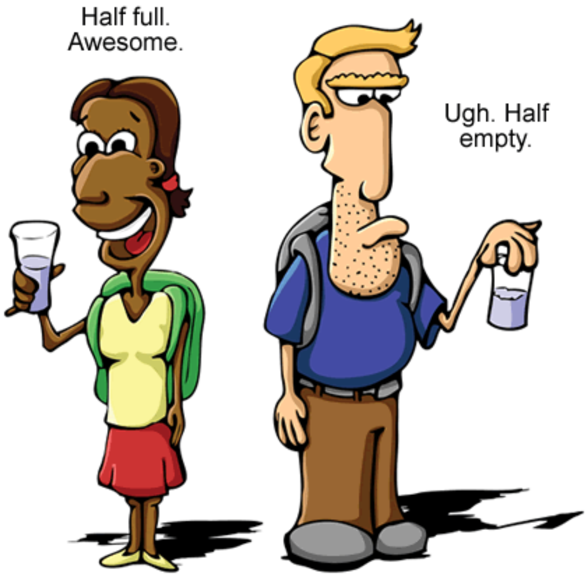
Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Challenges in human TQ eval.



- What's in a number?
  - People use the scales differently
  - Normalize?
- More reliable alternative:
  - Evaluate several systems at once
  - Which translation is better?



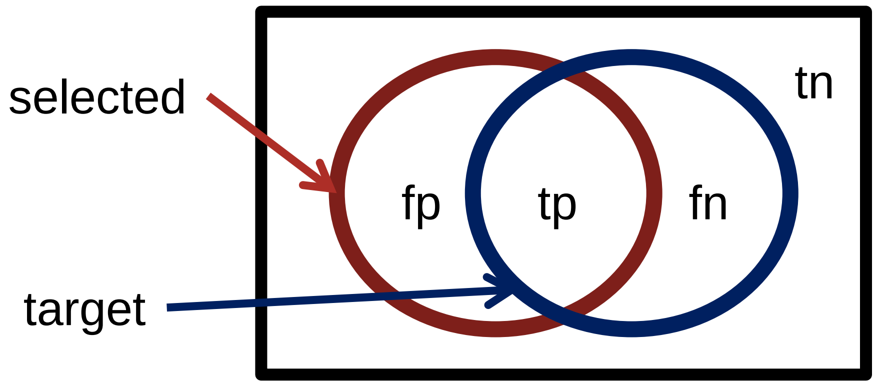
Machine Translation, lecture 2

- Why is (machine) translation hard?
  - Typological differences
  - Translational differences
- Evaluation in MT
  - Human evaluation of MT Quality
  - **Automatic evaluation in Language Technology**
  - Word precision and recall
  - BLEU

Evaluation in language technology

- Example 1: Tagging
  - Task: Assign part of speech tags to words in text
    - The/DT grand/JJ jury/NN commented/VBD …
  - Gold standard: A hand-annotaded corpus
  - Run your tagger on the gold standard
  - Compare the results with the gold standard
  - Accuracy: #(correct tags)/#words
- Experimental set up:
  - Split an annotaded corpus in two parts:
    - Training
    - Testing (=gold standard) not used in training

Common evaluation measures in LT



		Actual (gold)	
		target	Not target
System perform	selected	tp: True positive	fp: False positive
	Not selected	fn: False negative	tn: True negative

- Recall = $\frac{tp}{tp+fn}$
- Precision = $\frac{tp}{tp+fp}$
- F-score = $\frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}}$
- F_1 = $\frac{1}{0.5\frac{1}{P}+(1-0.5)\frac{1}{R}}=\frac{2PR}{R+P}$

Some remarks

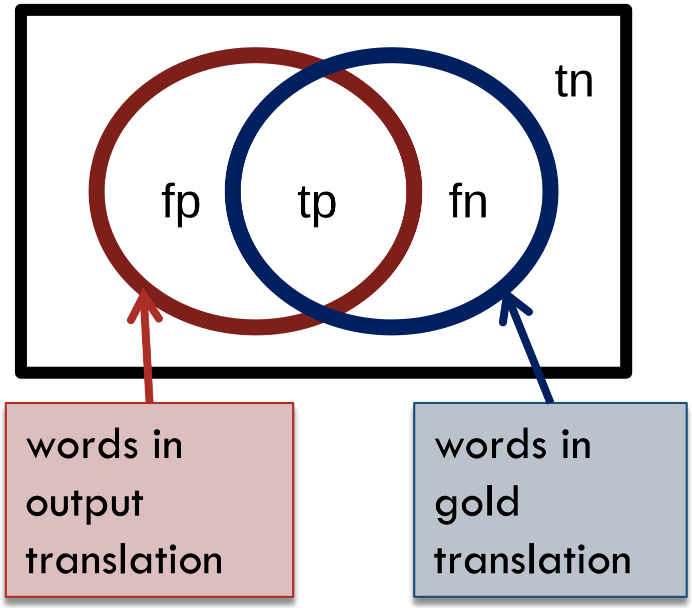
- Precision and recall:
  - Comes from Information Retrieval (IR)
  - Have become (too?) popular in language technology
- Useful when:
  - There is more than one target/correct answer
  - The targets are known
  - The true negatives are many, uninteresting or unknown
  - The targets are not ranked
- Statistical significance tests are more easily available for accuracy than for P, R, F

Machine Translation, lecture 2

- Why is (machine) translation hard?
  - Typological differences
  - Translational differences
- Evaluation in MT
  - Human evaluation of MT Quality
  - Automatic evaluation in Language Technology
  - **Word precision and recall**
  - BLEU

Adapting P, R, F to MT-eval

- Precision $= \dfrac{correct}{output.length}$
- Recall $= \dfrac{correct}{ref.length}$
- $F_1 =$



$$\frac{2}{\frac{1}{R}+\frac{1}{P}} = \frac{2}{\frac{ref.length}{correct}+\frac{output.length}{correct}} = \frac{2correct}{output.length + ref.length}$$

Precision and Recall of Words

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

- Precision

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

- Recall

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

- F-measure

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Precision and Recall

SYSTEM A: Israeli officials ~~responsibility~~ ~~of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible

Metric	System A	System B	
precision	50%	100%	
recall	43%	~~100%~~	$\frac{6}{7} \approx 0.86$
f-measure	46%	~~100%~~	$\frac{12}{13} \approx 0.92$

flaw: no penalty for reordering

Position-independent error rate

- Similar measure to (word) recall+precision
- Reports mistakes – not correctness
- We skip the details - formula

Word Error Rate

- Minimum number of editing steps to transform output to reference

  match: words match, no cost
  substitution: replace one word with another
  insertion: add word
  deletion: drop word

- Levenshtein distance

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference-length}}$$

Levenshtein distance used in

- spell-checking
- OCR
- Translation memory

Example

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
security	7	6	5	6	7	6	5

Metric	System A	System B
word error rate (WER)	57%	71%

Machine Translation, lecture 2

- Why is (machine) translation hard?
  - Typological differences
  - Translational differences
- Evaluation in MT
  - Human evaluation of MT Quality
  - Automatic evaluation in Language Technology
  - Word precision and recall
  - **BLEU**

BLEU

- A Bilingual Evaluation Understudy Score
- Main ideas:
  - Use several reference translations
  - Count precision of n-grams:
    - For each n-gram in output:
      does it occur in at least one reference?
  - Don’t count recall but use a penalty for brevity
    - Why not recall?

BLEU

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram, C, C.refs)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count\ (n-gram, C)}$$

- Candidates:
  - the set of sentences output by trans. system
- Count(n-gram, C):
  - the number of times *n-gram* occurs in C
- $\text{Count}_{\text{clip}}$(n-gram, C, C.refs):
  - the number of times the *n.gram* occurs in both
    - C and
    - the reference translation for the same sentence where *n.gram* occurs most frequent

- Technicality:
  - If the same n-gram has several occurrences in a candidate translation sentence, it should not be counted more times than the number of occurrences in the reference sentence with the largest number of occurrences of the same n-gram.

Example, p_3

- Hyp, C:
  - One of the girls gave one of the boys one of the boys.
- C-Refs:
  - A girl gave a boy one of the toy cars
  - One of the girls gave a boy one of the cars.

#

Example, p_3

- Hyp, C:
  - <u>One of the </u>girls gave <u>one of the </u>boys <u>one of the </u>boys.
- C-Refs:
  - A girl gave a boy <u>one of the </u>toy cars
  - <u>One of the </u>girls gave a boy <u>one of the </u>cars.
- Count_clip(one of the, C, C-refs)=2

one of the	of the girls	the girls gave	girls gave one
2 (3)	1	1	0 (1)

gave one of	of the boys	the boys one	boys one of
0 (1)	0 (2)	0 (1)	0 (1)

- $P_3 = 4/11$

BLEU

- How to combine the n-gram precisions?

$$p_1 \times p_2 \times \cdots \times p_n = \prod_{i=1}^{n} p_i$$

- Remember

$$\ln(\prod_{i=1}^{n} p_i) = \ln(p_1 \times p_2 \times \cdots \times p_n) = \ln(p_1) + \ln(p_2) + \cdots + \ln(p_n) = \sum_{i=1}^{n} \ln p_i$$

- One can add weights, typically $ai = 1/n$

$$\ln(p_1^{a1} \times p_2^{a2} \times \cdots \times p_n^{an}) = a1\ln(p_1) + a2\ln(p_2) + \cdots + an\ln(p_n)$$

- How long n-grams?
  - Max 4-grams seems to work best

Brevity penalty

- c is the length of the candidates
- r is the length of the reference translations:
  - for each C choose the R most similar in length

$$c = \sum_{C \in Candidates} length(C)$$

$$r = \sum_{C \in Candidates} length(R.sim.C)$$

- Penalty applies if c < r:
  - BP = 1 if c $\geq$ r
  - BP = $e^{(1-r/c)}$ otherwise

This is correct
Error in K:SMT

- $BLEU = BP \cdot \exp \sum_{i=1}^{n} w_n \ln p_i$
- $\ln BLEU = \min(1 - \frac{r}{c}, 0) + \sum_{i=1}^{n} w_n \ln p_i$