INF5820/INF9820

LANGUAGE TECHNOLOGICAL APPLICATIONS

Jan Tore Lønning, Lecture 2, 29 Sept.
jtl@ifi.uio.no

Machine Translation, lecture 2

- The challenge of MT
  - Why is (machine) translation hard?
    - Typological differences
    - Translational differences
  - MT in practice
  - The history of MT
- Evaluation in MT
  - Human evaluation of MT Quality
  - Evaluation in Language Technology
  - Automatic MT-evaluation:
    - Word precision and recall

Language typology

- 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

- 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!

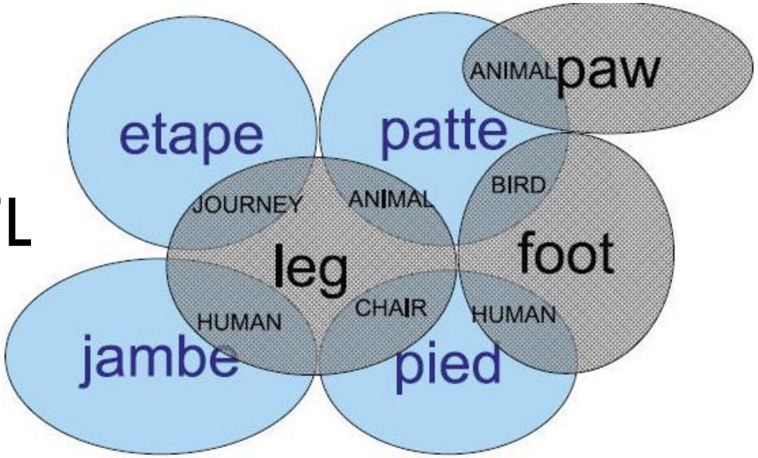
Lexical ambiguities in SL

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

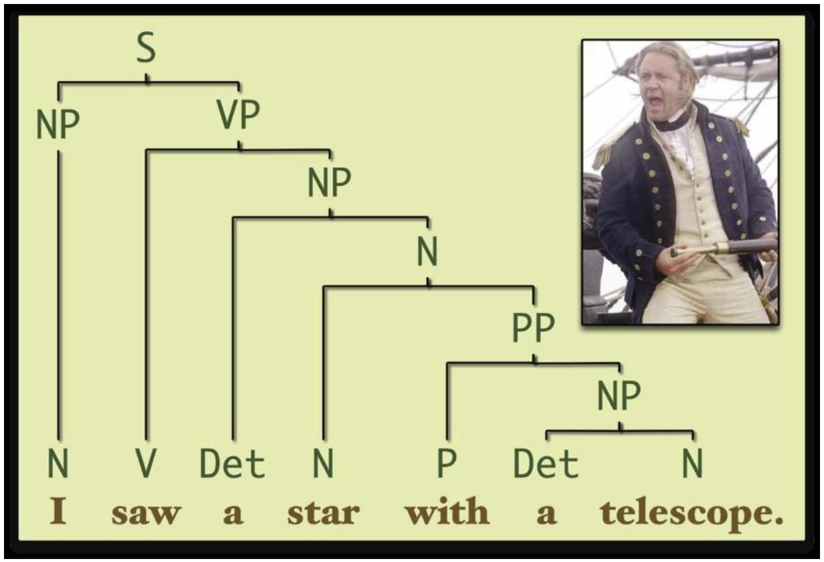
- The TL may make more distinctions than SL
  - No: *tak*, Eng: *ceiling/roof*
  - Eng: *grandmother*,
    No: *farmor/mormor*
- Context dependent choice in TL
  - *Strong tea, powerful government*
  - *Dekke på bordet* → *set the table*
  - *Dekke bordet* → *set/cover the table*

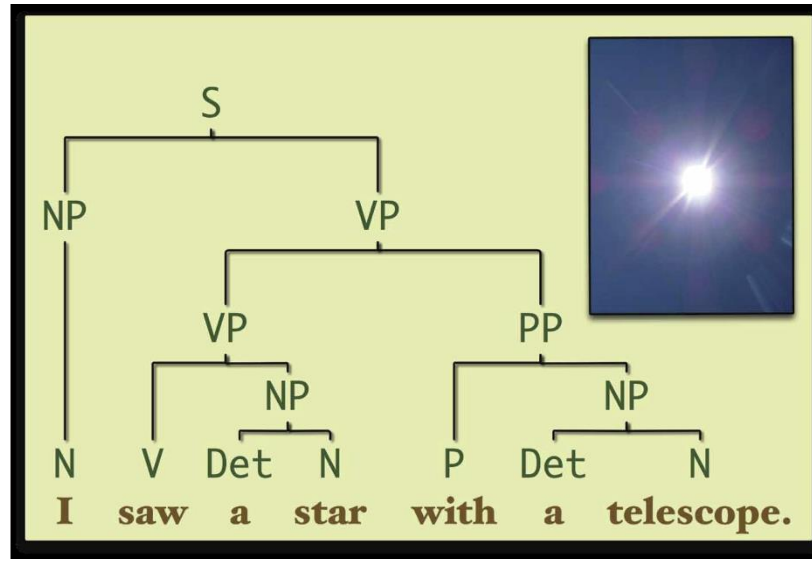


- Languages may draw different distinctions
  - *Morgen* – *morning*, *legg* – *leg*

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

- Larger units, paragraphs
- Tracking the referent, No: den/det
- Metaphors, idioms
- Changre,
- Rhime, rythm
- Deliberate ambiguity, humor
- …

Limitations

- På et grunnleggende nivå, utfører MT enkel substitusjon av ord i ett naturlig språk for ord i en annen, men det alene vanligvis ikke kan produsere en god oversettelse av en tekst, fordi anerkjennelse av hele setninger og deres nærmeste kolleger i målspråket er nødvendig. Løse dette problemet med korpus og statistisk teknikker er en raskt voksende felt som fører til bedre oversettelser, håndtering forskjeller i språklig typologi , oversettelse av idiomer , og isolering av anomalier.
- Google translate fra→

- On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus and statistical techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms, and the isolation of anomalies.
- Wikipedia: Machine translation

Machine Translation, lecture 2

- The challenge of MT
  - Why is (machine) translation hard?
  - MT in practice
  - The history of MT
- Evaluation in MT
  - Human evaluation of MT Quality
  - Evaluation in Language Technology
  - Automatic MT-evaluation:
    - Word precision and recall

Ultimate goal

Fully Automatic High-Quality (unrestricted) Translation (FAHQT)

- Not succeeded so far
- In practice, renounce on some of the goals

In practice

Fully Automatic High-Quality (~~unrestricted~~) Translation

- Restricted language
  - Example: METEO
    - Translated weather forecasts between English and French in Canada, 1981-2001

In practice

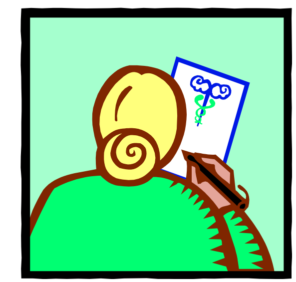
Fully Automatic ~~High-Quality~~ (unrestricted) Translation

- Lower Quality
  - Acceptable when:
    - To get an idea of a text (should I get it translated?)
    - Interactive communication where the parts may clarify
    - Web
    - Example: family letters

MT+human

~~Fully Automatic~~ High-Quality (unrestricted) Translation

Pre-processing



Post-processing

- Semi-automatic
- User-studies have indicated:
  - May be profitable
  - Boring and unpopular by translators

Machine-aided translation

~~Fully Automatic~~ High-Quality (unrestricted) Translation

- Machine-aided translation
  - Spell checker
  - Dictionary
  - Translation memory
    - (Ex: User manual for a new version of a system)
    - In common use since the 1990s
    - ”Trados” most used

Integrating human and machine

~~Fully Automatic~~ High-Quality (unrestricted) Translation

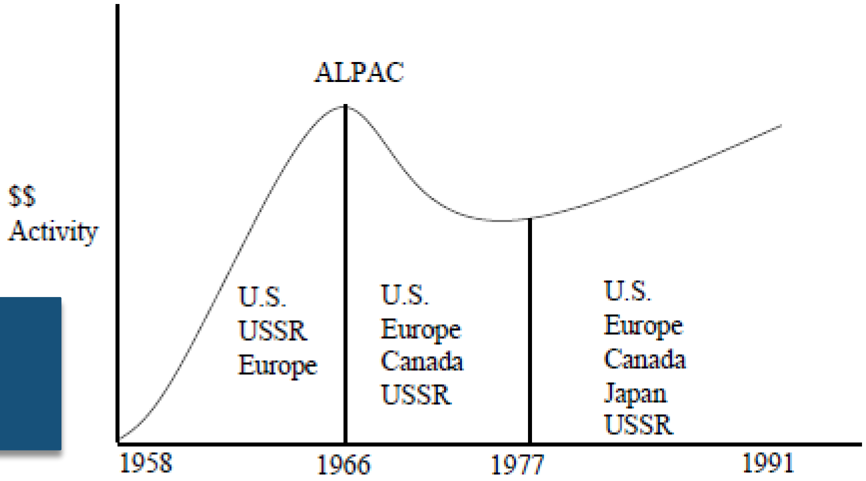
- "Translator's workbench"
  - Combining MT and human translation interactively
  - A long-time vision
- Starting to appear:
  - SDL: acquired and combines
    - Trados
    - Language Weaver, commercial SMT
  - Google Translator Toolkit

Machine Translation, lecture 2

- The challenge of MT
  - Why is (machine) translation hard?
  - MT in practice
  - The history of MT
- Evaluation in MT
  - Human evaluation of MT Quality
  - Evaluation in Language Technology
  - Automatic MT-evaluation:
    - Word precision and recall

History

- 1950s: great optimism(FAHQT)
  - First direct approach
  - Spawned interest in syntax
- 1960s: too difficult
  - Bar-Hillel lost faith
  - The ALPAC-report
- 1980s renew interest:
  - Japan
  - EU, Eurotra



From Dorr et al
A Survey of Current Paradigms in Machine Translation, 1999

Our time (1992→)

Applications:

- Off the shelf PC software
- WWW
- Mobile devices
- Interactive workbenches for translators
- New markets: China

Scientific:

- Speech translation
- SMT:
  - Developed since 1990
  - On the market 2003
  - Used by Google 2005:
    - Many pairs
    - English as IL
  - Predictable errors

Machine Translation, lecture 2

- The challenge of MT
  - Why is (machine) translation hard?
  - MT in practice
  - The history of MT
- Evaluation in MT
  - Human evaluation of MT Quality
  - Evaluation in Language Technology
  - Automatic MT-evaluation:
    - Word precision and recall

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

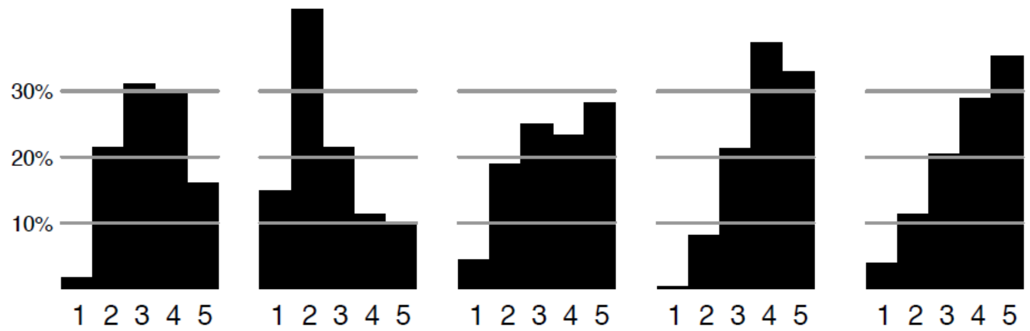
You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

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

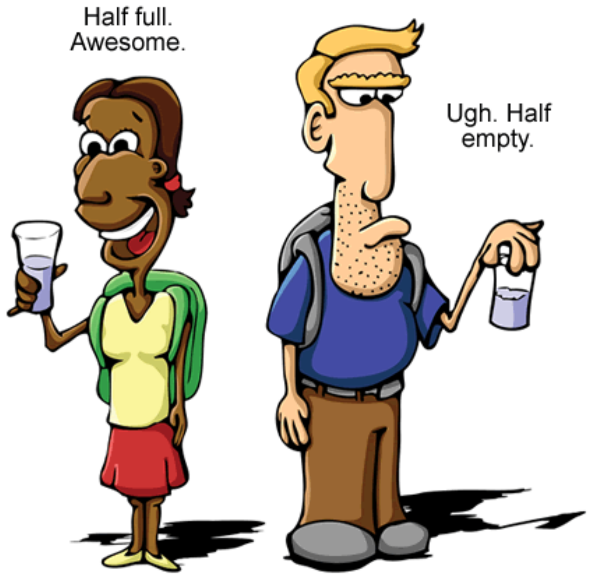
Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

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?



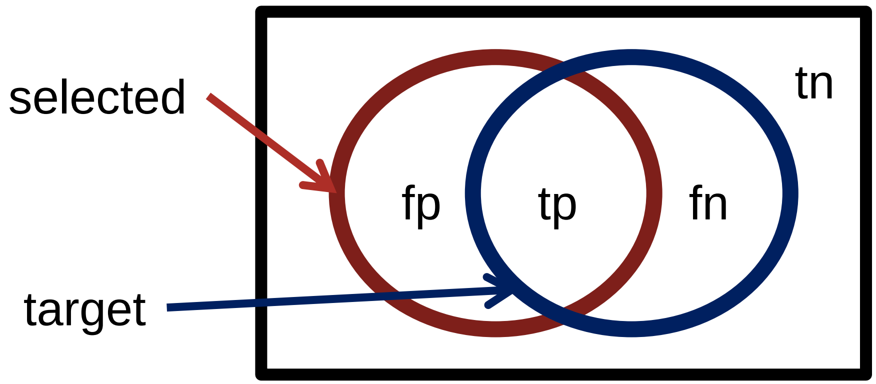
Machine Translation, lecture 2

- The challenge of MT
  - Why is (machine) translation hard?
  - MT in practice
  - The history of MT
- Evaluation in MT
  - Human evaluation of MT Quality
  - Evaluation in Language Technology
  - Automatic MT-evaluation:
    - Word precision and recall

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

Some remarks

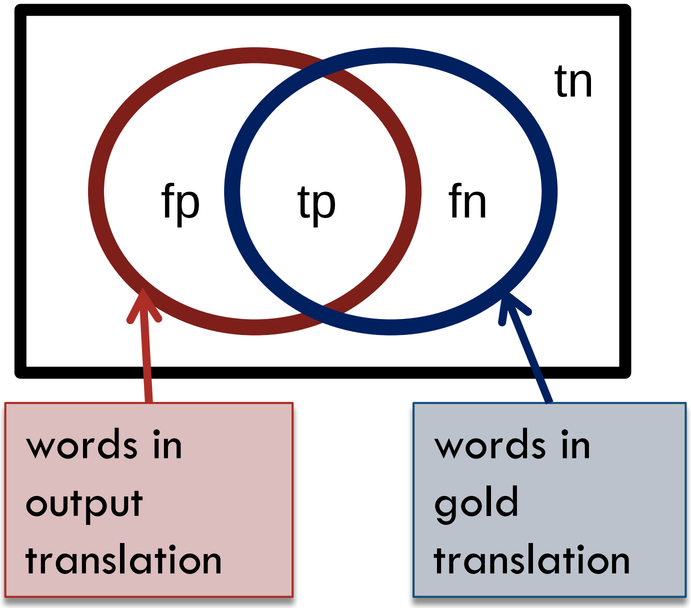
- Precision and recall:
  - Comes from Information Retrieval (IR)
  - Have become (too?) popular in language technology
- Useful when:
  - There is more than one target/correct answer
  - The targets are known
  - The true negatives are many, uninteresting or unknown
  - The targets are not ranked
-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

Machine Translation, lecture 2

- The challenge of MT
  - Why is (machine) translation hard?
  - MT in practice
  - The history of MT
- Evaluation in MT
  - Human evaluation of MT Quality
  - Evaluation in Language Technology
  - Automatic MT-evaluation:
    - Word precision and recall

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%~~
f-measure	46%	~~100%~~

$6/7 \approx 0.86$

$12/13 \approx 0.92$

flaw: no penalty for reordering

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

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

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%