INF5820/INF9820

LANGUAGE TECHNOLOGICAL APPLICATIONS

H2016

Jan Tore Lønning, Andrei Kutuzov
jtl@ifi.uio.no

Welcome

Today

- Course overview
- Starting Machine Translation

Two applications

Machine translation

- First part of semester
- Jan Tore Lønning (jtl)

Distributional Semantics: Extracting Meaning from Data

- Second part
- Andrei Kutuzov (andreku)

Language Technology Group (LTG), 7. floor

Classes

- Wednesdays 14.15-16
  - Lectures
  - OJD 2453 Perl
- Thursdays10.15-12
  - Group/lectures
  - OJD 1454 Sed

- In MT-part:
- On average meet 3 times in 2 weeks:
  - Two lectures
  - One «group»
- Some weeks
  - skip Wednesday
  - or skip Thursday
- Some lectures on Wednesdays

Obligatory assignments

- 3 obligatory assignments
  - MT 1: 23 September
  - MT 2: 21 October
  - Distributional Semantics: 18 November

Exam

- Written exam
- 12 December at 1430

- Next exams:
  - Spring 2017:
    - You must have completed oblig.s this fall (or earlier)
  - Fall 2018
    - But study reform and new courses from 2018

INF5820

- http://www.uio.no/studier/emner/matnat/ifi/INF5820/index-eng.xml
- Alternates with
  - INF5830 Natural Language Processing

Recommended prior knowledge

- INF4820 - Algorithms for artificial intelligence and natural language processing
  - The parts on distributional semantics
- Some knowledge of probabilities is an advantage (but we will provide crash course)
  - Probability theory
  - N-grams
  - Hidden Markov Models
- Useful:
  - Knowledge of linguistics/language
  - Computational linguistics, INF2820, INF1820

Why study Machine Translation?

- Importance:
  - Globalization
  - Most people don’t understand English
  - Most of the internet is not in English
    - and growing
  - Translation is a multi billion $ market
- Scientific:
  - Longest tradition in Language Technology
  - It is in use – and the use is growing
  - Interesting technology and algorithms
  - More to do!

What we will study in MT

1. MT overview
2. MT evaluation
3. Statistical MT,
   - The main part
4. Additional themes as time permits
   - Hybrid methods
   - Statistical syntactic transfer

Literature

- J&M, ch. 25
- Koehn, in particular, Part II Core methods: ch. 4-8
- A few papers

13

Machine Translation

Machine Translation

1. Translation – by humans and machines
2. Traditional approaches to MT
   1. Direct
   2. Interlingua
   3. Transfer
3. Empirical approaches:
   1. Example-based MT (EBMT)
   2. Statistical MT - SMT
4. History
5. Why is (machine) translation hard?

Machine Translation

- Active research field since 1949,
  - In the 1950s MT was not only the most important NLP/computational linguistics field, it was the only one
  - IBM 1954 press release
- Interest, results and funding have varied over time
- Today:
  - Fully-automatic text-translation: Systran, Google
  - Speech-translation: Mobile phones
  - Aid for professional translators: SDL trados

Two types of approaches to NLP

Rule-based	Empirical

Rule-based

- Build a declarative model using
  - Linguistics
  - Logic
- Algorithms
- How does it fit data?

Empirical

- Start with naturally occurring text
- What information can we get?
  - Statistics/Machine learning
- Use this to reproduce the examples

Applied to MT

Rule-based	Empirical

Rule-based

- Which linguistic information should be included,
  - syntax?
  - semantics?
- Approaches
  - Direct translation
  - Syntax-based transfer
  - Semantic-based transfer
  - ..

Empirical

- Example-based translation
- Statistical machine translation (SMT)
  - Word-based
  - Phrase-based
  - Syntactic

Translation

- What does it mean to translate a text T from a source language SL to a target language TL?

- T’ should express ”the same” in TL as T does in SL
  - What does it mean to say the same?
  - Is it possible?

Goal of translation

What to preserve?

- What to preserve?
  - Content
  - Transfer the same ”feeling”
  - Genre, style, rhyme
    - Slang vs. church language
- Should Ibsen be translated into contemporary English or late 19th century English?

Consequences for MT

- Some problems are avoided if we stick to technical texts.
- But a lesson:
  - There is not always a unique best translation!
  - ”Give and take”

How to translate?

- We all know (at least) 2 languages?
  - How do we proceed if we are to translate between them

- How would you proceed to translate between two languages you do not know?

”Realskolealgoritmen”

S.N.def.sg			V.pr	V.pa.part	H.D.3p.sg	O.A.indef.pl	
Jenta	fra	byen	har	gitt	ham	noen røde	epler
Mädchen	von	Stadt	haben	geben	er	einige rot	Apfel

Das Mädchen von der Stadt hat ihm einige rote Äpfel gegeben

1. Identify verb, syntactic function, case
2. And morphosyntactic features:
   - definiteness, number, person, form, tense, …
3. Translate the lexemes (dictionary)
4. Properties of the target lexemes: gender, arguments, agreement
5. Inflection: Case, number, person, gender, def., tense, agr. …
6. Word order

Does it work?

- All language pairs aren’t as similar as N & German
- All Norwegian-German translations aren’t that similar to e.o.
- The ”algorithm” is not run by a machine as is:
  - Identify verb(s)
  - Identify syntactic function
  - Word order
  - Lexical ambiguities

Machine Translation

1. Translation – by humans and machines
2. Traditional approaches to MT
   1. Direct
   2. Interlingua
   3. Transfer
3. Empirical approaches:
   1. Example-based MT (EBMT)
   2. Statistical MT - SMT
4. History
5. Why is (machine) translation hard?

1. Direct MT

- Bilingual, one direction
- Basic steps:
  1. Morph. analysis of source sentence
  2. Dict. lookup
  3. Morph. processing of target words
  4. Word reordering
- Possible refinements:
  - Homograph analysis
  - Compound analysis
  - Preposition translation
  - Idioms
  - …

2.Interlingua

- A universal meaning representation language (lingua franca)
- Steps:
  - Analyze the source language sentence
  - Resulting in an interlingua representation
  - From this, generate sentence in target language

```
(*BE-PREDICATE
  (attribute
       (*REQUIRED
          (degree positive)))
  (mood declarative)
  (predicate-role attribute)
  (punctuation period)
  (qualification
        (*QUALIFYING-EVENT
            (event
               (*PERSIST
                   (argument-class theme)
                   (mood declarative)
                   (tense present)
                   (theme
                      (*ERROR
                         (number  (:OR mass singular))
                         (reference definite)))))
            (extent  (*CONJ-if))
            (topic +)))
   (tense present)
   (theme
       (*SERVICE
           (number  (:OR mass singular))
           (reference no-reference))))
```

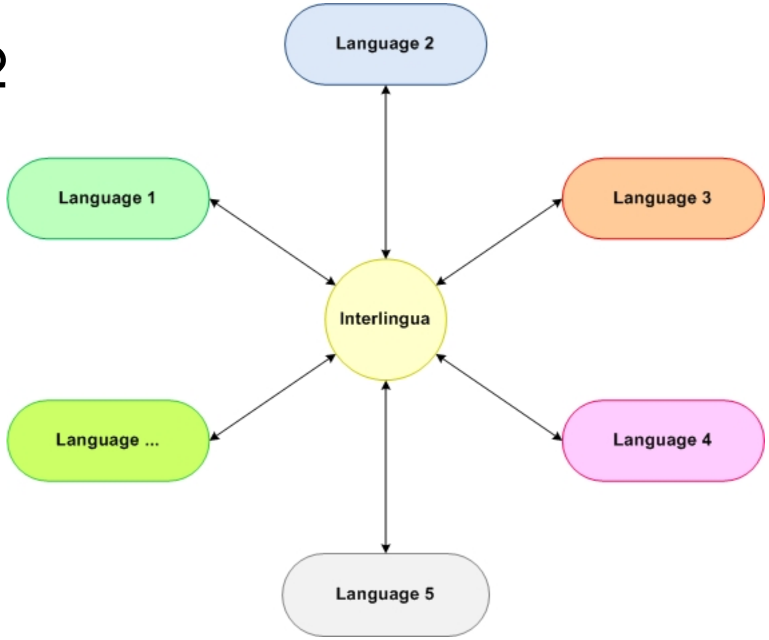
Figure 3: KANT Representation of *If the error persists, service is required.*

IL example from Dorr, Hovy, Levin:

Natural Language Processing and Machine Translation
Encyclopedia of Language and Linguistics, 2nd ed. (ELL2).
Machine Translation: Interlingual Methods

Interlingua strength

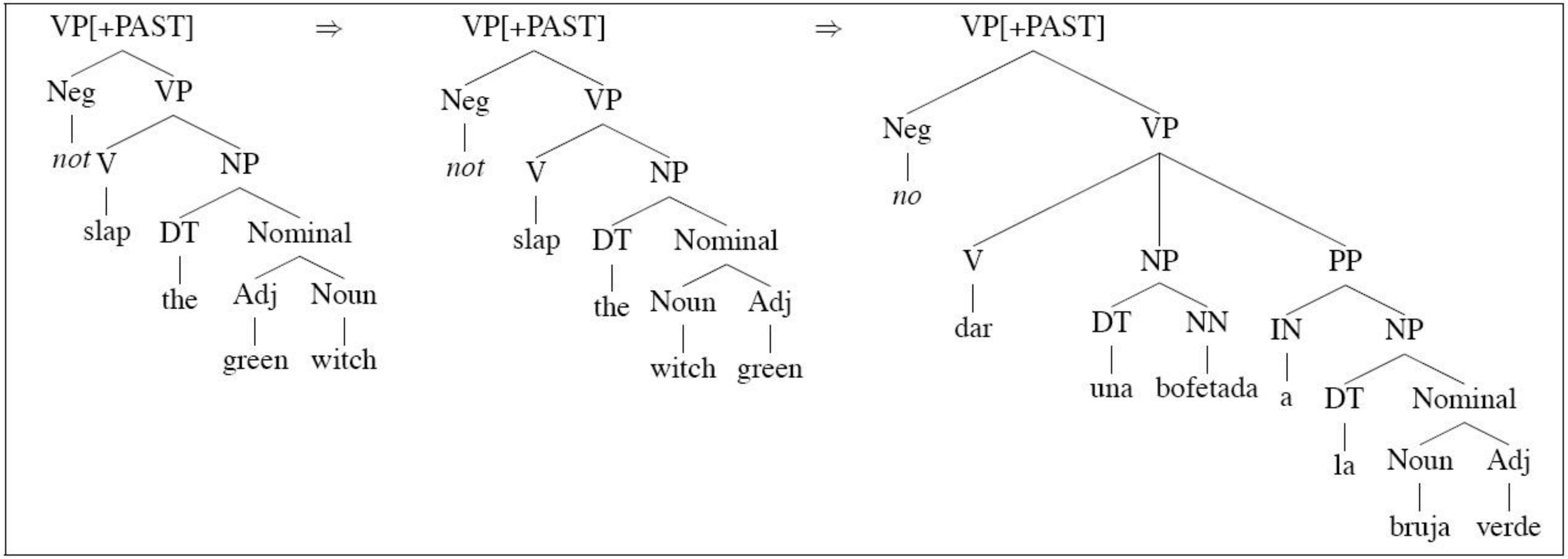
- Translation between many languages.
- One analysis module and one generation module per languages
- Example 17 languages:
  - Direct 17*16 modules (=272
  - Interlingua 2*17 (=34)
- Language18:
  - Direct +(2*17)
  - Interlingua +2



3. Transfer

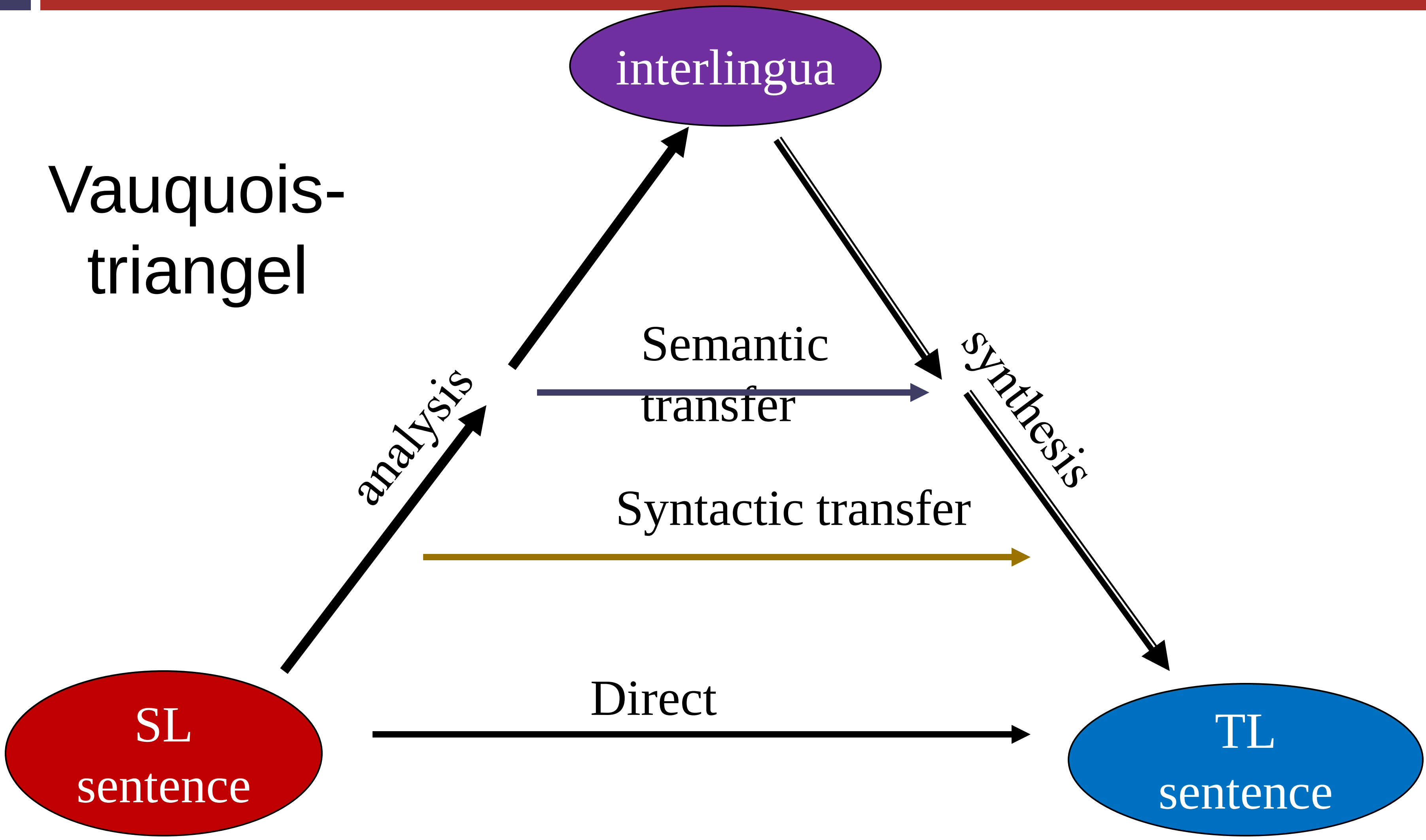
- Problem for interlingua:
  - A language independent meaning representation
  - Has to encode all distinctions in all languages
  - What should the lexical items be?
- Transfer approach:
  - Language specific representations
  - Contrast between pair of languages as transfer rules
- Syntactic transfer:
  - Extends the direct approach with a syntactic analysis
- Semantic transfer
  - Semantic representations, but language independent

Syntactic transfer



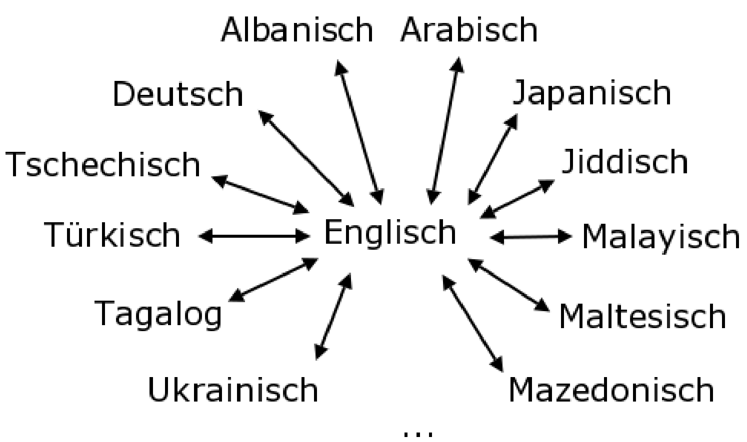
Alternative strategies

Vauquois-triangel



How different are the strategies?

- From direct to interlingua:
  - Choose one language as interlingua
  - (Google translate seems to do this:)



- From transfer to interlingua:
  - Choose the syntactic (or semantic) representations of one language as interlingua.
- But: In general,
  - two translation steps: L1 → L3 → L2
  - are inferior to one step L1 → L2
  - Why?

Machine Translation

1. Translation – by humans and machines
2. Traditional approaches to MT
    1. Direct
    2. Interlingua
    3. Transfer
3. Empirical approaches:
    1. Example-based MT (EBMT)
    2. Statistical MT - SMT
4. History
5. Why is (machine) translation hard?

Example-based MT

- No: Jenta har lest lekser i en time.
- Eng: ?
- Eksempler:
  - Jenta har spist et eple hver dag
  - The girl has eaten an apple a day
  - Per hadde lest lekser
  - Per had studied
  - Kari sang i en time.
  - Kari sang for an hour.
- Find the longest overlapping sequences
- Not necessarily constituents

SMT main principles

dekket	
the tire	0.314
the deck	0.118
covered	0.072
the cover	0.066
hid	0.045
set	0.029

- Bilingual
- Two parts:
  - Translation model
  - Language model
- Translation model:
  - Large amounts of text translated from SL to TL
  - Try to determine which word (phrase) in TL which translates which word in SL
  - Construct a translation dictionary with probabilities

SMT main principles 2

- Language model:
  - Huge amounts of text in TL
  - Count n-gram frequencies
- Translation
  - Given an input string
  - Construct (in principle) all possible strings of words in TL
  - Assign a probability by combining probabilities from translation model and language model
  - Choose the most probable result

SMT simplified example

En	kokk	lagde	en	rett	med	bygg	.
a 0.9	chef 0.6	made 0.3	a 0.9	right 0.19	with 0.4	building 0.45	
…	cook 0.3	created 0.25	…	straight 0.17	by 0.3	construction 0.33	
	…	prepared 0.15		court 0.12	of 0.2	barley 0.11	
		constructed 0.12		dish 0.11	…	…	
		cooked 0.05		course 0.07			
		…		…			

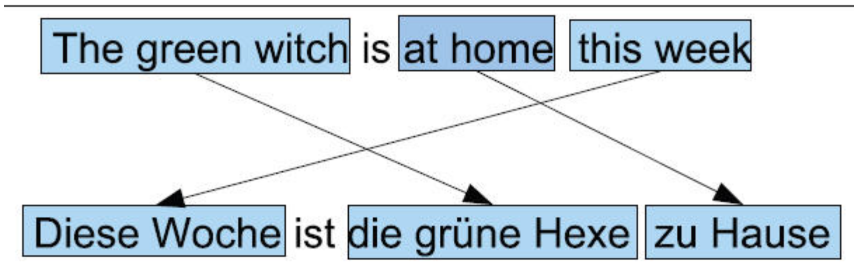
Similarly for:

- pos 0-2 (2x3)
- pos 1-3
- pos 2-4
- pos 3-5 (4x5)
- pos 6-8

Pos4 – pos 6 (1x3x3 many)		Pos5 – pos 7 (5x3x3 many)	
a right with	2.7×10^{-12}	right with building	1.7×10^{-18}
a right of	1.5×10^{-10}	right with construction	5.4×10^{-18}
a right by	9.7×10^{-12}	right with barley	8.7×10^{-19}
…		…	
a course of	1.5×10^{-14}	course of barley	1.5×10^{-16}

Refinements

- Word order
- LM with more than 3 words (4, 5,…)
- phrases:
  - dommeren – the judge
  - en dommer – a judge
  - god dag – nice day



Examples

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→

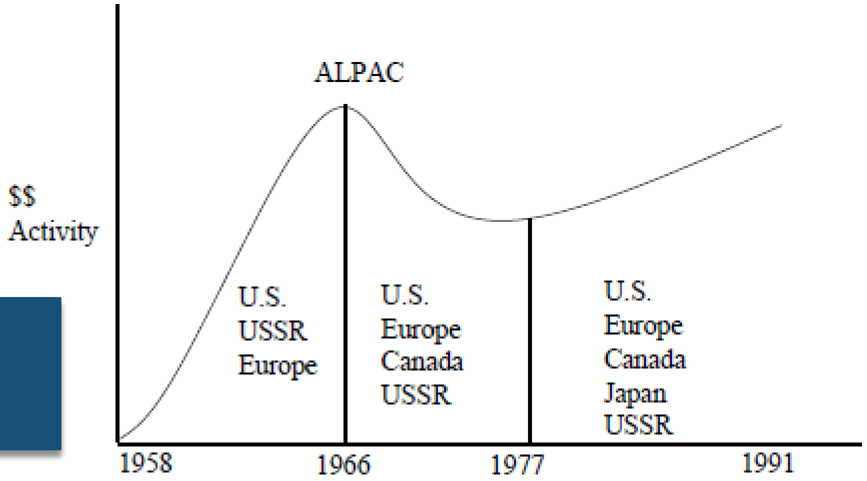
- On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with 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

1. Translation – by humans and machines
2. Traditional approaches to MT
   1. Direct
   2. Interlingua
   3. Transfer
3. Empirical approaches:
   1. Example-based MT (EBMT)
   2. Statistical MT - SMT
4. History
5. Why is (machine) translation hard?

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 for PCs
- 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

1. Translation – by humans and machines
2. Traditional approaches to MT
   1. Direct
   2. Interlingua
   3. Transfer
3. Empirical approaches:
   1. Example-based MT (EBMT)
   2. Statistical MT - SMT
4. History
5. Why is (machine) translation hard?

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!

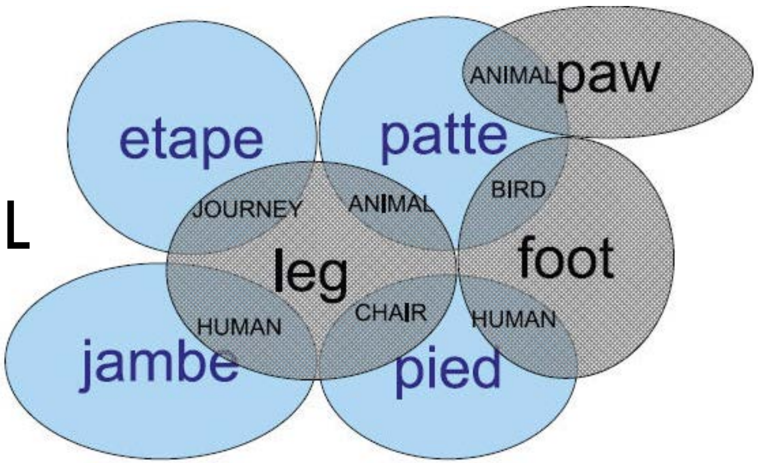
Lexical ambiguities in SL

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)

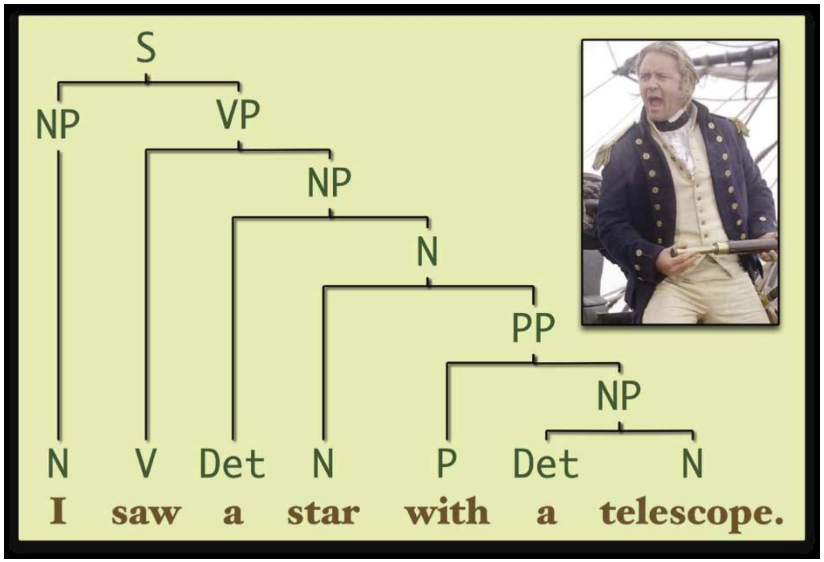
Lexical choice in transfer

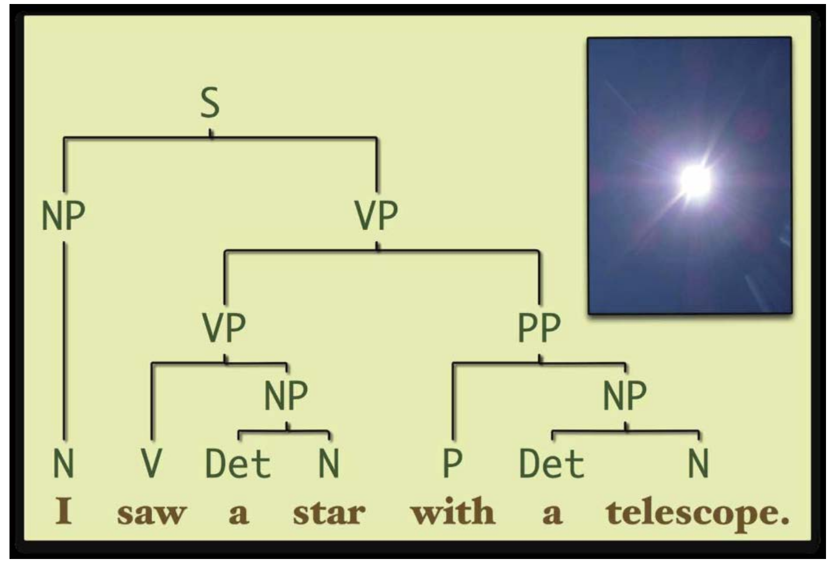
- The TL may make more distinctions than SL
  - No: *tak*, Eng: *ceiling/roof*
  - Eng: *grandmother*,
    No: *farmor/mormor*
- Context dependent choice in TL
  - *Strong tea, powerful government*
  - *Dekke på bordet* → *set the table*
  - *Dekke bordet* → *set/cover the table*
- Languages may draw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