

## Elements of Language Processing and Learning

Dr. Khalil Sima'an

Statistical Language Processing and Learning  
Institute for Logic, Language and Computation

Universiteit van Amsterdam



"My computer doesn't understand me!"

Natural language: a dense packaging documenting human state:

- Print (books, newspapers,...), Recorded speech (audio, songs, ...), WWW (webpages, Wikipedia,...)
- E.g., Ancient writings (e.g., Greece, Egypt, India, Persia); Political documents (law, parliament proceedings); History; Science; Safety guidelines; Cooperation in commerce and business; etc.

Bottleneck/challenge for economy, science, human exchange:

- Too many texts, too little time (Retrieval, Summarization, ...).
- Resources only in certain languages (Translation: see WWW in English).
- Scientific/commercial/cultural exchange (Translation).

# Natural Language Processing and Learning

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

## Two kinds of overlapping motivations

- **Theoretical:** A unique cognitive capacity.  
Language understanding; human communication;  
Finger-print of “Thinking”.
- **Engineering:** Need computer programs that conduct tasks:
  - Help-tools: Retrieval, Extraction, Q&A etc.
  - ⋮
  - Human-like: Translation, Summarization, Generation,  
Dialogue, etc.

This course: Human-like Text Processing and Learning

# Artificial Intelligence: Natural Language Systems

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Human language in input and output:

- Systems that extract information from text/speech.  
Examples: inf retrieval, info extraction, text mining etc.
- Systems that transform text/speech to text/speech  
Examples: translation systems, summarization, dictation, etc.
- Systems that communicate with people through language.  
Examples: question-answering, dialogue systems.

This course: Statistical Parsing and Translation

# Translation between Languages

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Human translators translate texts from a source language to a target language.

I don't smoke.

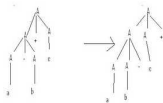
Je ne fume pas



Can we build a computer program that translates texts from one language to another?



What challenges will we face and how do we tackle them?



# History I: Premature Optimism and Failure

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Some history on translation and speech recognition:

- During 50's and 60's

First computers; Chomsky's grammars; programming languages; big optimism and huge funding

Translation is Easy: we can program this!!

ALPAC (Automatic Language Processing Advisory Committee) Report 1966 (U.S. Government).

Failure: AI abandons NLP, NLP abandons Translation

- During 70's and 80's:

AI: "You need world-knowledge: build an ontology"

CS: Concentrate on Information Retrieval

Linguistics: We need better theory

# History II: Renewed Optimism

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

- During 70's: A group of statisticians at IBM TJ Watson “digs up” an old idea ([Weaver 1948, 1949]):  
*When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.*

Communication and Information Theory (Shannon, Weaver);  
Code breaking (Turing).

- During 80's: Success in ASR; Look at Translation
- During 90's: Success in parsing and Translation
- By 2006: Google introduces “Google Translate”!

Next: Example Phenomena from Statistical MT

# Compound noun in Dutch vs. English

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Google Translate

http://translate.google.com/translate\_t



Text and Web

[Translated Search](#)

[Dictionary](#)

[Tools](#)

[Help](#)

## Translate Text

Original text:

De regering heeft aangekondigd dat de hypotheekrenteaftrek zal worden afgeschaft. Hypotheek Rente Aftrek.

Het IBM concern zal binnenkort veel winst gaan maken, heeft de Telegraaf gemeld.

Translation: Dutch » English

The government has announced that the hypotheekrenteaftrek will be abolished. Mortgage Interest Deduction.

The IBM group will soon be making much profit, the Telegraaf reported.

Dutch » English **Translate**

[Suggest a better translation](#)

## Translate a Web Page

http://

Spanish » English **Translate**

# Fixed expressions and other stuff: Dutch to English

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Google vertalen

Van: Nederlands ▼

Naar: Engels ▼

Vertaal

Als wij in details treden dan is het einde zoek.

Als wij in details treden dan komt er geen einde aan.

---

Als wij in details treden dan zult u dat niet waarderen.

Als wij in details treden dan zult u dat niet apprecieeren.

**Vertaling van het Nederlands in het Engels**

If we go into detail then no end.

If we go into details then there is no end.

---

If we go into details you will not like it.

If we go into details you will not appreciate.

# More fixed expressions or negation: Dutch to English

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Google vertalen

Van: Nederlands ▼

Naar: Engels ▼

Vertaal

Hij eet zijn broodje op en vertrekt naar school.

Meneer van den Bosch neemt waar voor meneer den Uil.

Ik vind dit geen stijl.

**Vertaling van het Nederlands in het Engels**

He eats his sandwich and leaves for school.

Mr. van den Bosch takes the value for Mr. Owl.

I think this style.

# Relative clause in Dutch vs. English



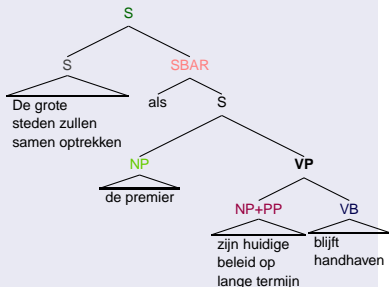
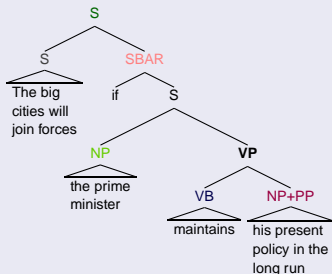
Van:  Naar:

De grote steden zullen samen optrekken als de premier zijn huidige beleid op lange termijn blijft handhaven.

## Vertaling van het Nederlands in het Engels

The big cities will join forces as the prime minister his present policy on maintaining long-term stays.

## A role for syntax in translation



# Morphology Arabic vs. English

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

Google vertalen

Van:  Naar:

-  
The wanted man  
The wanted men  
The men wanted  
---  
The most wanted man  
The most wanted men  
---  
The man wanted a car  
The men wanted a car

Vertaling van het Engels in het Arabisch

الرجل المطلوب  
وأراد رجال  
وأراد رجال  
---  
الرجل المطلوبين  
الرجال المطلوبين  
---  
أراد رجل سيارة  
أراد الرجل سيارة

Fonetisch lezen

# Some challenges in Translation between Languages

To translate between languages effectively we need to model differences between them, e.g.,

- Word senses (e.g., Dutch "bank" is English "couch" or "bank").
- Morphological inflection (e.g., Slavic/Semitic vs. English).
- Word-order/syntax (e.g., Slavic/Semitic/Chinese/Dutch/German vs English).
- Pronouns that disappear (e.g., Spanish/Arabic/... vs English).
- Language-specific combinations of words, fixed expressions.
- Idioms: culture-bound, non-compositional.
- ⋮
- **Ambiguity!**

# So many languages, so little time

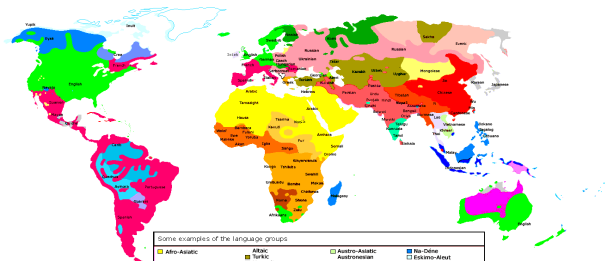
Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course



Some examples of the language groups

■ Afro-Asiatic	■ Altaic	■ Austro-Asiatic	■ No-Dine
■ Niger-Congo	■ Turkic	■ Austronesian	■ Eskimo-Aleut
■ Hamitic	■ Mongolic	■ Borneo-Philippine/Formosan	■ American Indian
■ Nilo-Saharan	■ East Siberian languages	■ Nuclear Malayo-Polynesian	■ Algonquian
■ Khoisan	■ Uralic	■ Papuan	■ Uto-Aztecian
■ Indo-European	■ Dravidian	■ Formo-Nyungan	■ Andean
■ Germanic	■ Sino-Tibetan	■ Tai-Kadaï	■ Naban
■ Albanic	■ Chinese	■ Isolabe	■ Brazilian indigenous
■ Romance	■ Burmese-Tibetan		
■ Slavic			
■ Indo-Iranian			
■ Baltic			
■ Caucasian			

- Are the differences between languages arbitrary?
- Are there shared regularities between different languages?

How should we automatically translate?

# Modeling Human Translation Expertise

- Translators are Experts in Translation
- Humans: Study, work, acquire by experience ...

## Can we model “expertise acquisition” from experience?

- Observe and learn how humans translate?
- Use input-output translation examples: Parallel corpora  
No access to what happens in between
- How do we build and select the correct translation?  
**Ambiguity** is stalking us all the way.

How can we learn translation regularities from data?

# Data and Statistical Models

Parallel corpus = a collection of text-chunks and their translations.  
Parallel corpora are the by-product of *human translation*.  
Every source chunk is paired with a target chunk.

Dutch	English
De prijs van het huis is gestegen.	The price of the house has risen.
Het huis kan worden verkocht.	The house can be sold.
Als het de marktprijs daalt zullen sommige gezinnen een zware tijd doormaken.	If the market price goes down, some families will go through difficult times.
.	.
.	.
.	.
.	.
.	.
.	.

- Hansards Canadian Parliament Proc. (English-French).
- European Parliament Proc. (23 languages).
- United Nations documents.
- Newspapers: Chinese-English; Arabic-English; Urdu-English.

# The hidden structure of translation

## How to model the translation mapping in the data?

The big cities will join forces if the prime minister maintains his present policy in the long run.

????

De grote steden zullen samen optrekken als de premier zijn huidige beleid op lange termijn blijft handhaven.

## What is the nature of the mapping?

$$\text{“} Translate(sentence) = \hat{\sum}_i Translate(part_i) \text{”}??$$

- What are  $part_i$  and  $Translate(part_i)$  in the data?
- What is  $\hat{\sum}_i$ ?
- How to model differences in word-order, morphology etc?
- What about ambiguity, idioms etc?

# Probabilistic Modeling: Simple Noisy Channel

Source sentence  $\mathbf{s} = s_1, \dots, s_n$

Target sentence  $\mathbf{t} = t_1, \dots, t_n$

$$\arg \max_{\mathbf{t}} P(\mathbf{t} | \mathbf{s}) = \arg \max_{\mathbf{t}} P(\mathbf{t}) \times P(\mathbf{s} | \mathbf{t})$$

- **Target Language Model  $P(\mathbf{t})=?$**

How regular is a given string  $\mathbf{t}$  in the target language?

$$P(\mathbf{t}) = \sum_{\mathbf{d}} P(\mathbf{t}, \mathbf{d})$$

Derivation  $\mathbf{d}$ : Formal device (Finite-State, Context-Free, etc).

- **Translation Model  $P(\mathbf{s} | \mathbf{t})=?$**

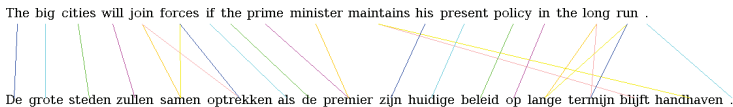
How to model the mapping  $\mathbf{t} \rightarrow \mathbf{s}$ ?

How to learn good language models from data?

How to learn good translation models from data?

# The hidden structure of translation

## Simple approach 1: Words as basic units



Let  $\mathbf{a} \subseteq \{ \langle i, j \rangle \mid i \in [0..|\mathbf{s}|] \text{ and } j \in [0..|\mathbf{t}|] \}$

$$P(\mathbf{s} \mid \mathbf{t}) = \sum_{\mathbf{a}} P(\mathbf{s}, \mathbf{a} \mid \mathbf{t}) \approx \sum_{\mathbf{a}} P_o(\mathbf{a} \mid \mathbf{t}) \prod_i P_l(s_i \mid t_{a_i})$$

Where  $\mathbf{a} = a_1, \dots, a_{|\mathbf{s}|}$

Where does the word-alignment come from?

What about word-reordering  $P_o$ ? What about the lexicon  $P_l$ ?

Unsupervised learning on parallel corpus

# The hidden structure of translation

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

## Word-order differences: Impoverished Models

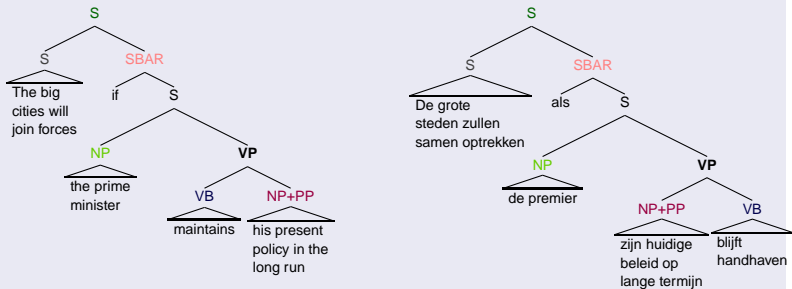
$$P(\mathbf{s} \mid \mathbf{t}) = \sum_{\mathbf{a}} P(\mathbf{s}, \mathbf{a} \mid \mathbf{t}) \approx \sum_{\mathbf{a}} P_o(\mathbf{a} \mid \mathbf{t}) \prod_i P_l(s_i \mid t_{a_i})$$

Reordering:  $P_o(\mathbf{a} \mid \mathbf{t})$  may be defined based on, e.g.,

- relative positions of  $a_i$  to  $a_{i+1}$  (Markovian)
- include lexical content from target
- syntactic information of current position in  $\mathbf{t}$ .

# The hidden structure of translation

## Synchronous Tree-based Rewrite Productions as basic units?



Where do the trees and node alignments come from?

Which trees to use (syntax, other)?

Probabilistic model? Training?

Our courses: How do we learn the structures and use them?

# Modeling Parallel Corpus Data

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

- How to represent the source sentence?
- How to represent the target sentence?
- How to model the mapping between these representations?  
We need to model sentence pairs!!
- Is translation compositional?
- Some options: Probabilistic Synchronous Grammars, Probabilistic Tree Transducers, etc.
- What learning algorithms?
- How to automatically evaluate translation output?

Statistical Machine Translation is Happening Now

# Course structure

Elements of  
Language  
Processing and  
Learning

Dr Khalil  
Sima'an

Motivation

Challenges

This course

- Statistical parsing: Treebanks; Grammars; Parsing algorithms; Estimation
- Intro to: Unsupervised learning of richer probabilistic grammars
- Statistical Translation: Word-alignment; Phrase-based models;
- Intro to: Hierarchical and syntax-driven translation models.