

NATURAL LANGUAGE PROCESSING

Gayane Petrosyan
McGill University

Luys DAP 2012

Alan Turing

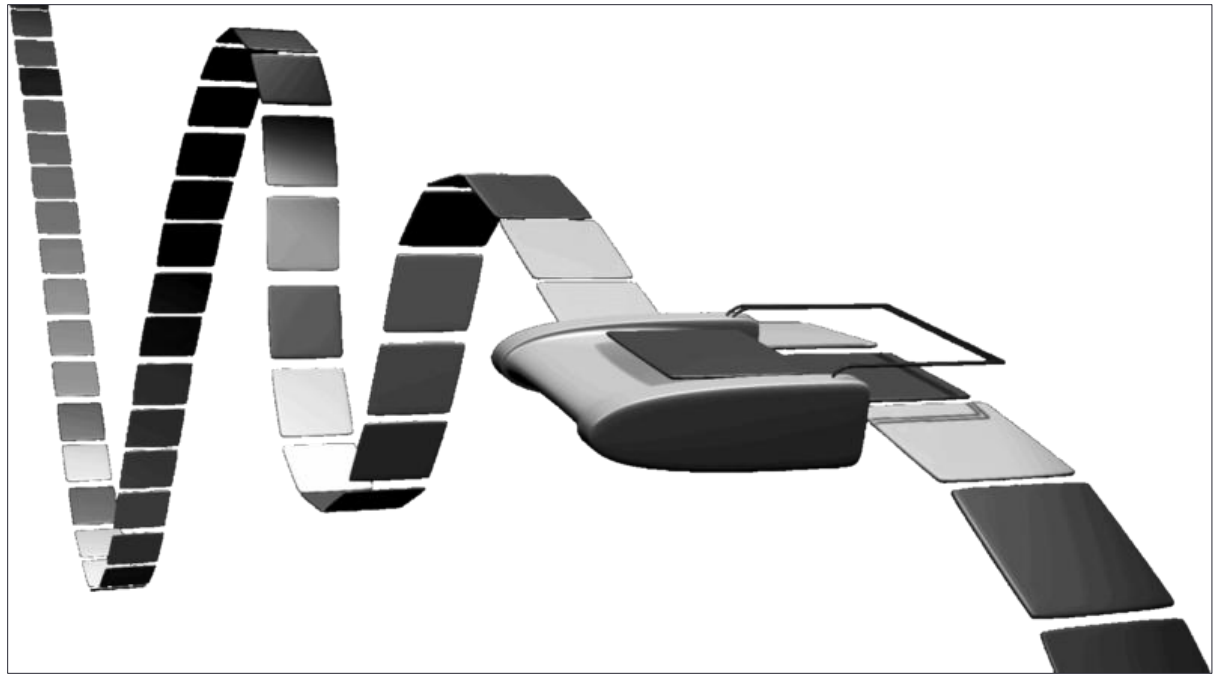
- Enigma Machine



Bundesarchiv, Bild 103-2007-0106-402
Foto: Wabner, 10. Dezember 1943

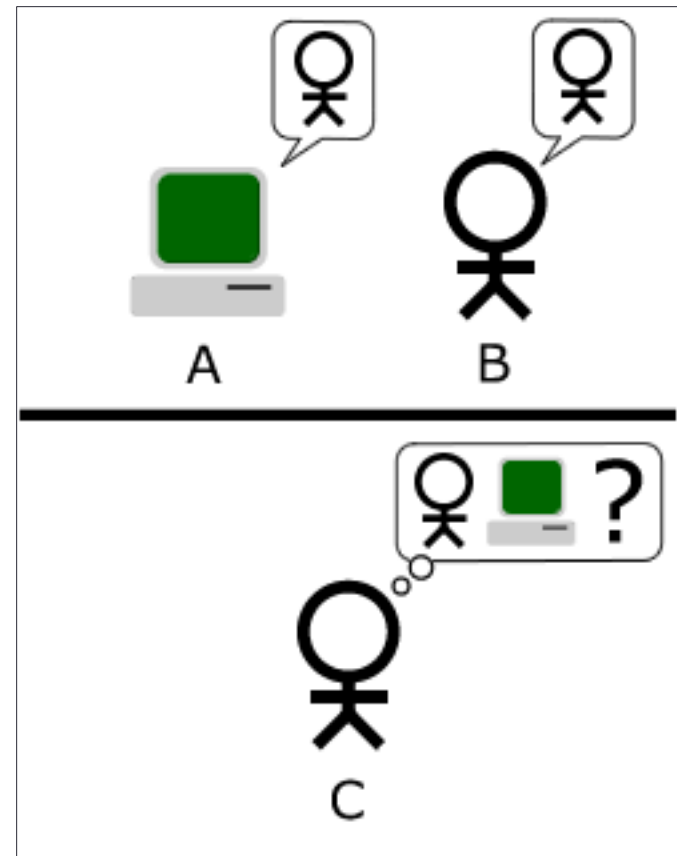
Alan Turing

- Turing Machine



Alan Turing

- "[Computing Machinery and Intelligence](#)," in 1950
- 'Can machines think?'



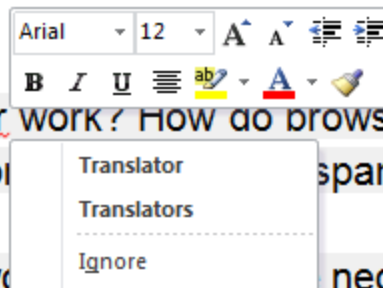
- [Loebner Prize](#) since 1991.

Natural Language Processing

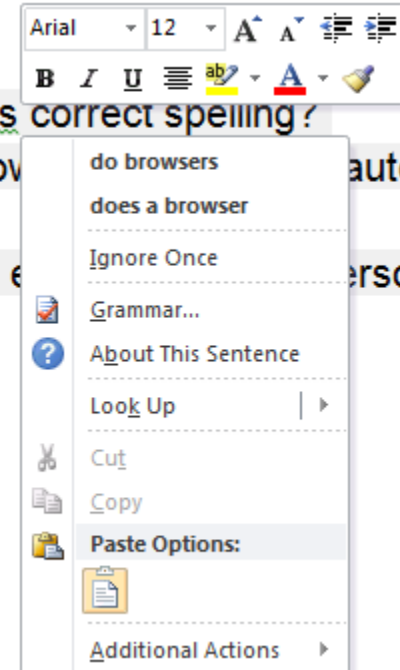
- APPLICATIONS: Question Answering, Information Extraction, Sentiment Analysis, Machine translation,
- MOSTLY SOLVED: Spam classification, POS tagging, NER
- GOOD PROGRESS: Sentiment analysis, Information extraction, machine translation,
- HARD: Question Answering, paraphrase, summarization, Dialog
- WHY IS NLP DIFFICULT?: Ambiguities, idioms, Non-English word, world knowledge.

Spelling correction

How does Google Tanslator work? How do browsers correct spe
How do email services auto spam? How do onlin
classify product reviews?
How does Google search w negative effects of



How does Google Translator work? How does browsers correct spelling?
How do email services automatically classify spam? How
classify product reviews?
How does Google search work? What are the negative e on individuals and democracy?



Machine Translation

Translate From: English To: Armenian Translate

Russian French English English Russian Armenian Alpha

interest

հետաքրքրություն

New! Click the words above to edit and view alternate translations. Dismiss

noun
հետաքրքրություն interest, gusto, hobby
շահ benefit, interest, gain, profit, behalf, a

Translate From: English To: Armenian Translate

Russian French English English Russian Armenian Alpha

interest rate

տոկոսադրույք

New! Click the words above to edit and view alternate translations. Dismiss

noun
տոկոսադրույք interest rate

Sentiment Analysis



Nikon Zoom-Nikkor Zoom lens - 24 mm - 70 mm - F/2.8 - Nikon F

\$1,489 [online](#) ★★★★★ [1,205 reviews](#)

December 2007 - Nikon - Wide Angle - Zoom - Nikon F - f/2.8 - Autofocus - Ultrasonic Motor - Aspherical

▶ [See more details](#)

Reviews


Summary - Based on 1,205 reviews









What people are saying

picture/video	<div><div></div><div></div><div></div><div></div><div></div></div>	"The image quality is great."
value	<div><div></div><div></div><div></div><div></div><div></div></div>	"This is a great lens and well worth the price."
images	<div><div></div><div></div><div></div><div></div><div></div></div>	"Sharp, Fast, great images..."
construction	<div><div></div><div></div><div></div><div></div><div></div></div>	"Construction is robust."
size	<div><div></div><div></div><div></div><div></div><div></div></div>	"Very light weight, sharp, wonderful for events!"
bokeh	<div><div></div><div></div><div></div><div></div><div></div></div>	"Sharp, versatile and excellent bokeh."
sharpness	<div><div></div><div></div><div></div><div></div><div></div></div>	"Lens with excellent sharpness and magnificent."

Question Answering

 **WolframAlpha** computational...
knowledge engine

What is life expectancy in Armenia ?  

    [Examples](#) [Random](#)

Input interpretation:

Armenia life expectancy

Result:

72.7 years (world rank: 120th)

Demographics: [Show rates](#) [Show distribution](#) [Show non-metric](#)

population	3.09 million people (world rank: 137 th) (2010 estimate)
population density	110 people/km ² (people per square kilometer) (world rank: 100 th) (2010 estimate)
population growth	0.151 %/yr (world rank: 201 st) (2008 estimate)
life expectancy	72.7 years (world rank: 120 th) (2009 estimate)
median age	31.5 years (world rank: 81 st) (2009 estimate)

Units »

Spelling correction

- Non dictionary word -> error
- Find similar words from dictionary -> candidate set
- Dictionary word -> consider all words
- Find similar pronunciation, spelling words -> candidate set

The Noisy channel model of Spelling

- Operations
 - Insertion
 - Deletion
 - Substitution
- $D(n, m)$ - distance between n length string and m length string is the solution
- Bottom-up
 - Compute $D(i, j)$ for small i, j
 - Compute $D(l, j)$ from previously computed values

Real systems

- If very confident in correction
 - Autocorrect, hte->the
- Less confident
 - Give the best correction
- Even less confident
 - Give a correction list
- Unconfident
 - Mark as an error

Computing Probabilities

Coin
H
T

$$P(H) = \frac{\text{count}(H)}{\text{Count}(\text{all cases})} = \frac{1}{2}$$

L	R
H	H
H	T
T	H
T	T

$$P(HH) = \frac{\text{count}(HH)}{\text{Count}(\text{all cases})} = \frac{1}{4}$$

$$P(HH/H_L) = \frac{\text{count}(HH)}{\text{Count}(H_L)} = \frac{1}{2}$$

Language Models

- The goal is to assign probability to the sentence
- $P(\text{Water is so transparent}) = P(\text{transparent/ Water is so}) * P(\text{so/ water is}) * P(\text{is/ Water}) * P(\text{Water/ <Start>})$
- Bigram model
$$P(\text{Water is so transparent}) \approx P(\text{transparent/ so}) * P(\text{so/ is}) * P(\text{is/ Water}) * P(\text{Water/ <Start>})$$
- $P(\text{is / Water}) = \frac{\text{count}(\text{Water is})}{\text{count}(\text{Water})}$
- 3-gram, 4-gram, 5-gram and that's it.

Example

- Corpus
 - `<s> I am Sam </s>`
 - `<s> Sam I am </s>`
 - `<s> I do not like green ham </s>`
- $P(I / \text{<s>}) = ?$
- $P(\text{</s>} / \text{Sam}) = ?$
- $P(\text{Sam} / \text{<s>}) = ?$
- $P(\text{am} / I) = ?$
- $P(\text{Sam} / \text{am}) = ?$
- $P(\text{not} / \text{do}) = ?$

Next Lecture

- Date:
 - 2-Aug-2012 9:00am
- Topics:
 - Sentiment Analysis and Text Classification more in detail,
 - How does Google search work? What are the negative effects of search personalization on individuals and democracy?