NATURAL LANGUAGE PROCESSING

Gayane Petrosyan
McGill University

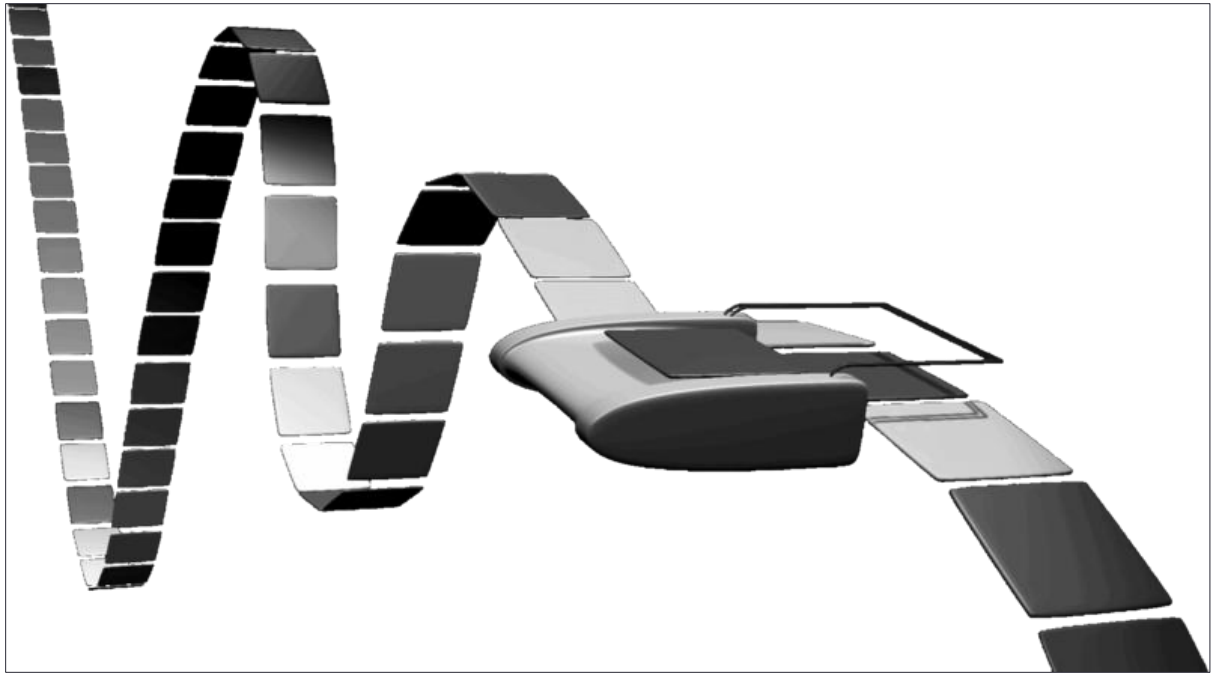
Luys DAP 2012

Alan Turing

- Enigma Machine

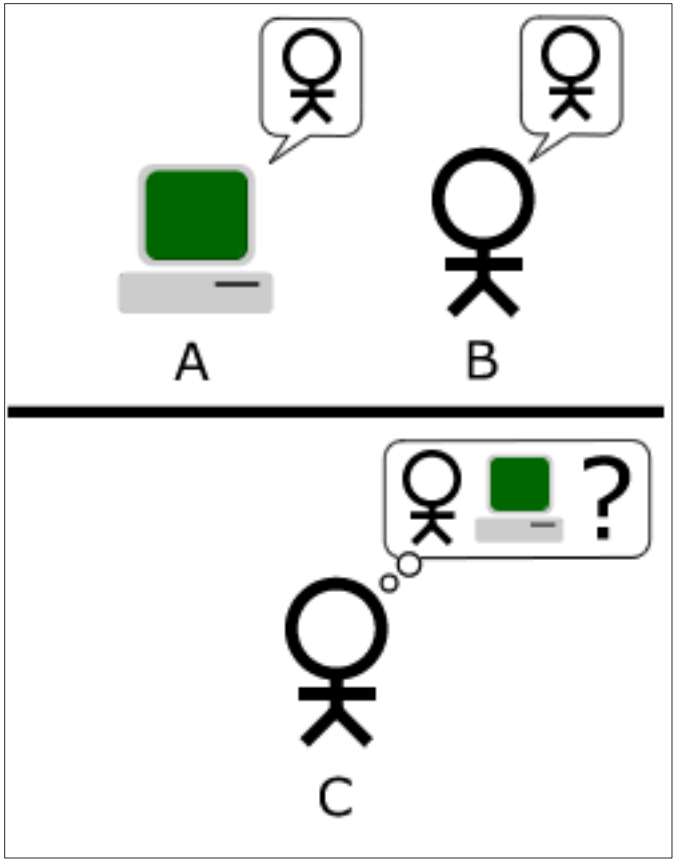
Alan Turing

- Turing Machine



Alan Turing

- "Computing Machinery and Intelligence," in 1950
- 'Can machines think?'



- *Loebner Prize since 1991.*

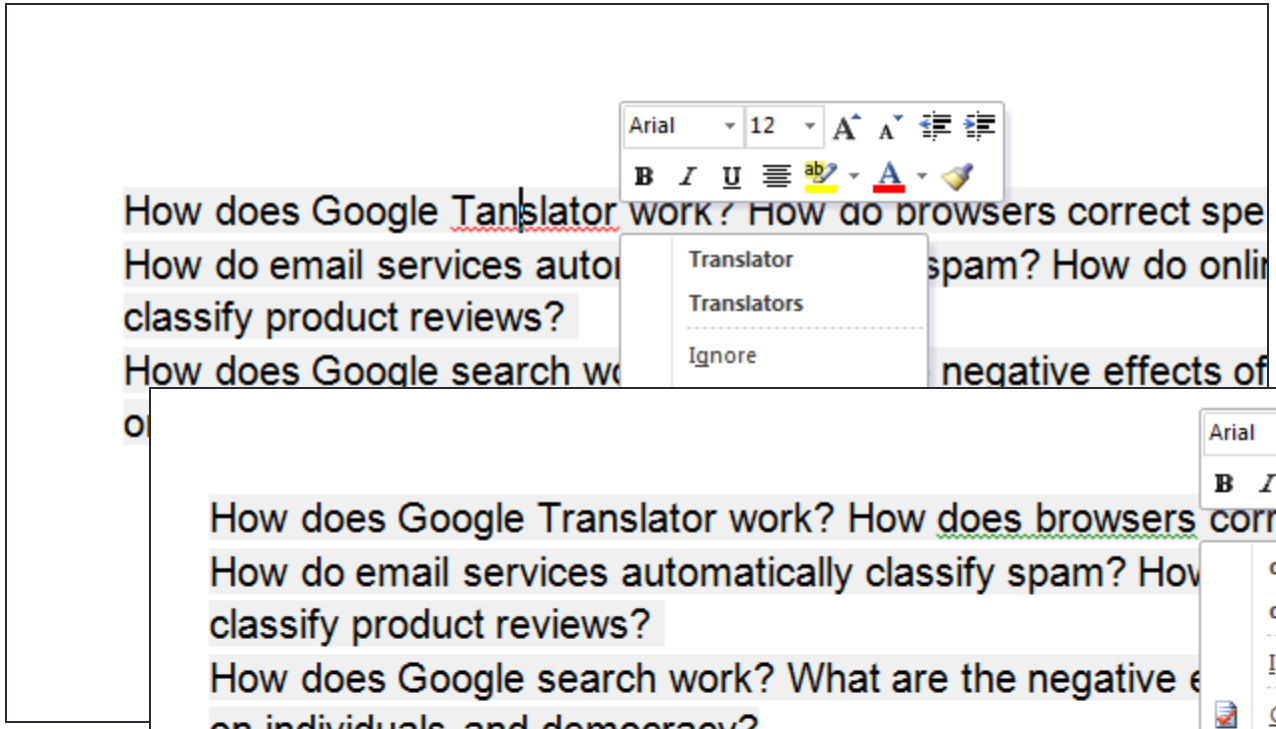
Natural Language Processing

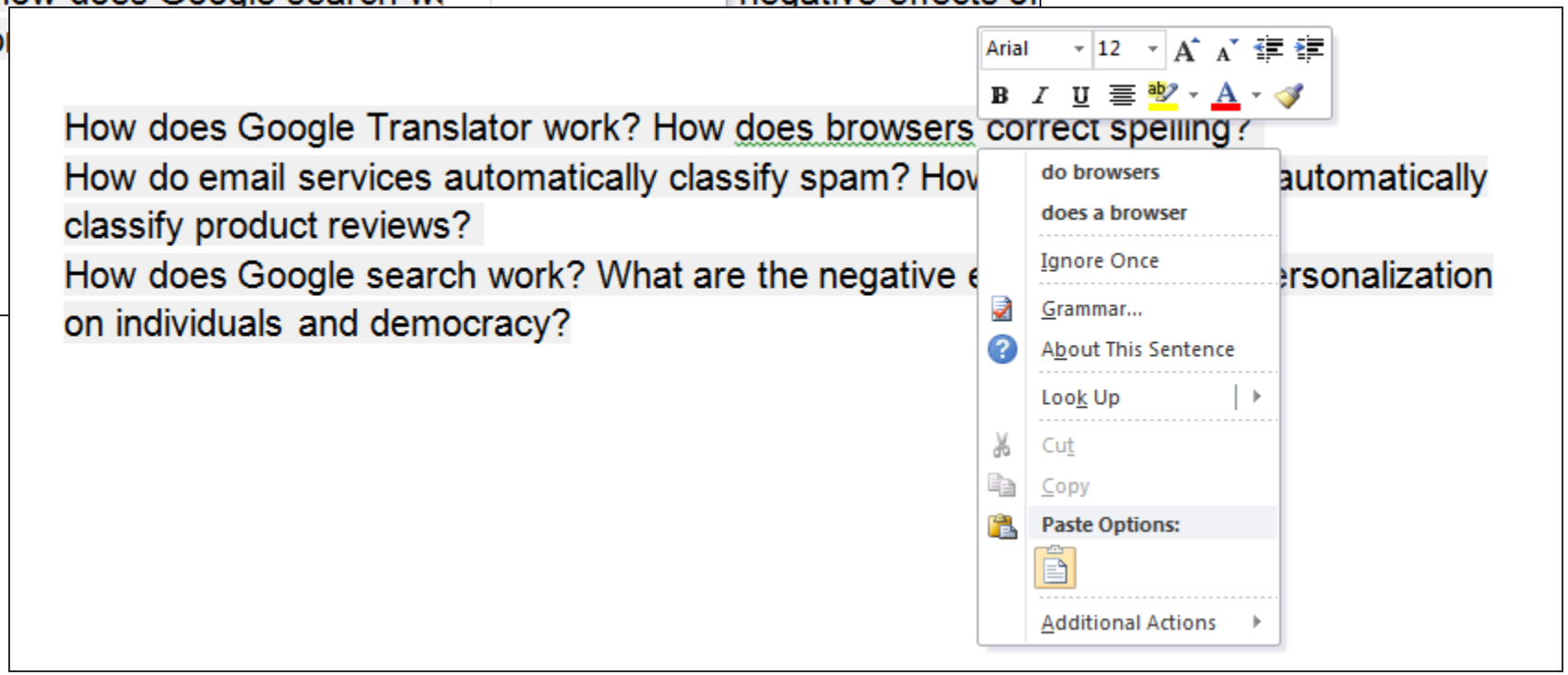
- APPLICATIONS: Question Answering, Information Extraction, Sentiment Analysis, Machine translation,

- MOSTLY SOLVED: Spam classification, POS tagging, NER
- GOOD PROGRESS: Sentiment analysis, Information extraction, machine translation,
- HARD: Question Answering, paraphrase, summarization, Dialog

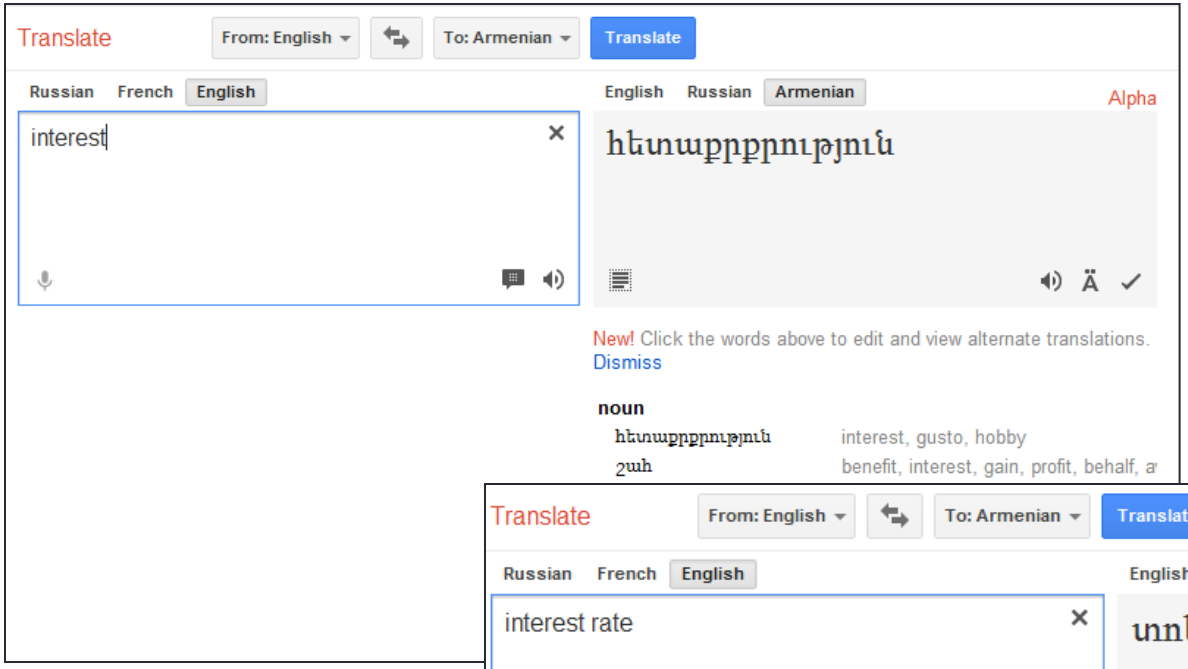
- WHY IS NLP DIFFICULT?: Ambiguities, idioms, Non-English word, world knowledge.

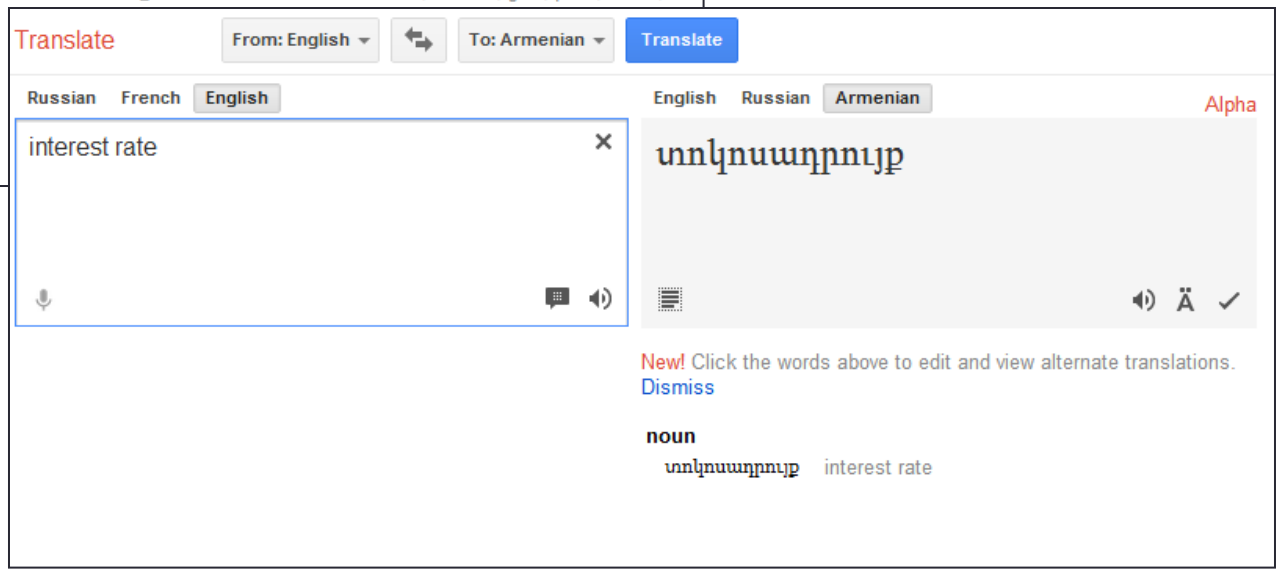
Spelling correction





Machine Translation





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

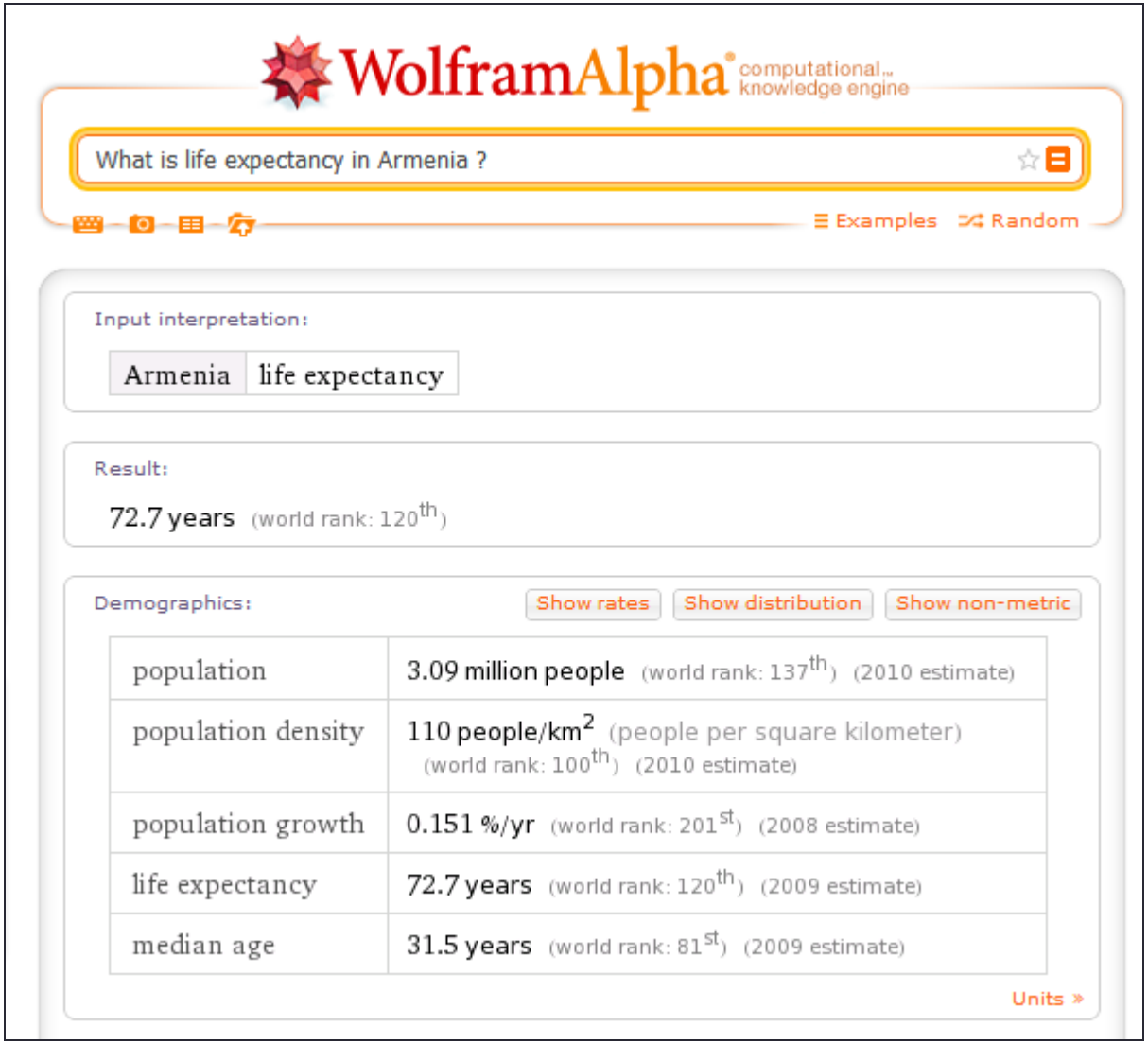
Summary - Based on 1,205 reviews

1 2 3 4 5 stars

What people are saying

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

Question Answering



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(I,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

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

L	R
H	H
H	T
T	H
T	T

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

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

Language Models

- The goal is to assign probability to the sentence

- P(Water is so transparent) = P(transparent/ Water is so) *
  P(so/ water is) * P(is/ Water) * P (Water/ <Start>)

- Bigram model
  P(Water is so transparent) ≈ P(transparent/ so) *
  P(so/ is) * P(is/ Water) * P (Water/ <Start>)

- $P(is\ /\ Water) = \frac{count(Water\ is)}{count(Water)}$

- 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 / <s>) = ?
- P(</s> / Sam) = ?
- P(Sam / <s>) = ?
- P(am / I) = ?
- P(Sam / am) = ?
- P(not / 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?