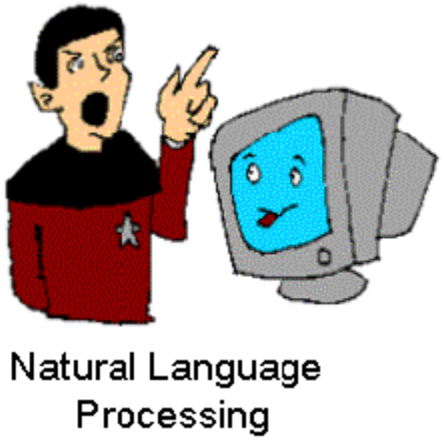
4TH INTERNATIONAL FORUM
BIG DATA DAY
- BAKU 2018 -

Natural Language Processing and its Application

Dr., Samir Rustamov,
Assistant Professor, School of IT & Engineering, ADA University

What is Natural Language Processing (NLP)?

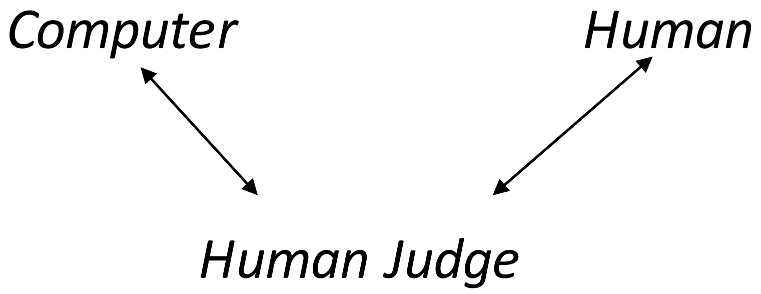
- Natural Language Processing (NLP) is a field of artificial intelligence that enables computers interact with human in natural language.



Ultimate goal: Natural human-to-computer communication

The Turing Test

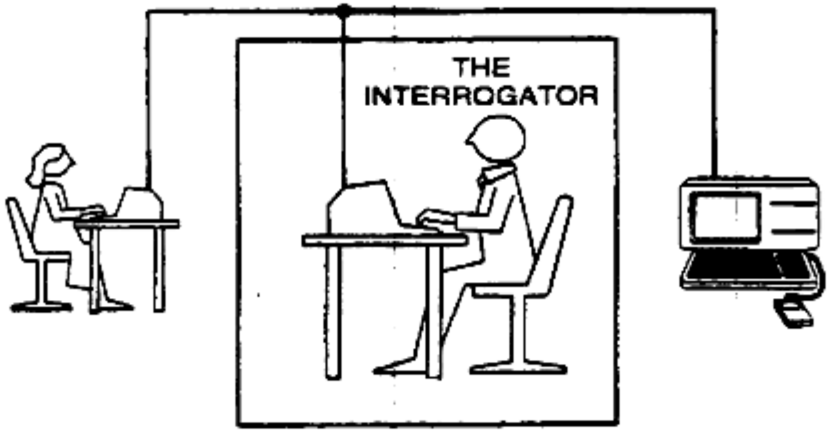
(Can Machine think? A. M. Turing, 1950)



Computer

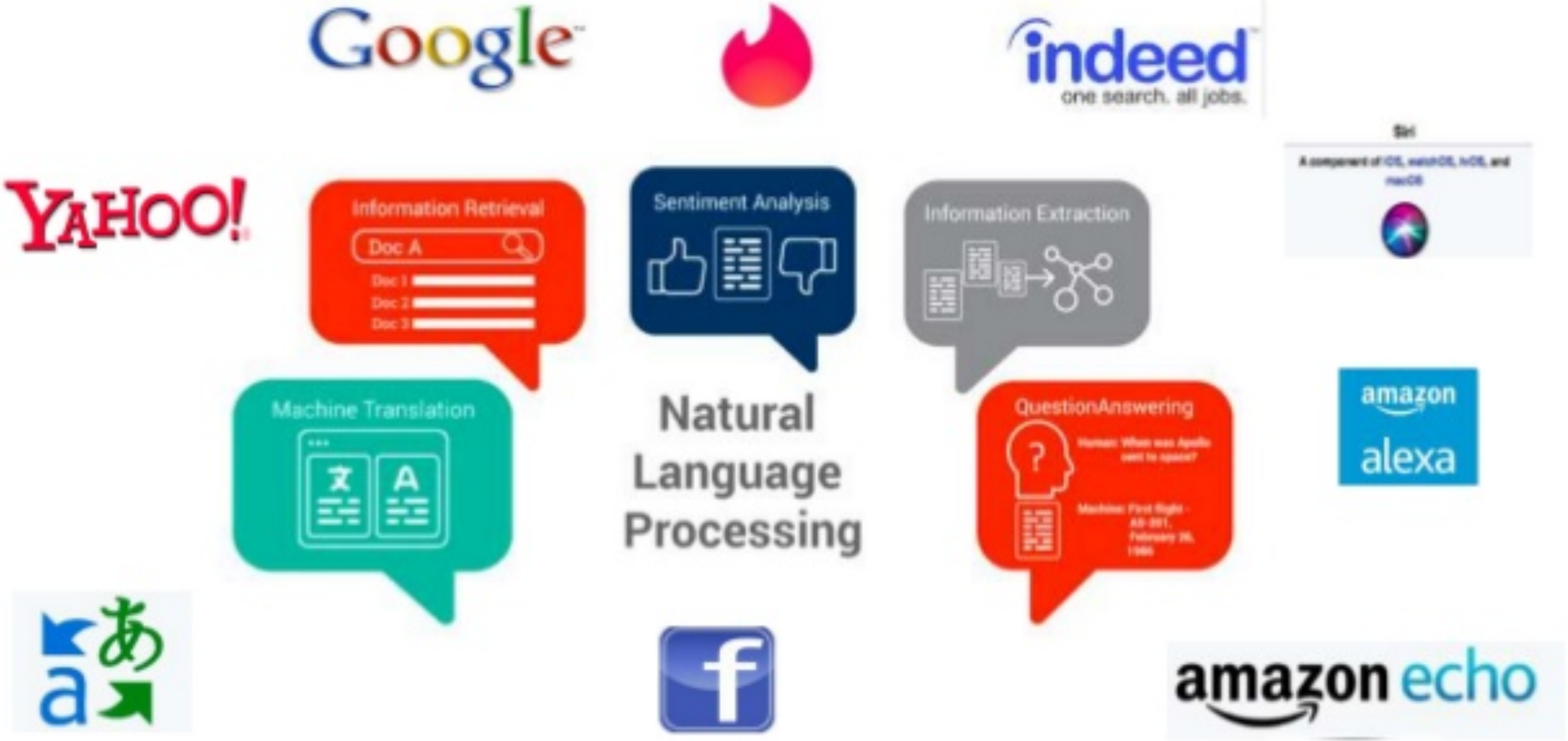
Human

Human Judge



- *Human Judge* asks tele-typed questions to *Computer* and *Human.*
- *Computer's* job is to act like a human.
- *Human's* job is to convince Judge that he is not machine.
- *Computer* is judged "intelligent" if it can fool the judge
- Judgment of intelligence is linked to appropriate answers to questions from the system.

NLP in the Commercial World



Information Retrieval

Doc A

Doc 1

Doc 2

Doc 3

Sentiment Analysis

Information Extraction

Machine Translation

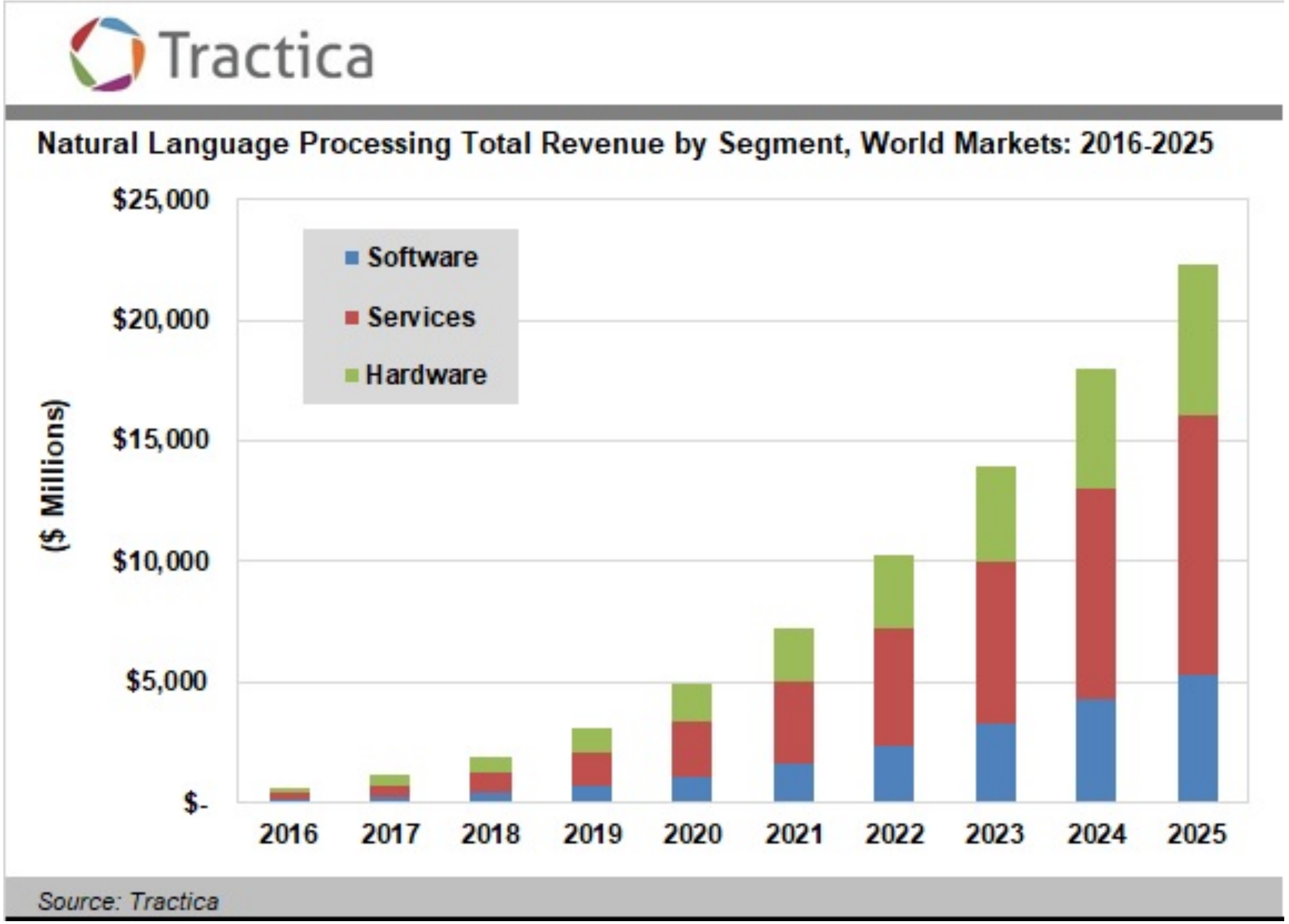
Natural Language Processing

QuestionAnswering

Human: When was Apollo sent to space?

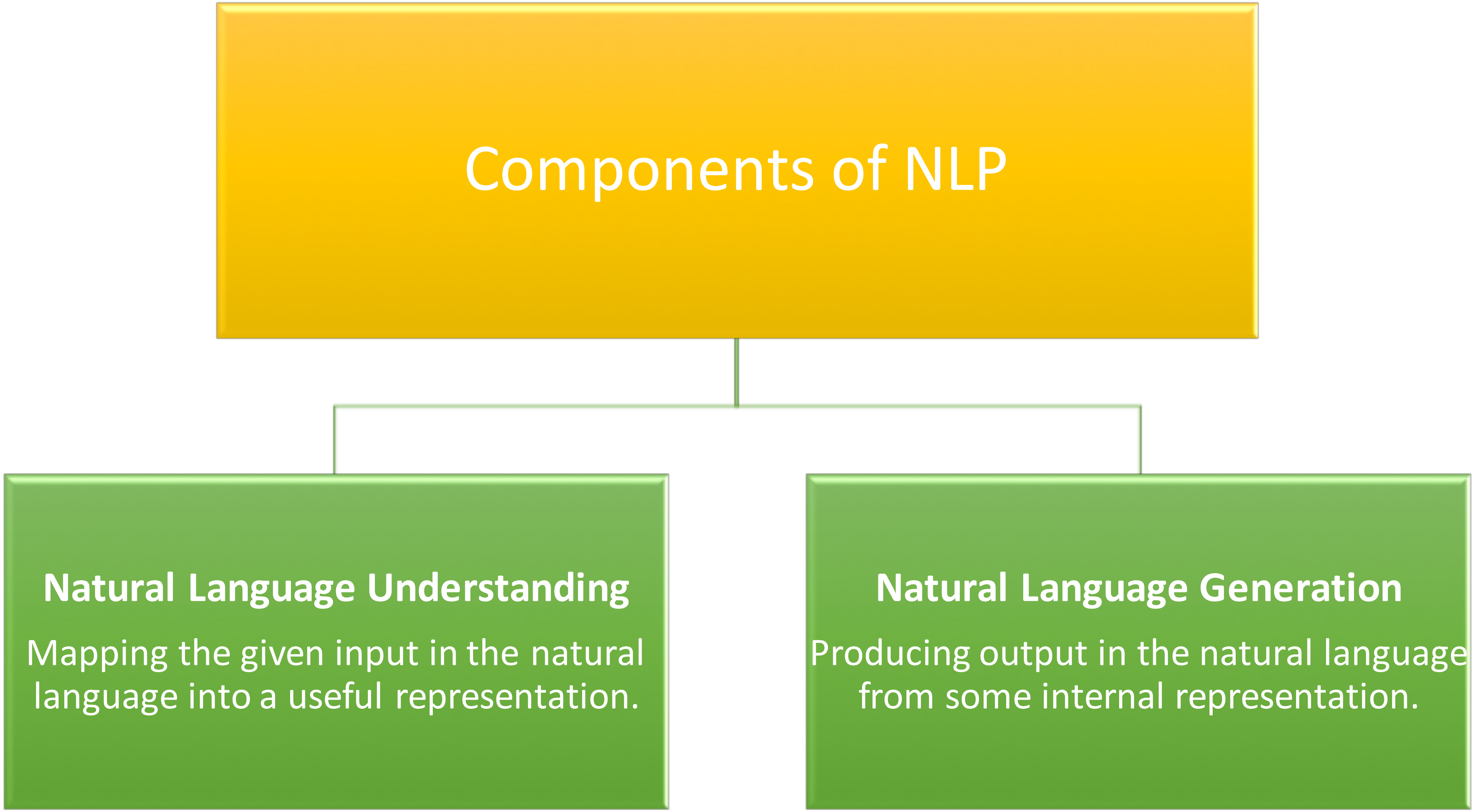
Machine: First flight - AS-201, February 26, 1966

Natural Language Processing Market to Reach $22.3 Billion by 2025



Forms of Natural Language

- The input/output of a NLP system can be:
  - **written text**
  - **speech**
- To process written text, we need:
  - **lexical, syntactic, semantic knowledge about the language**
  - **discourse information, real world knowledge**
- To process spoken language, we need everything required to process written text, plus the challenges of speech recognition and speech synthesis.



Components of NLP

Natural Language Understanding

Mapping the given input in the natural language into a useful representation.

Natural Language Generation

Producing output in the natural language from some internal representation.

NL Understanding is much harder than NL Generation. But, still both of them are hard.

Why NL Understanding is hard?

- Natural language is extremely rich in form and structure, and **very ambiguous***.*
- One input can mean many different things. Ambiguity can be at different levels.
  - Lexical (word level) ambiguity -- different meanings of words
  - Syntactic ambiguity -- different ways to parse the sentence
  - Interpreting partial information -- how to interpret pronouns
- Many input can mean the same thing.
- Interaction among components of the input is not clear.

Example of ambiguity

- Some interpretations of: **`Adamı gördüm.`**
  1. I saw the man.
  2. I saw Adam
  3. I saw my island.
  4. I visited my island.
  5. I saw my ADA
  6. I visited my ADA
  7. I bribed the man.
- Semantic Ambiguity:

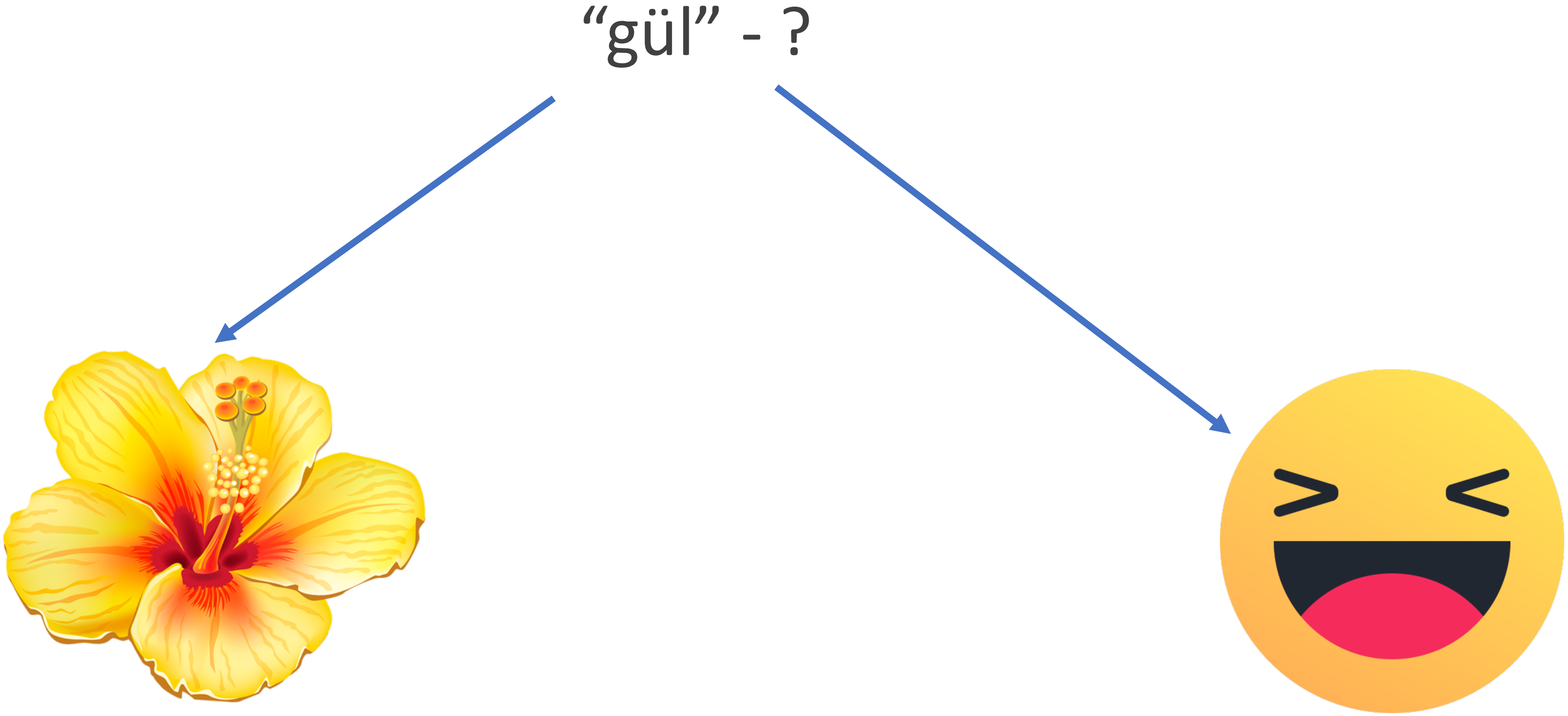
• gör	to see
• gör	to visit
• gör	to bribe

Resolve Ambiguities

- **lexical disambiguation** -- Resolution of part-of-speech and word-sense ambiguities are two important kinds of lexical disambiguation.
- **syntactic ambiguity** -- can be addressed by probabilistic parsing.

What is PoS tagging? Why is it important?

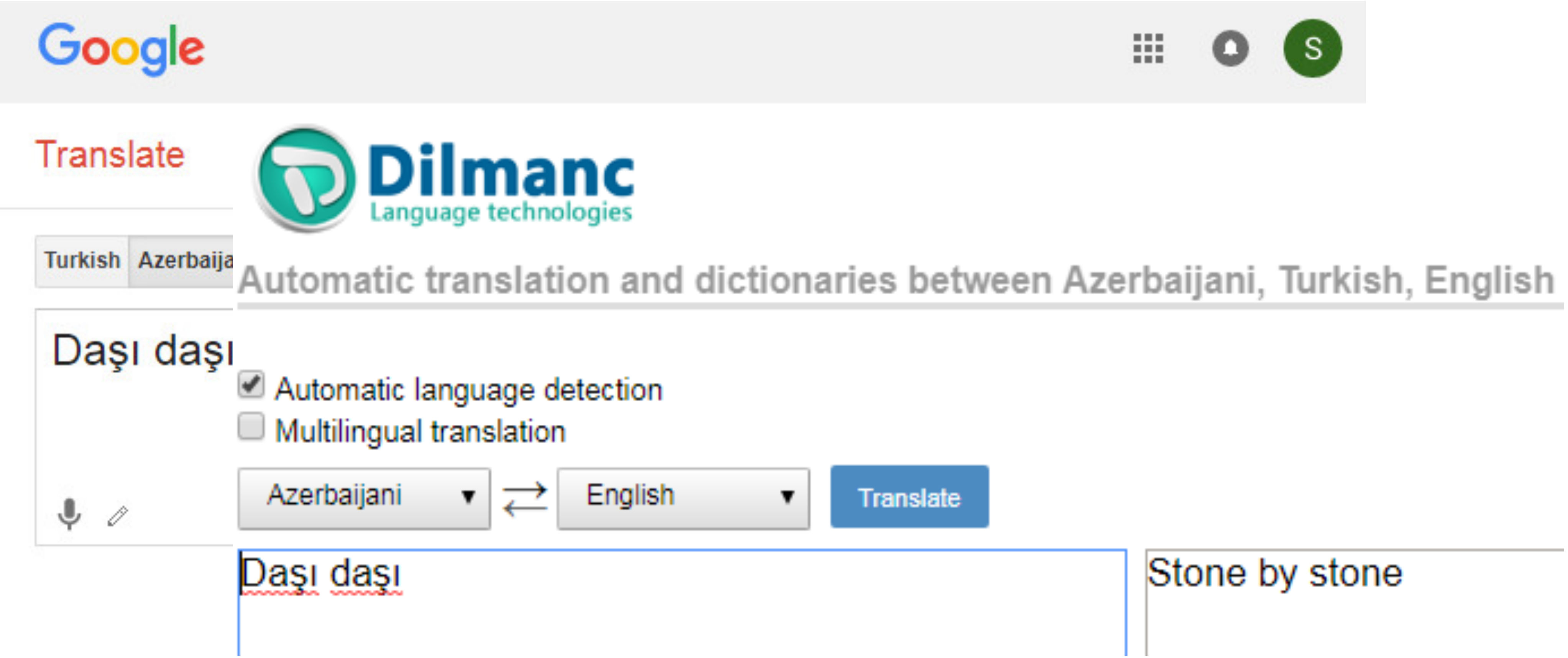
Each word has a part-of-speech tag to describe its category. POS Taggers try to find POS tags for the words.



“gül” - ?

Why it matters? Applications

- Machine translation – “Daşı daşı”



Daşı daşı

Automatic translation and dictionaries between Azerbaijani, Turkish, English

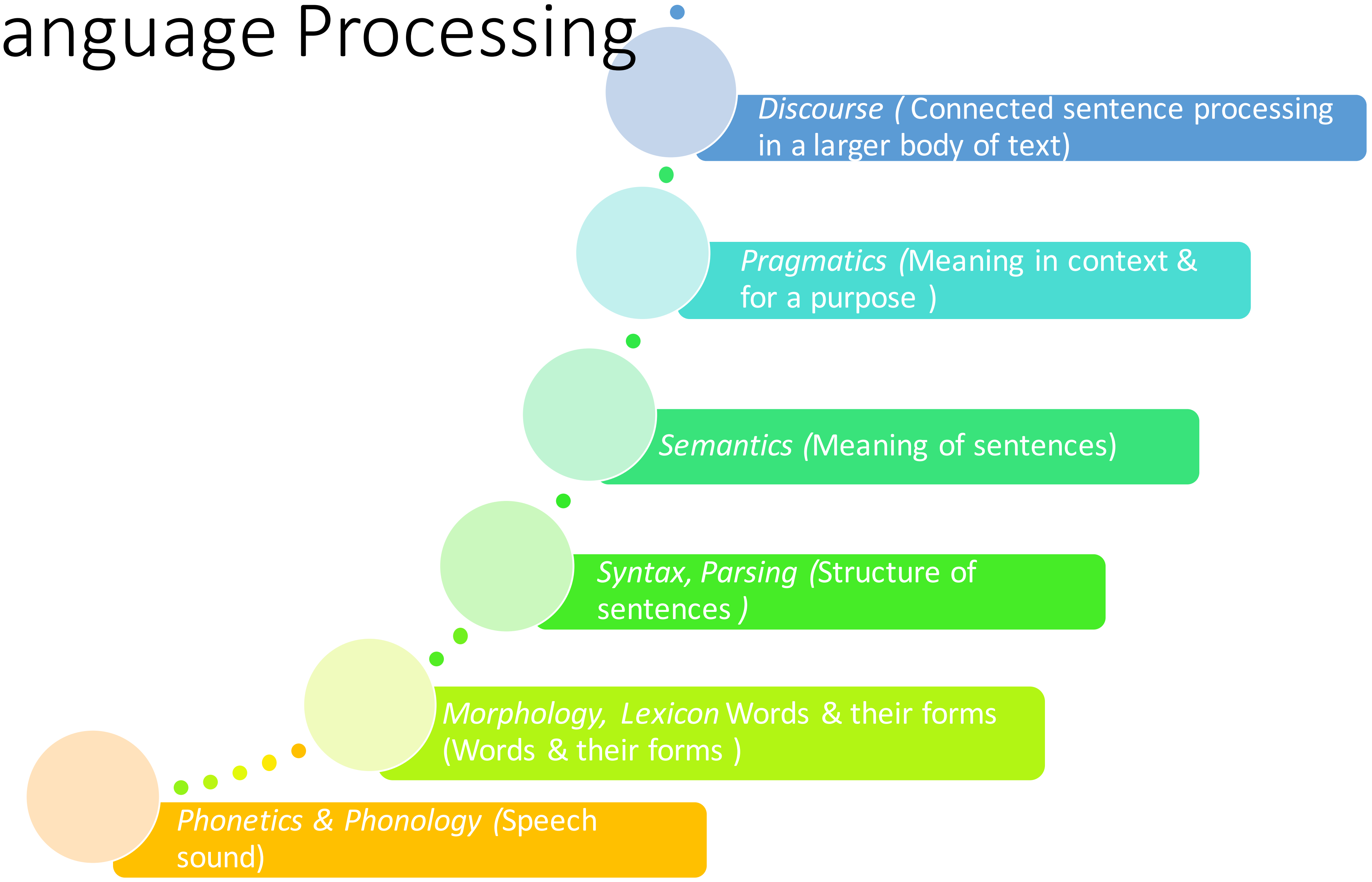
☑ Automatic language detection
□ Multilingual translation

Azerbaijani ▾ ⇄ English ▾

Daşı daşı

Stone by stone

Language Processing



Discourse (Connected sentence processing in a larger body of text)

*Pragmatics (*Meaning in context & for a purpose)

*Semantics (*Meaning of sentences)

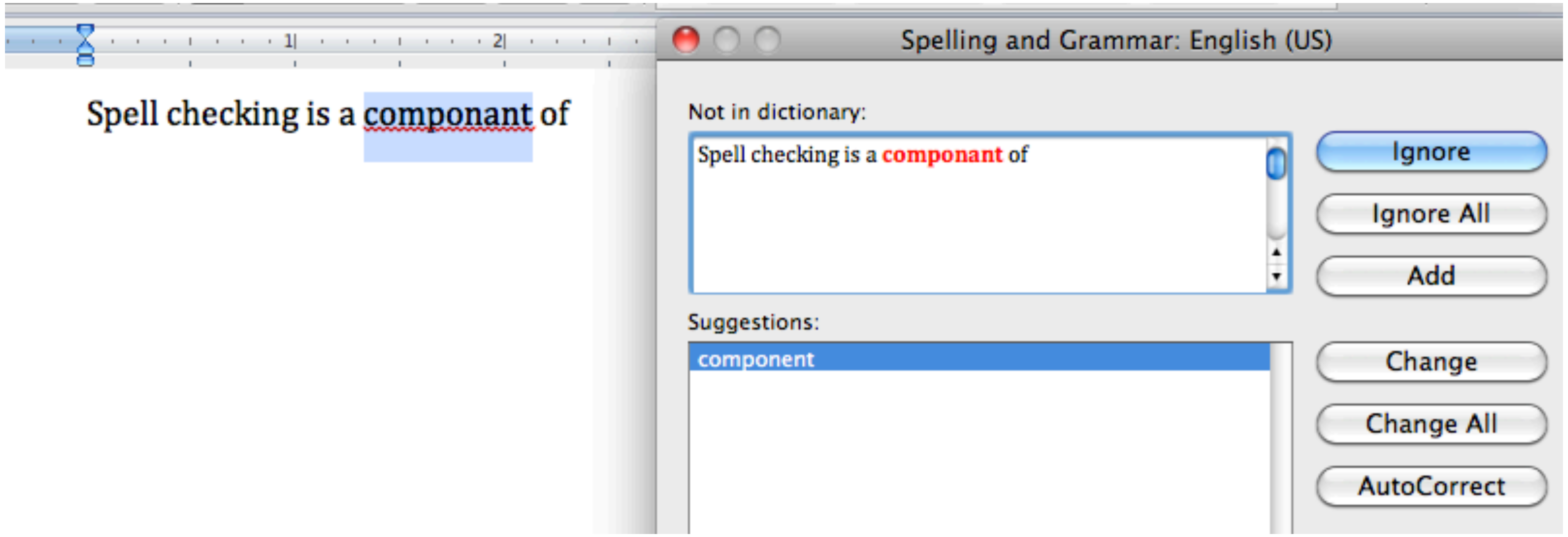
*Syntax, Parsing (*Structure of sentences *)*

Morphology, Lexicon Words & their forms (Words & their forms)

*Phonetics & Phonology (*Speech sound)

Applications for spelling correction

Word processing



Spell checking is a componant of

Spelling and Grammar: English (US)

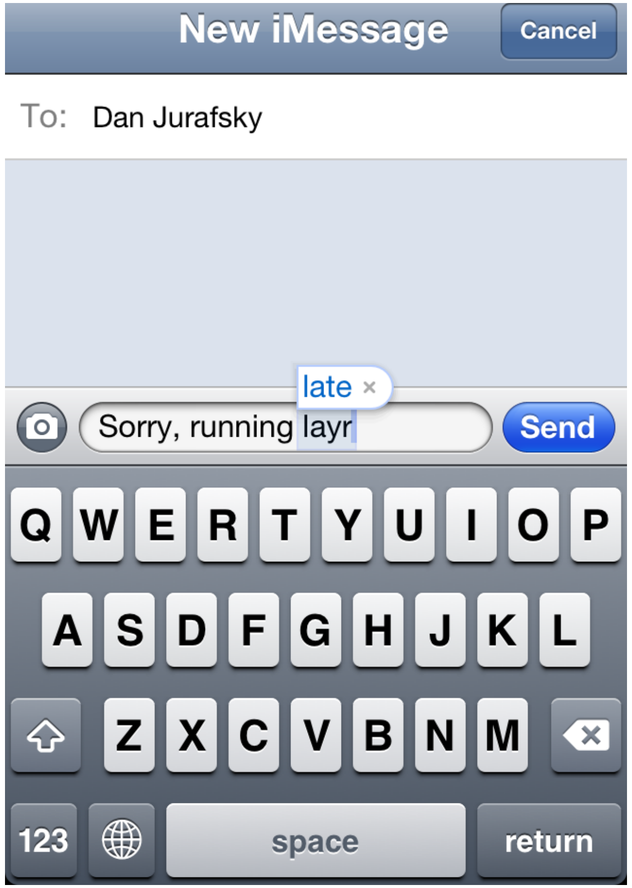
Not in dictionary:

Spell checking is a componant of

Suggestions:

component

Phones



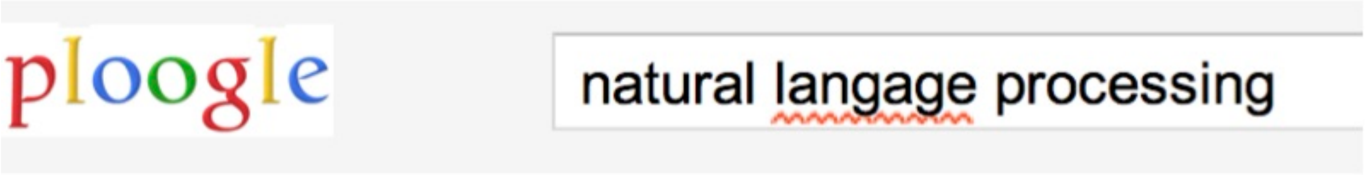
New iMessage

To: Dan Jurafsky

Sorry, running layr

late ×

Web search



natural langage processing

Showing results for natural ***language*** processing

Search instead for natural langage processing

Spelling correction

f(

- Cin seddi dunyanın yedi mocusəsindən birdir.
- Niye?
- Cunki duzəldikleti en uzunomurlu sheydi.

)=

-Çin səddi dünyanın yeddi möcüzəsindən biridir.
- Niyə?
- Çünki düzəltdikləti ən uzunömürlü şeydir.

Machine translation

Original text	When the space organization NASA first started sending up astronaunts they discovered ballpoint pens would not work in zero gravity. To solve the problem, NASA scientists spent ten years and 12 billion to develop a pen that would write in zero gravity, upside down, under water, on all types of surface, and at temperatures ranging from below freezing to 300C. Russians used a pencil.
Google	NASA kosmik təşkilatı ilk dəfə astronavtların göndərilməsinə başladıqları zaman, ballpoint qələmləri sıfır çəkisi ilə işləməyəcəkdi. Problemi həll etmək üçün, NASA alimləri sıfır ağırlıq, baş aşağı, su altında, bütün səthlərdə və aşağıda dondurmadan 300 dərəcə qədər dəyişən temperaturda yazacaq bir qələm hazırlamaq üçün on il və 12 milyard dollar sərf etmişdir. Ruslar bir qələm istifadə edirdi.
Dilmanc	Kosmos təşkilatı NASA ilk astronaunts qaldıraraq başlayanda onlar kəşf diyircəkli qələmlər çəkisizlik şəraitində işləməyəcək. Problemin həlli, NASA alimləri çəkisizlik yazacaq qələm inkişaf etdirmək, on il və 12 milyard xərcləyib, tərsinə, su altında, bütün səthinin növləri, 300c-a şaxta tutmuş və temperaturda . karandaş istifadə olunan Ruslar.

KU
PLANET
PARNI IZ BAKU
PLANET
PARNI IZ BAKU
KIA
KIA MOTORS
The Power to Surprise™
Beynəlxa

Why is machine translation hard?

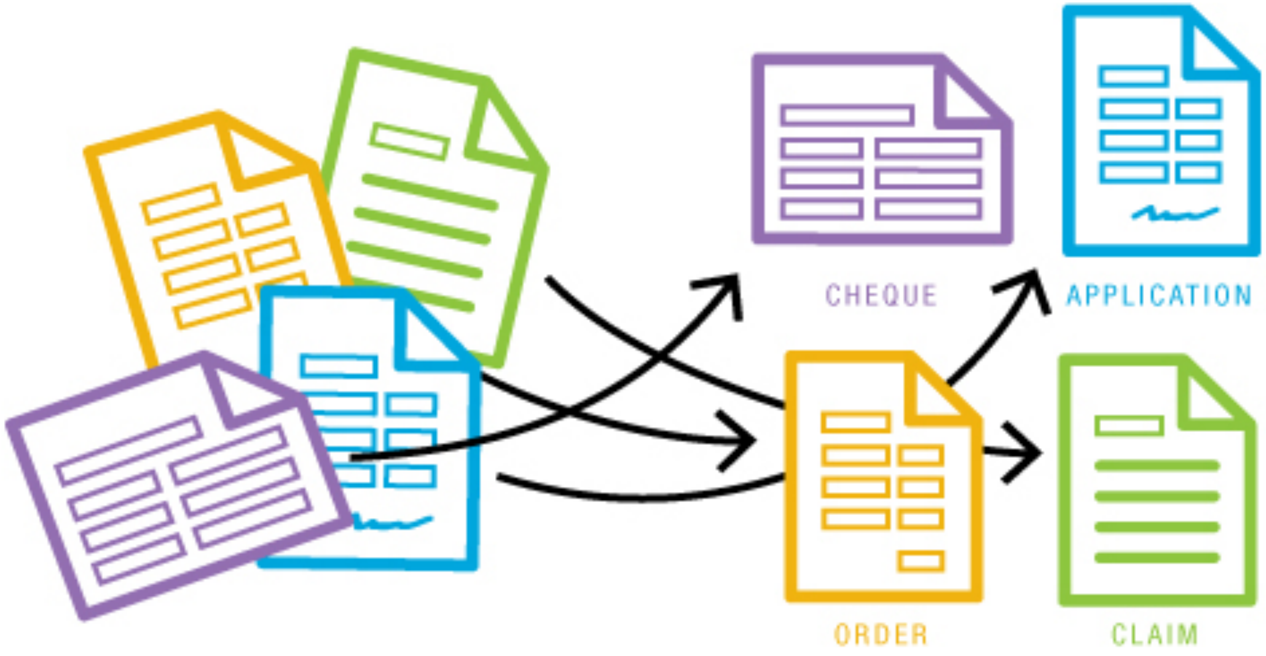
- Requires both understanding the “from” language and generating the “to” language.

- How can we teach a computer a “second language” when it doesn’t even really have a first language?

- Can we do machine translation without solving *natural language understanding* and *natural language generation* first?

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …



SENTIMENT ANALYSIS

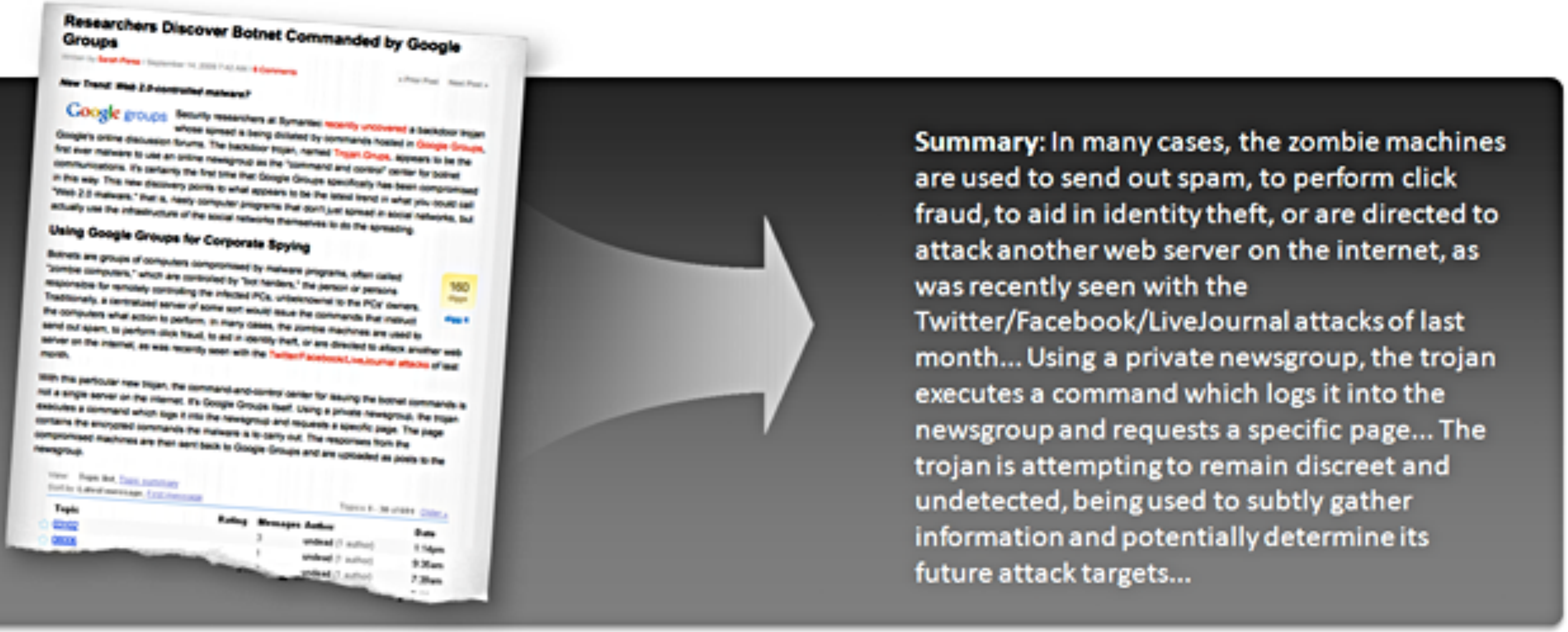
Discovering people opinions, emotions and feelings about a product or service

Information retrieval

- Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources.

Text Summarization

- **Goal**: produce an abridged version of a text that contains information that is important or relevant to a user.



Researchers Discover Botnet Commanded by Google Groups

[illegible]

New Trend: Web 2.0-controlled malware?

Security researchers at Symantec recently uncovered a backdoor trojan whose spread is being dictated by commands hosted in Google Groups, Google's online discussion forums. The backdoor trojan, named Trojan.Grups, appears to be the first ever malware to use an online newsgroup as the "command and control" center for botnet communications. It's certainly the first time that Google Groups specifically has been compromised in this way. This new discovery points to what appears to be the latest trend in what you could call "Web 2.0 malware," that is, nasty computer programs that don't just spread in social networks, but actually use the infrastructure of the social networks themselves to do the spreading.

Using Google Groups for Corporate Spying

Botnets are groups of computers compromised by malware programs, often called "zombie computers," which are controlled by "bot herders," the person or persons responsible for remotely controlling the infected PCs, unbeknownst to the PCs' owners. Traditionally, a centralized server of some sort would issue the commands that instruct the computers what action to perform. In many cases, the zombie machines are used to send out spam, to perform click fraud, to aid in identity theft, or are directed to attack another web server on the internet, as was recently seen with the Twitter/Facebook/LiveJournal attacks of last month.

With this particular new trojan, the command-and-control center for issuing the botnet commands is not a single server on the internet. It's Google Groups itself. Using a private newsgroup, the trojan executes a command which logs it into the newsgroup and requests a specific page. The page contains the encrypted commands the malware is to carry out. The responses from the compromised machines are then sent back to Google Groups and are uploaded as posts to the newsgroup.

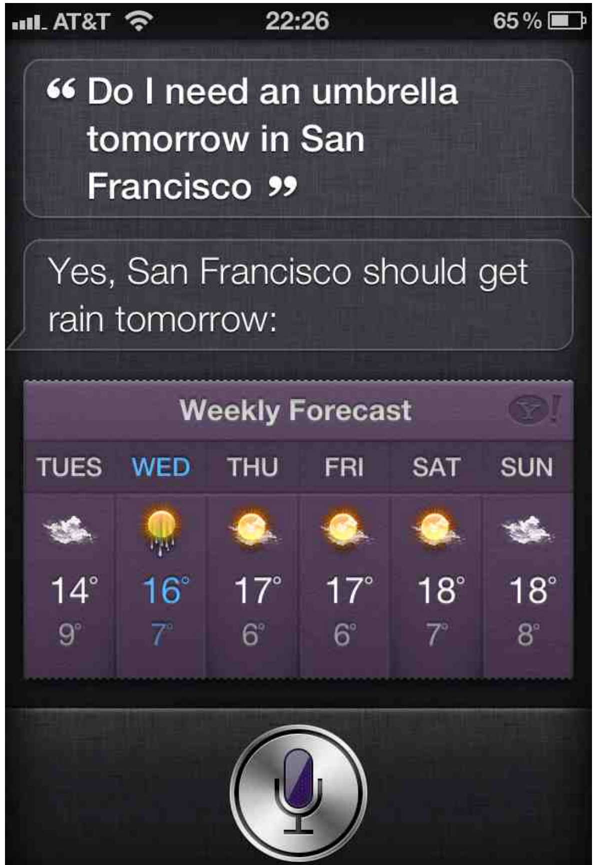
Topic Rating Messages Author Date

Summary: In many cases, the zombie machines are used to send out spam, to perform click fraud, to aid in identity theft, or are directed to attack another web server on the internet, as was recently seen with the Twitter/Facebook/LiveJournal attacks of last month... Using a private newsgroup, the trojan executes a command which logs it into the newsgroup and requests a specific page... The trojan is attempting to remain discreet and undetected, being used to subtly gather information and potentially determine its future attack targets...

Dialogue systems

- A **dialog system** or **conversational agent** (**CA**) is a computer system intended to converse with a human, with a coherent structure.

Question Answering:



“ Do I need an umbrella tomorrow in San Francisco ”

Yes, San Francisco should get rain tomorrow:

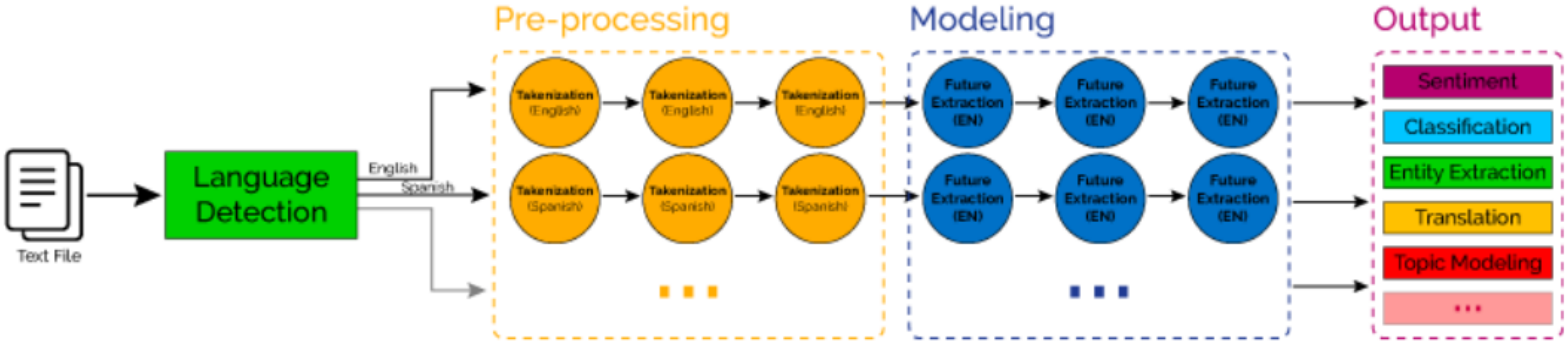
Weekly Forecast

TUES	WED	THU	FRI	SAT	SUN
14°	16°	17°	17°	18°	18°
9°	7°	6°	6°	7°	8°

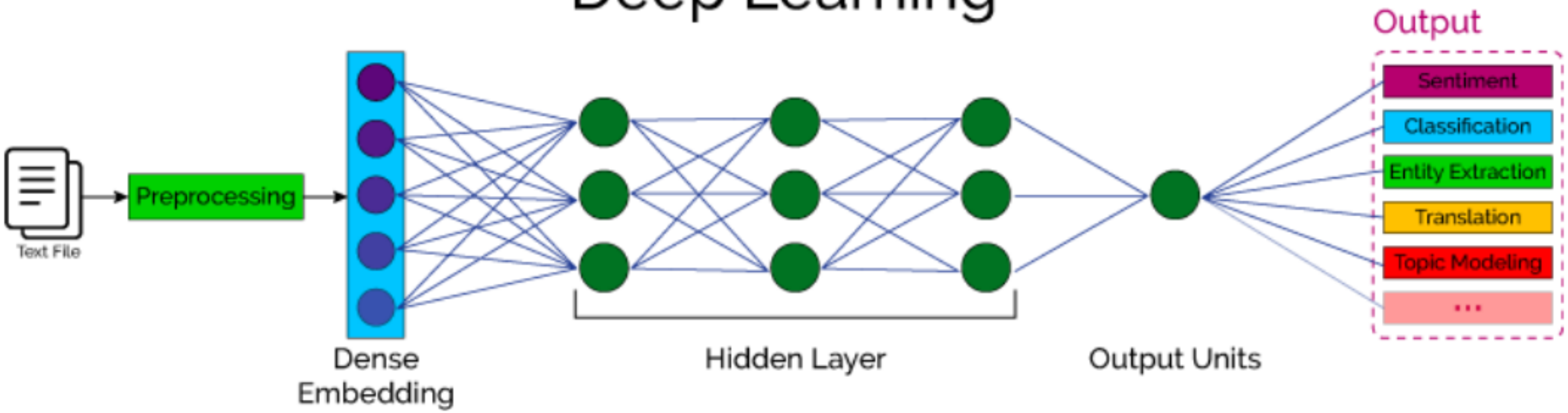
Deep Learning Algorithms	NLP Usage
Neural Network – NN (feed)	•Part-of-speech Tagging •Tokenization •Named Entity Recognition •Intent Extraction
Recurrent Neural Networks -(RNN)	•Machine Translation •Question Answering System •Image Captioning
Recursive Neural Networks	•Parsing sentences •Sentiment Analysis •Paraphrase detection •Relation Classification •Object detection
Convolutional Neural Network -(CNN)	•Sentence/ Text classification •Relation extraction and classification •Spam detection •Categorization of search queries •Semantic relation extraction

Difference Between Classical NLP & Deep Learning NLP

Classical NLP



Deep Learning



References:

- 1. Overview of Artificial Intelligence and Natural Language Processing. NAVDEEP SINGH GILL. https://www.upwork.com/hiring/for-clients/artificial-intelligence-and-natural-language-processing-in-big-data/

2. **Natural Language Processing Market to Reach $22.3 Billion by 2025**

https://www.tractica.com/newsroom/press-releases/natural-language-processing-market-to-reach-22-3-billion-by-2025/

3. Dan Jurafsky. Natural Language Processing Lectures.

4. BİL711 Natural Language Processing. Prof. Dr. İlyas Çiçekli.