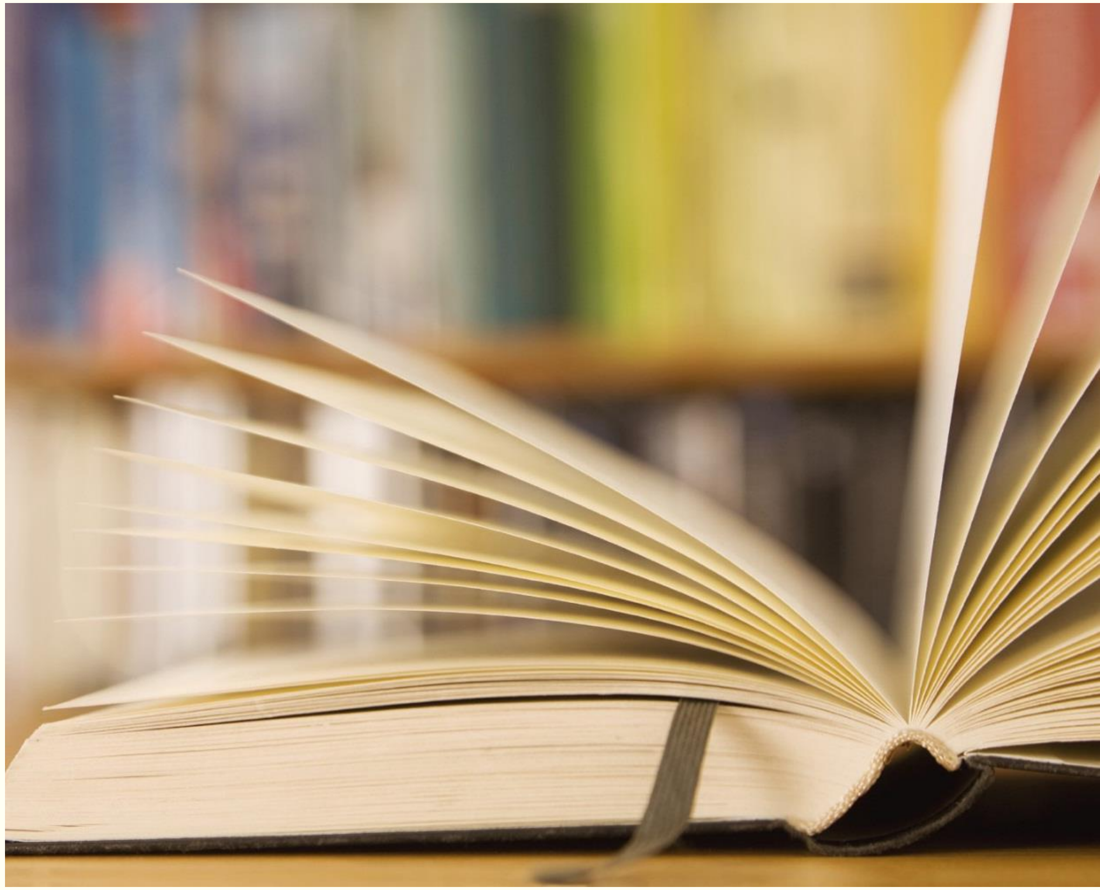
NATURAL LANGUAGE PROCESSING

Dr. G. Bharadwaja Kumar

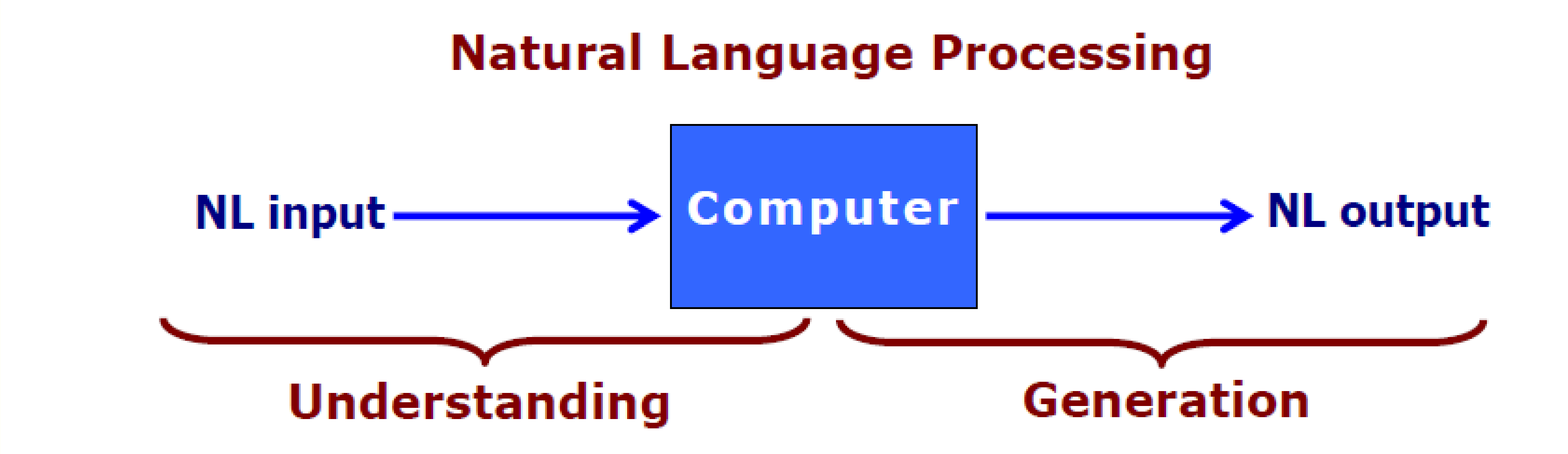


NATURAL?

- Natural Language?
  - Refers to the language spoken by people, e.g. English, Telugu, Tamil, as opposed to artificial languages, like C++, Java, etc.
- Natural Language Processing
  - Applications that deal with natural language in a way or another
- Computational Linguistics
  - Deals Linguistics aspects of languages and closely related NLP

NATURAL LANGUAGE PROCESSING?

- Goal is to enable computers to understand, generate and communicate with people using human languages.



Why Natural Language Processing

- Computers "see" text in English the same you have seen the text here!
    - kJfmmfj mmmvvv nnnffn333
    - Uj iheale eleee mnster vensi credur
    - Baboi oi cestnitze
    - Coovoel2^ ekk; ldsllk lkdf vnnjfj?
    - Fgmflmllk mlfm kfre xnnn!
- People have no trouble understanding language
    - Common sense
    - Reasoning capacity
    - World knowledge

- Computers does not have all the above
- The goal of NLP is to go way beyond just string processing or keyword matching!

Why NLP is difficult

- Many hidden variables
  - Knowledge about the world
  - Knowledge about the context
  - Knowledge about human communication techniques
    - *Can you tell me the time?*
- Problem of scale
  - Many (infinite?) possible words, meanings, context
- Problem of sparsity
  - Very difficult to do statistical analysis, most things (words, concepts) are never seen before
- Long range correlations

Natural Languages vs. Computer Languages

Ambiguity is the primary difference between natural and computer languages.

Formal programming languages are designed to be unambiguous, i.e. they can be defined by a grammar (deterministic context-free languages (DCLFs))that produces a unique parse for each sentence in the language.

Any Light at The End of The Tunnel?

- Yahoo, Google, Microsoft → Information Retrieval
- Monster.com, HotJobs.com (Job finders) → Information Extraction + Information Retrieval
- Systran powers Babelfish → Machine Translation
- Ask Jeeves → Question Answering
- Myspace, Facebook, Blogspot → Processing of User-Generated Content
- All “Big Guys” have (several) strong NLP research labs:
  - IBM, Microsoft, AT&T, Xerox, Sun, etc.

Various Levels of Natural Language Processing

Phonetics and phonology	The study of language sounds
Orthography	The study of language conventions for punctuation, script and encoding
Morphology	The study of meaningful components of words
Syntax	The study of structural relationships among words
Lexical semantics	The study of word meaning
Compositional semantics	The study of the meaning of sentences
Pragmatics	The study of the meaning in terms of the situational context
Discourse	The study of comprehending the intension and meaning more than sentences

Phonetics & phonology

- Phonetics is the study of language at the level of sounds while phonology is the study of combination of sounds into organized units of speech, the formation of syllables and larger units.

- Phonetic and phonological knowledge are essential for speech based systems as they deal with how words are related to the sounds that realize them.

Orthography -- Word & Sentence Segmentation

- Breaking a string of characters (graphemes) into a sequence of words and sentences.
- Generally words are separated by spaces and sentences are separated by boundary markers.
- Even in English, more than one sentence boundary marker {; ! ?}

Morphological Analysis

- *Morphology* is the field of linguistics that studies the internal structure of words.
- A *morpheme* is the smallest linguistic unit that has semantic meaning or grammatical function
- Morphological analysis is the task of segmenting a word into its morphemes:
  - carried ⇒ carry + ed (past tense)
  - independently ⇒ in + (depend + ent) + ly
  - Googlers ⇒ (Google + er) + s (plural)
  - Antidisestablishmentarianism ⇒ (Anti+dis+**establish+**ment+ary+an+ism)
  - Infix and Circumfix (**Abso-frickin-lutely)** (em+bold+en)in other languages

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech

I ate the spaghetti with meatballs.
Pro V Det N Prep N

John saw the saw and decided to take it to the table.
PN V Det N Con V Part V Pro Prep Det N

- Useful for subsequent syntactic parsing and word sense disambiguation.

Syntax

- Syntax concerns the proper ordering of words and its affect on meaning.
  - The dog bit the boy.
  - The boy bit the dog.
  - * Bit boy dog the the.

Syntax

•If a verb takes objects, then it is a **transitive verb**.

Example:
They played **soccer**. → (The verb ***play*** takes ONE object '**soccer**')
They sent **him** **a postcard**. → (The verb ***send*** takes TWO objects '**him**' and '**a postcard**')

•If a verb doesn't take an object, then it is **an intransitive verb**.
Example:
She lies. → (The verb '***lie***' doesn't take any object)
The building collapsed. → (The verb ***'collapse'*** doesn't take any object)

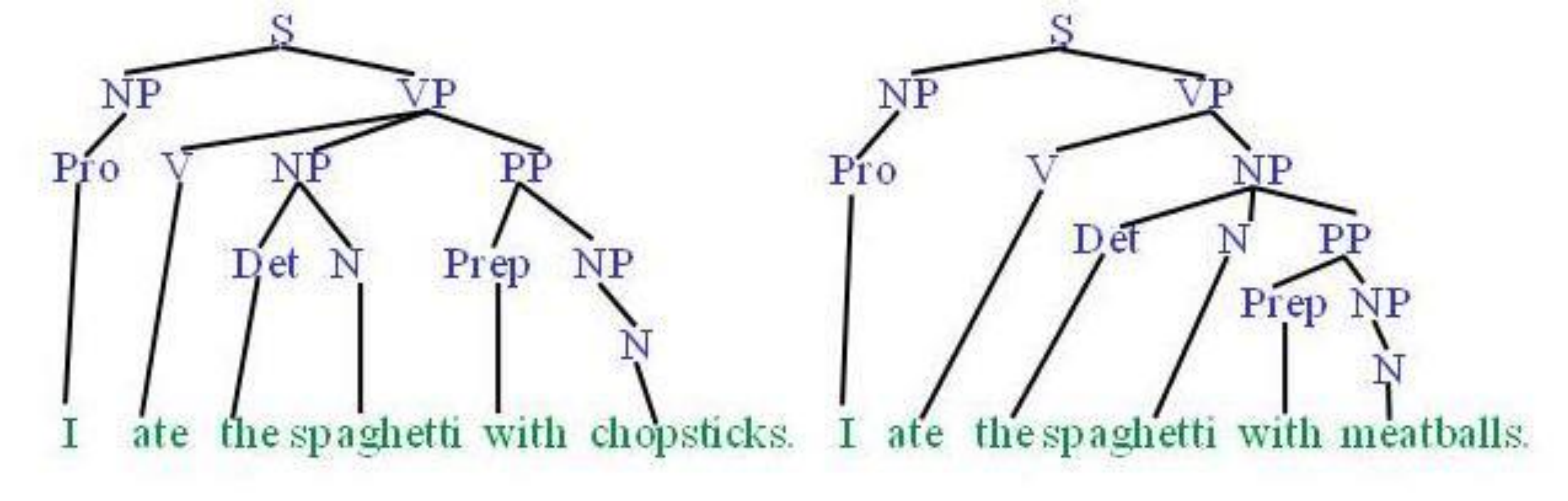
·They named **the boy Christopher**.

Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.


S
NP
VP
Pro
V
NP
PP
Det
N
Prep
NP
N
I ate the spaghetti with chopsticks.
S
NP
VP
Pro
V
NP
Det
N
PP
Prep
NP
N
I ate the spaghetti with meatballs.

Semantics: Selectional Restrictions

- The restaurant serves green-lipped mussels.
  - THEME is some kind of food
- Which airlines serve Denver?
  - THEME is an appropriate location

Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
  - Ellen has a strong interest in computational linguistics.
  - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is an argument to the verb.

  agent patient source destination instruments

  - John drove Mary from Austin to Dallas in his Toyota Prius.
  - The hammer broke the window.
- Also referred to a "case role analysis," "thematic analysis," and "shallow semantic parsing"

Compositional Semantic Tasks

- Semantics at phrase and sentence level
  - *The old horse finally **kicked the bucket**.*
  - I think my sewing machine has **kicked the bucket**.
  - *My old backpack finally **bit the dust** the other day.*
  - *Colorless green ideas sleep furiously.*
  - *The rat killed the snake and swallowed it.*

Discourse Tasks

- Anaphora Resolution/ Co-Reference
- Determine which phrases in a document refer to the same underlying entity.
  - John put the carrot on the plate and ate it.
  - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
  - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Pragmatics -

- "Jill and Mary are mothers." – (each is independently a mother).
- "Jill and Mary are sisters." – (they are sisters of each other)

'What time do you call this?'

- Literal Meaning: What time is it?
  Literal Response: A time (e.g. 'twenty to one.')
- (Pragmatic Meaning: a different question entirely, e.g. Why are you so late?
  Pragmatic Response: Explain the reason for being so late.)

- Peter: Is John a good accountant?
- Mary: John is a computer.

- Peter: I heard that you have moved from Manhattan to Brooklyn.
- Mary: The rent is lower.

World Knowledge

- Some times the computers may need human level thinking to understand.

- Can computers think like humans?

- Does computer has common sense like humans?

- A snake killed the rat and swallowed it.

- *What if I say the below sentence?*

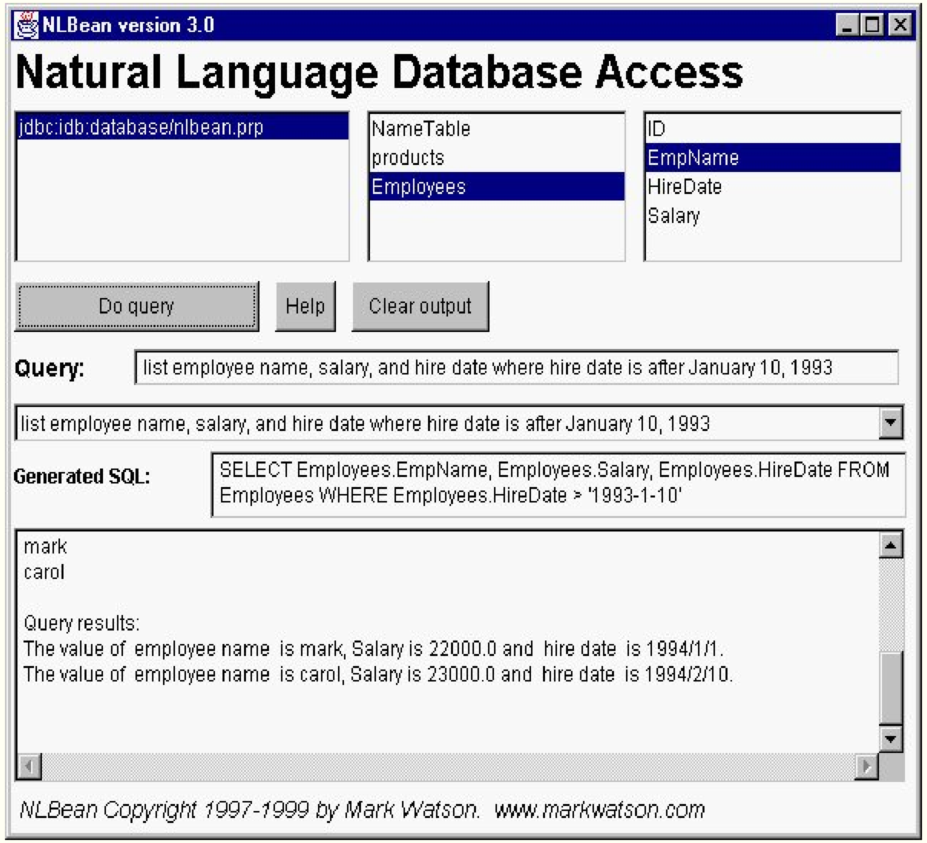
- A rat killed the snake and swallowed it.
    - Can a computer say the above sentence is ridiculous?

APPLICATIONS OF NLP

Applications of NLP

- Natural Language Interfaces

Spell Checking



Natural Language Database Access

Query: list employee name, salary, and hire date where hire date is after January 10, 1993

list employee name, salary, and hire date where hire date is after January 10, 1993

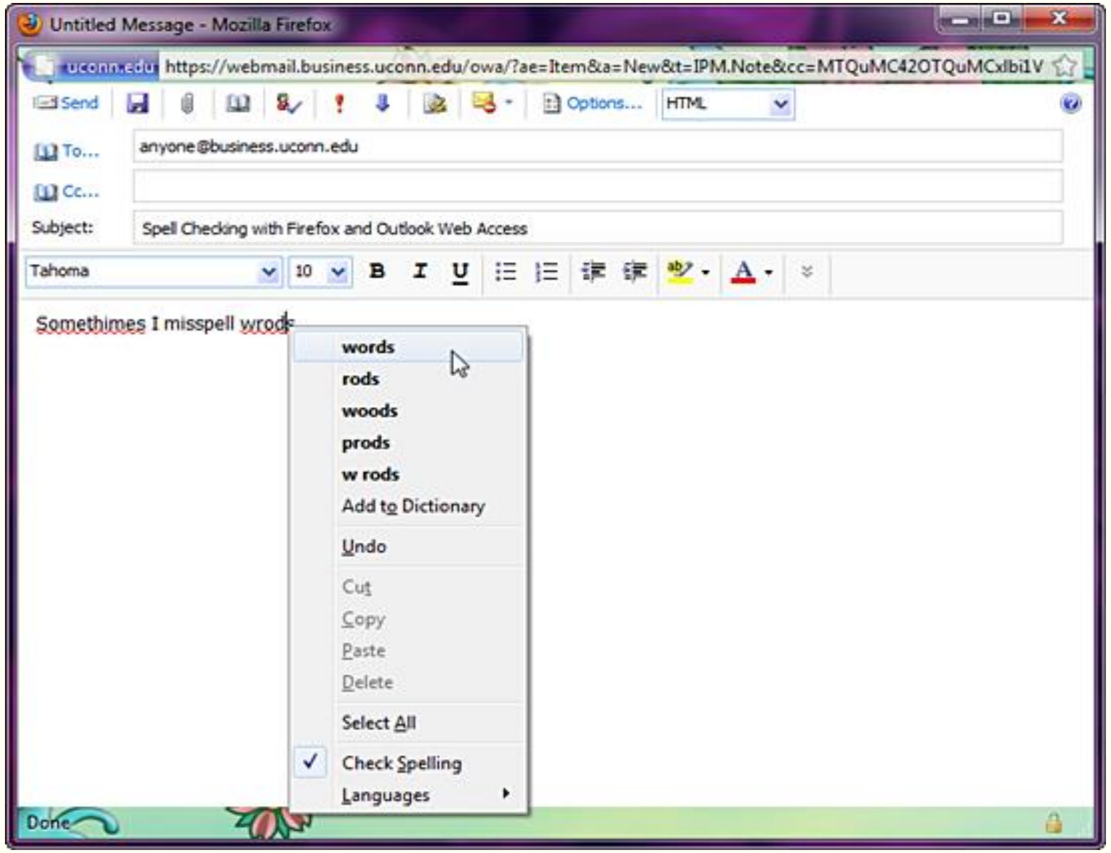
Generated SQL:

```
SELECT Employees.EmpName, Employees.Salary, Employees.HireDate FROM
Employees WHERE Employees.HireDate > '1993-1-10'
```

```
mark
carol

Query results:
The value of  employee name  is mark, Salary is 22000.0 and  hire date  is 1994/1/1.
The value of  employee name  is carol, Salary is 23000.0 and  hire date  is 1994/2/10.
```

NLBean Copyright 1997-1999 by Mark Watson. www.markwatson.com



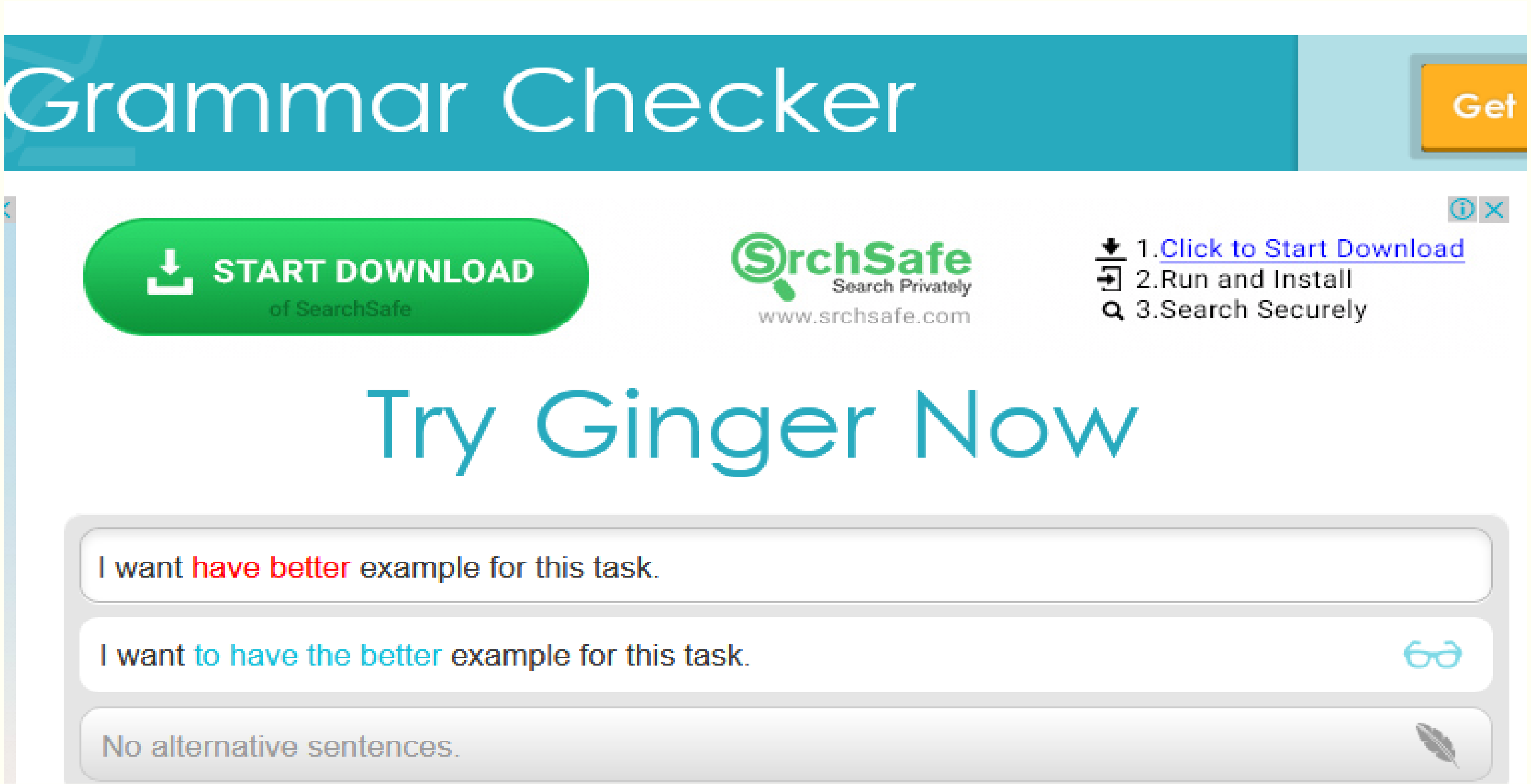
https://webmail.business.uconn.edu/owa/?ae=Item&a=New&t=IPM.Note&cc=MTQuMC42OTQuMCxlbi1V

To...: anyone@business.uconn.edu

Subject: Spell Checking with Firefox and Outlook Web Access

Somethimes I misspell wrods

Automatic Grammar Checkers



Grammar Checker

Get

Try Ginger Now

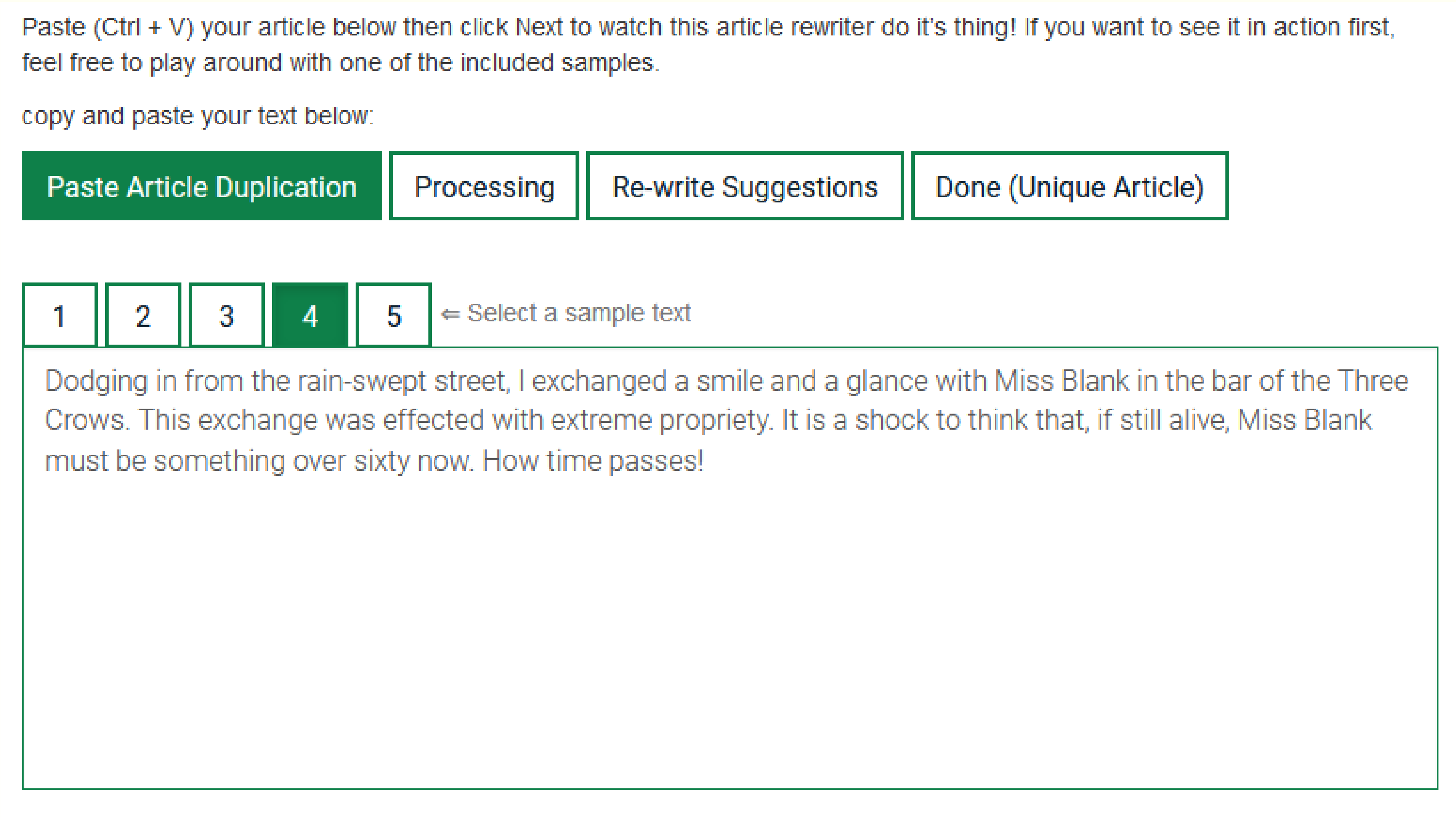
I want have better example for this task.

I want to have the better example for this task.

No alternative sentences.

Article Rewriter

Small SEO Tools
Free Online SEO Tools



Paste (Ctrl + V) your article below then click Next to watch this article rewriter do it's thing! If you want to see it in action first, feel free to play around with one of the included samples.

copy and paste your text below:

Dodging in from the rain-swept street, I exchanged a smile and a glance with Miss Blank in the bar of the Three Crows. This exchange was effected with extreme propriety. It is a shock to think that, if still alive, Miss Blank must be something over sixty now. How time passes!

Cross-Language IR

Retrieving information written in a language different from the language of the user's query

يا ليلي يا عيني

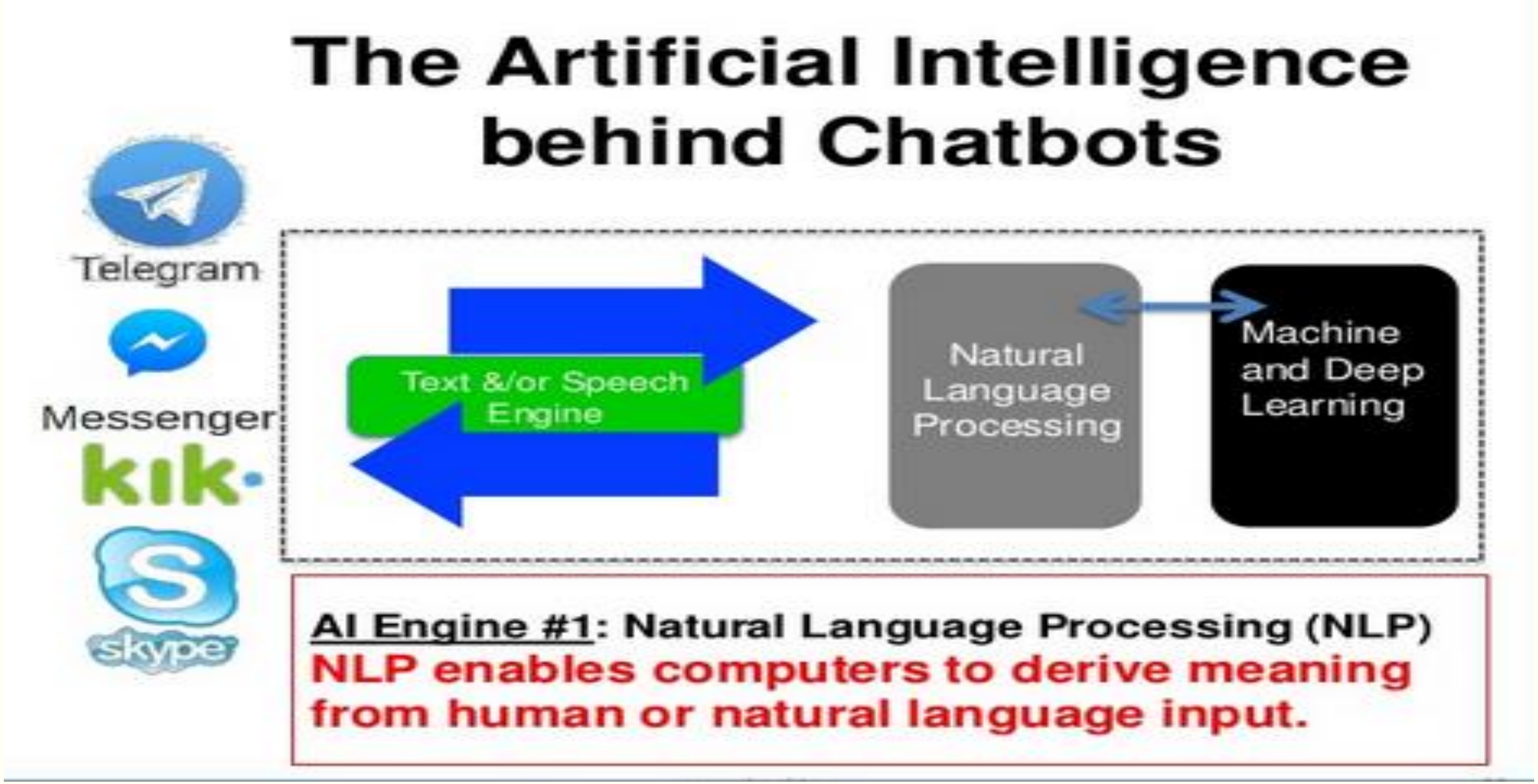
Исследований

高等学校

att förstå

których można

Chatbots



The Artificial Intelligence behind Chatbots

Telegram

Messenger

kik·

skype

Text &/or Speech Engine

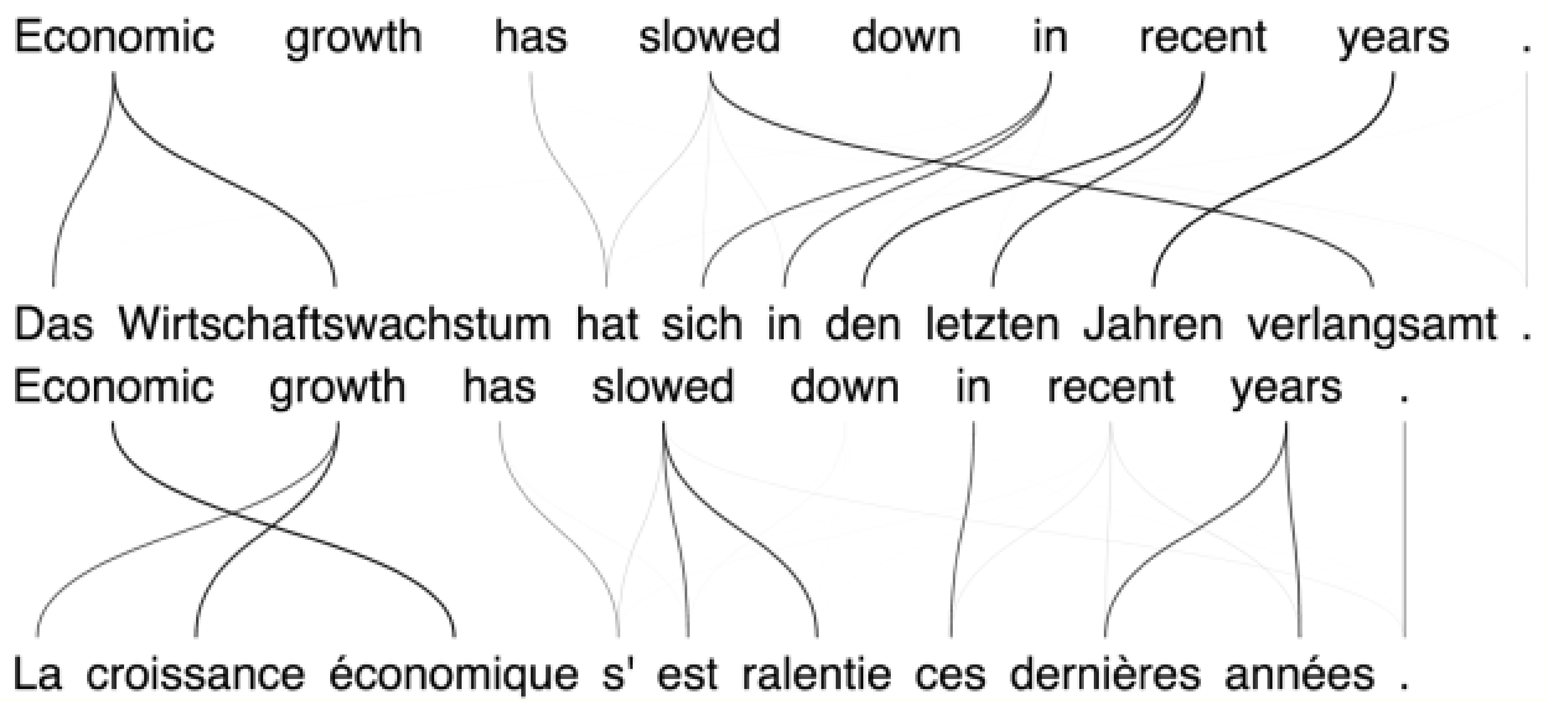
Natural Language Processing

Machine and Deep Learning

AI Engine #1: Natural Language Processing (NLP)

NLP enables computers to derive meaning from human or natural language input.

Machine Translation



Economic growth has slowed down in recent years .

Das Wirtschaftswachstum hat sich in den letzten Jahren verlangsamt .

Economic growth has slowed down in recent years .

La croissance économique s' est ralentie ces dernières années .

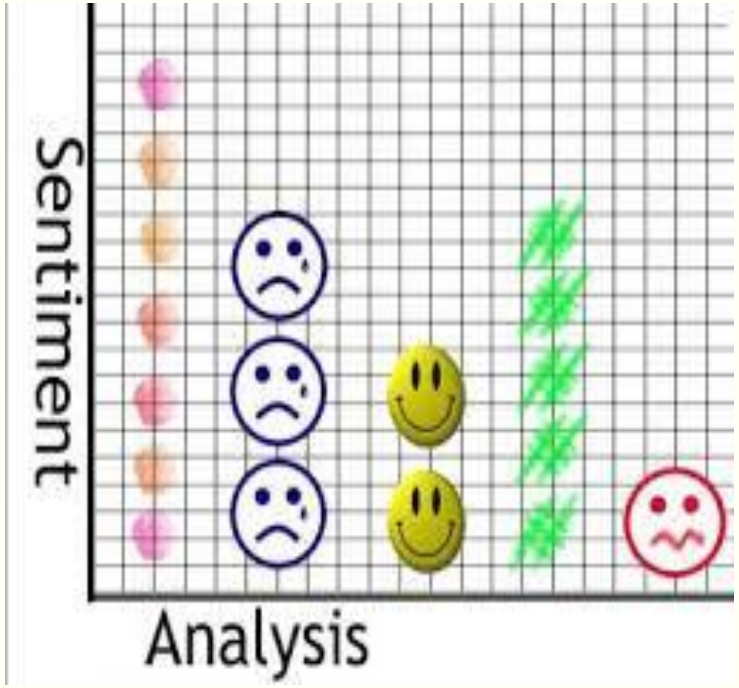
Google Translate

Break through language barriers

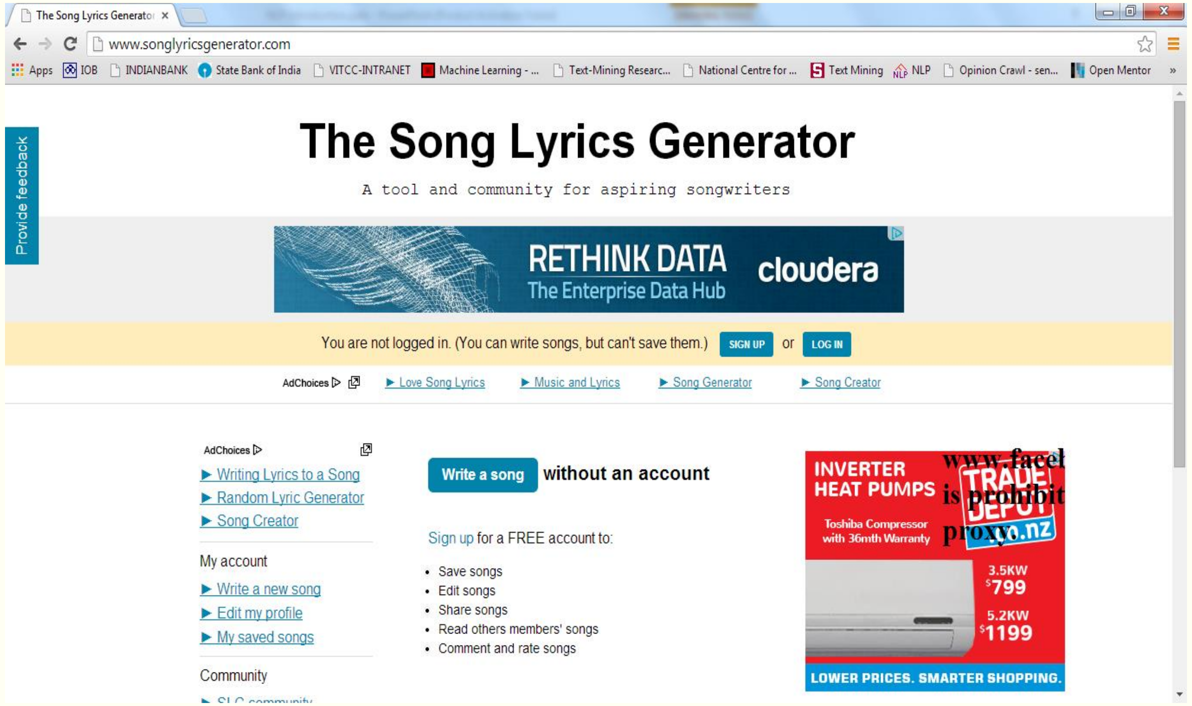
https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-3/

Applications of NLP

- Sentiment Analysis



Automatic Lyrics Generation



The Song Lyrics Generator

A tool and community for aspiring songwriters

You are not logged in. (You can write songs, but can't save them.) SIGN UP or LOG IN

Write a song without an account

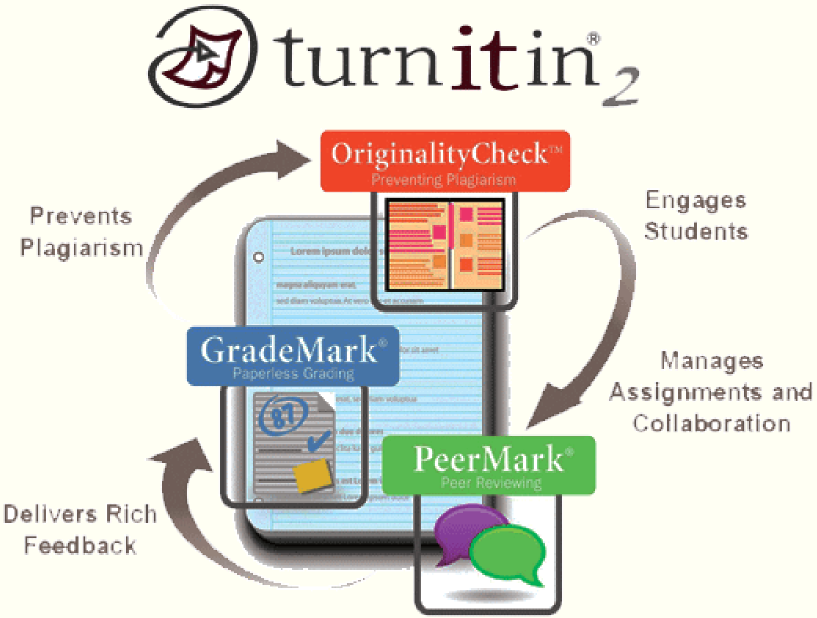
Sign up for a FREE account to:

- Save songs
- Edit songs
- Share songs
- Read others members' songs
- Comment and rate songs

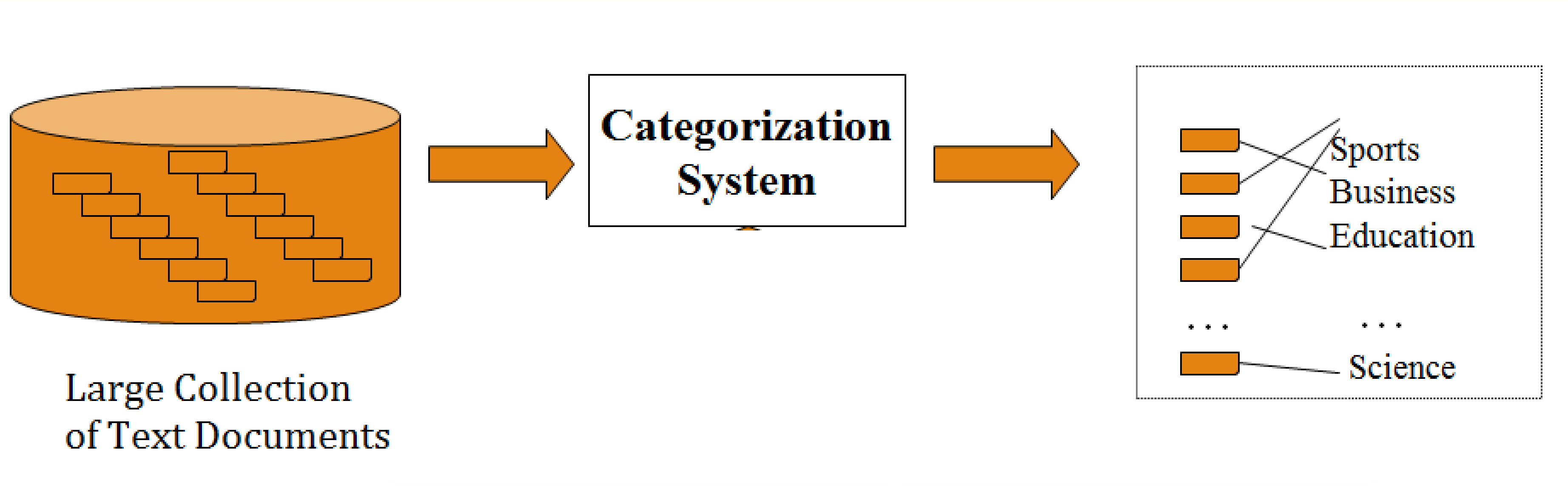
Applications of NLP

- Plagiarism Detection

Speech Recognition



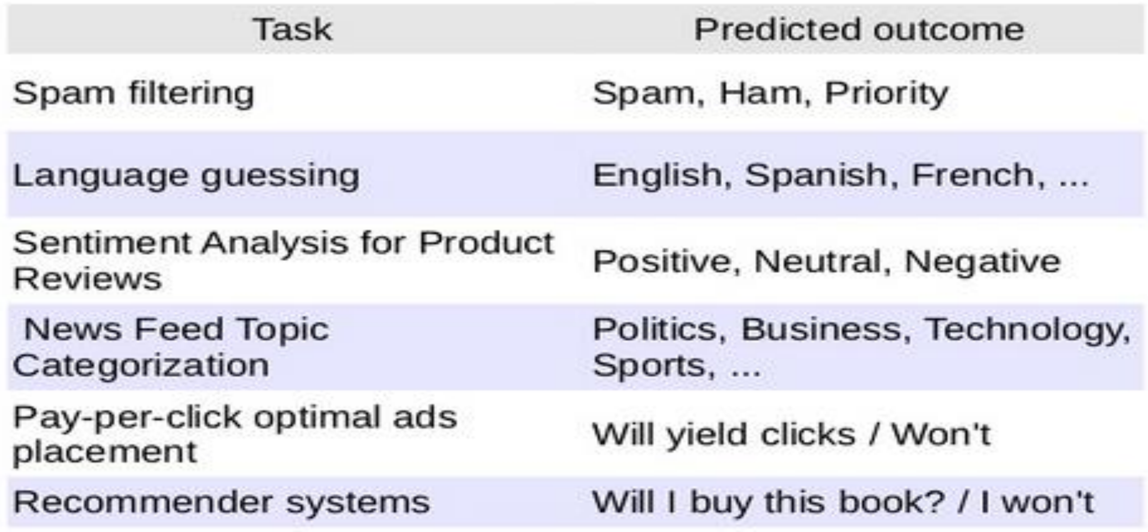
Text Categorization



Categorization System

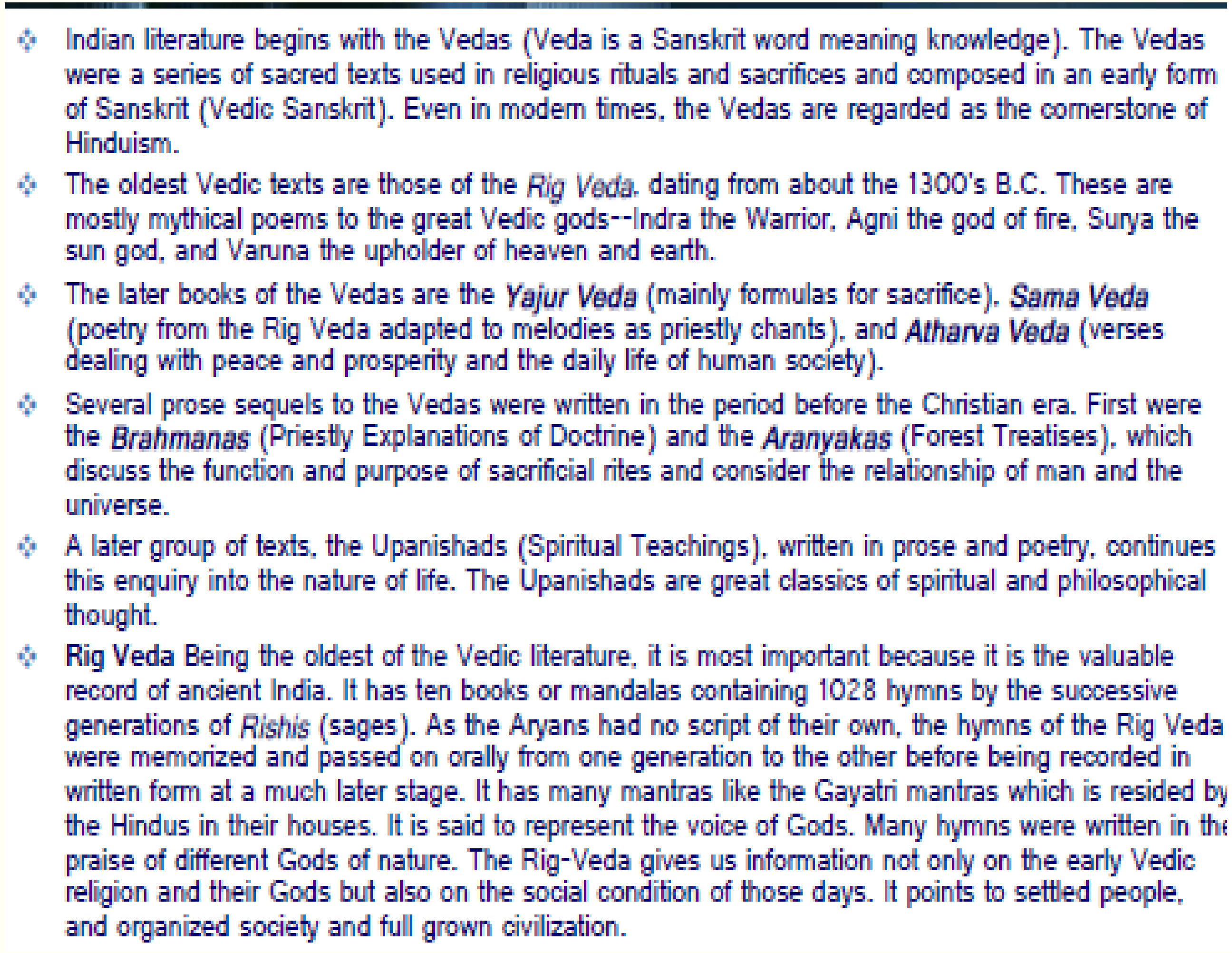
Sports
Business
Education
. . .
. . .
Science

Large Collection
of Text Documents

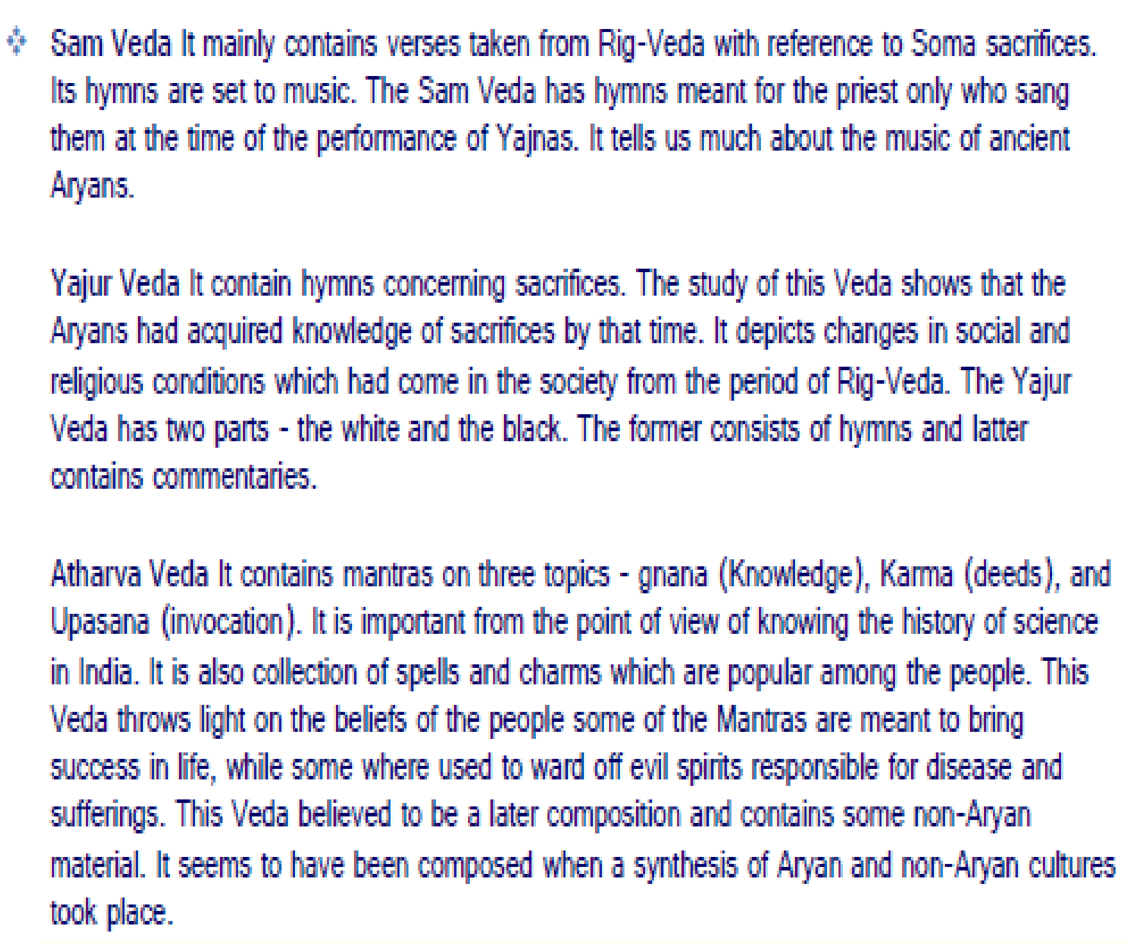


Task	Predicted outcome
Spam filtering	Spam, Ham, Priority
Language guessing	English, Spanish, French, ...
Sentiment Analysis for Product Reviews	Positive, Neutral, Negative
News Feed Topic Categorization	Politics, Business, Technology, Sports, ...
Pay-per-click optimal ads placement	Will yield clicks / Won't
Recommender systems	Will I buy this book? / I won't

Text Summarization



- Indian literature begins with the Vedas (Veda is a Sanskrit word meaning knowledge). The Vedas were a series of sacred texts used in religious rituals and sacrifices and composed in an early form of Sanskrit (Vedic Sanskrit). Even in modern times, the Vedas are regarded as the cornerstone of Hinduism.
- The oldest Vedic texts are those of the *Rig Veda*, dating from about the 1300's B.C. These are mostly mythical poems to the great Vedic gods--Indra the Warrior, Agni the god of fire, Surya the sun god, and Varuna the upholder of heaven and earth.
- The later books of the Vedas are the ***Yajur Veda*** (mainly formulas for sacrifice). ***Sama Veda*** (poetry from the Rig Veda adapted to melodies as priestly chants), and ***Atharva Veda*** (verses dealing with peace and prosperity and the daily life of human society).
- Several prose sequels to the Vedas were written in the period before the Christian era. First were the ***Brahmanas*** (Priestly Explanations of Doctrine) and the ***Aranyakas*** (Forest Treatises), which discuss the function and purpose of sacrificial rites and consider the relationship of man and the universe.
- A later group of texts, the Upanishads (Spiritual Teachings), written in prose and poetry, continues this enquiry into the nature of life. The Upanishads are great classics of spiritual and philosophical thought.
- **Rig Veda** Being the oldest of the Vedic literature, it is most important because it is the valuable record of ancient India. It has ten books or mandalas containing 1028 hymns by the successive generations of *Rishis* (sages). As the Aryans had no script of their own, the hymns of the Rig Veda were memorized and passed on orally from one generation to the other before being recorded in written form at a much later stage. It has many mantras like the Gayatri mantras which is resided by the Hindus in their houses. It is said to represent the voice of Gods. Many hymns were written in the praise of different Gods of nature. The Rig-Veda gives us information not only on the early Vedic religion and their Gods but also on the social condition of those days. It points to settled people, and organized society and full grown civilization.



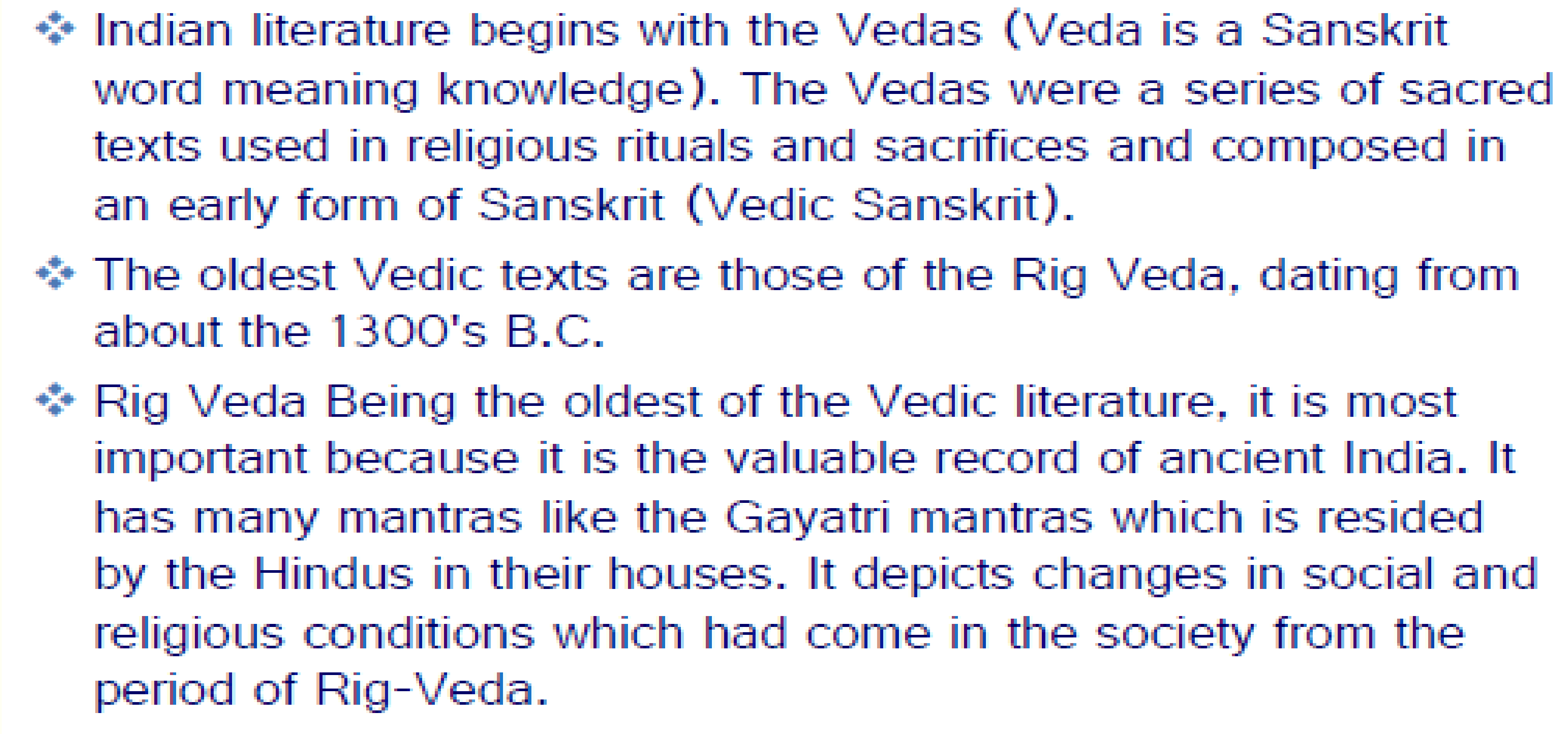
- Sam Veda It mainly contains verses taken from Rig-Veda with reference to Soma sacrifices. Its hymns are set to music. The Sam Veda has hymns meant for the priest only who sang them at the time of the performance of Yajnas. It tells us much about the music of ancient Aryans.

Yajur Veda It contain hymns concerning sacrifices. The study of this Veda shows that the Aryans had acquired knowledge of sacrifices by that time. It depicts changes in social and religious conditions which had come in the society from the period of Rig-Veda. The Yajur Veda has two parts - the white and the black. The former consists of hymns and latter contains commentaries.

Atharva Veda It contains mantras on three topics - gnana (Knowledge), Karma (deeds), and Upasana (invocation). It is important from the point of view of knowing the history of science in India. It is also collection of spells and charms which are popular among the people. This Veda throws light on the beliefs of the people some of the Mantras are meant to bring success in life, while some where used to ward off evil spirits responsible for disease and sufferings. This Veda believed to be a later composition and contains some non-Aryan material. It seems to have been composed when a synthesis of Aryan and non-Aryan cultures took place.

Summary (24%) Given by Automatic Text Summarizer

Start generating your online summary

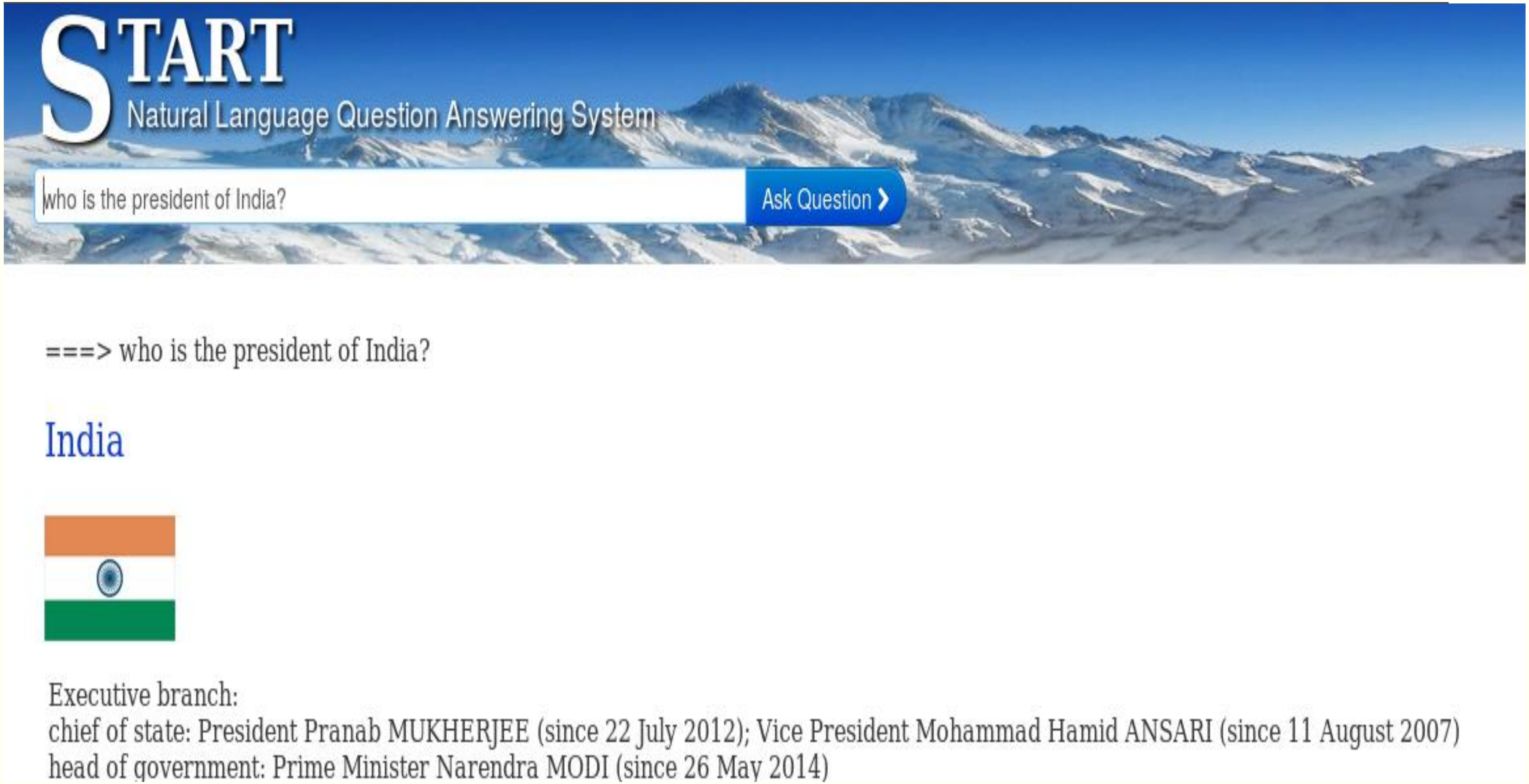


- Indian literature begins with the Vedas (Veda is a Sanskrit word meaning knowledge). The Vedas were a series of sacred texts used in religious rituals and sacrifices and composed in an early form of Sanskrit (Vedic Sanskrit).
- The oldest Vedic texts are those of the Rig Veda, dating from about the 1300's B.C.
- Rig Veda Being the oldest of the Vedic literature, it is most important because it is the valuable record of ancient India. It has many mantras like the Gayatri mantras which is resided by the Hindus in their houses. It depicts changes in social and religious conditions which had come in the society from the period of Rig-Veda.

Question Answering

IBM Watson took on 2 of the best humans in the game of Jeopardy, the famous TV quiz show in the US.

Used natural language processing and machine learning

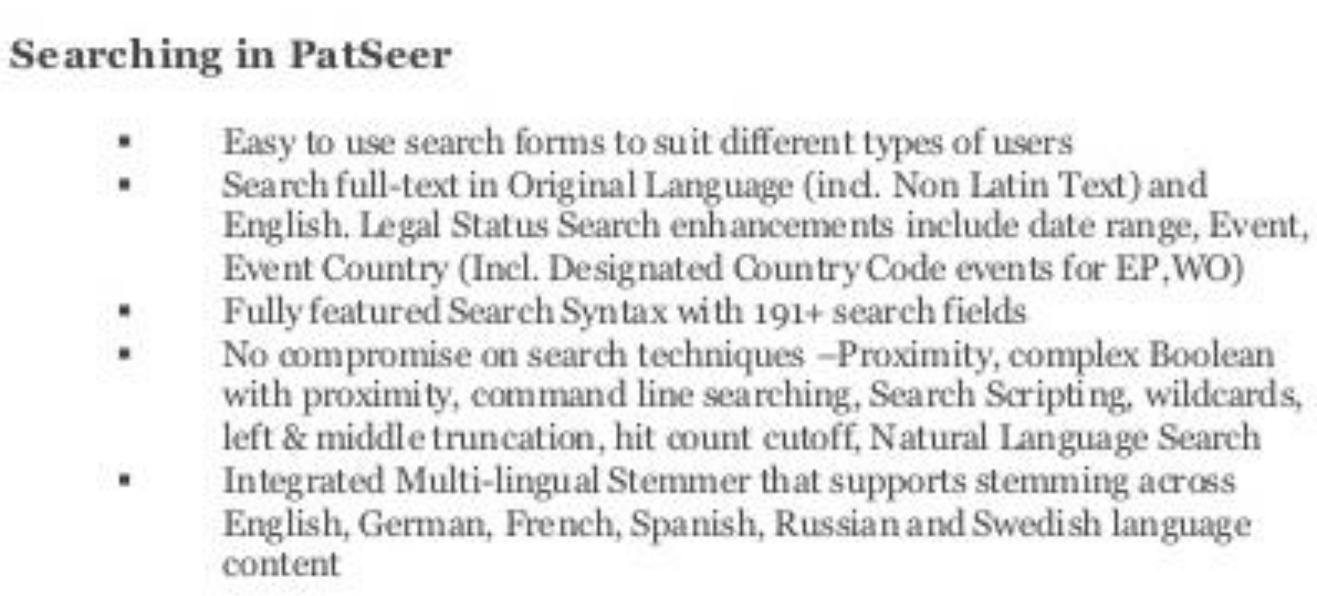

START
Natural Language Question Answering System
who is the president of India?
Ask Question ❯

===> who is the president of India?

India

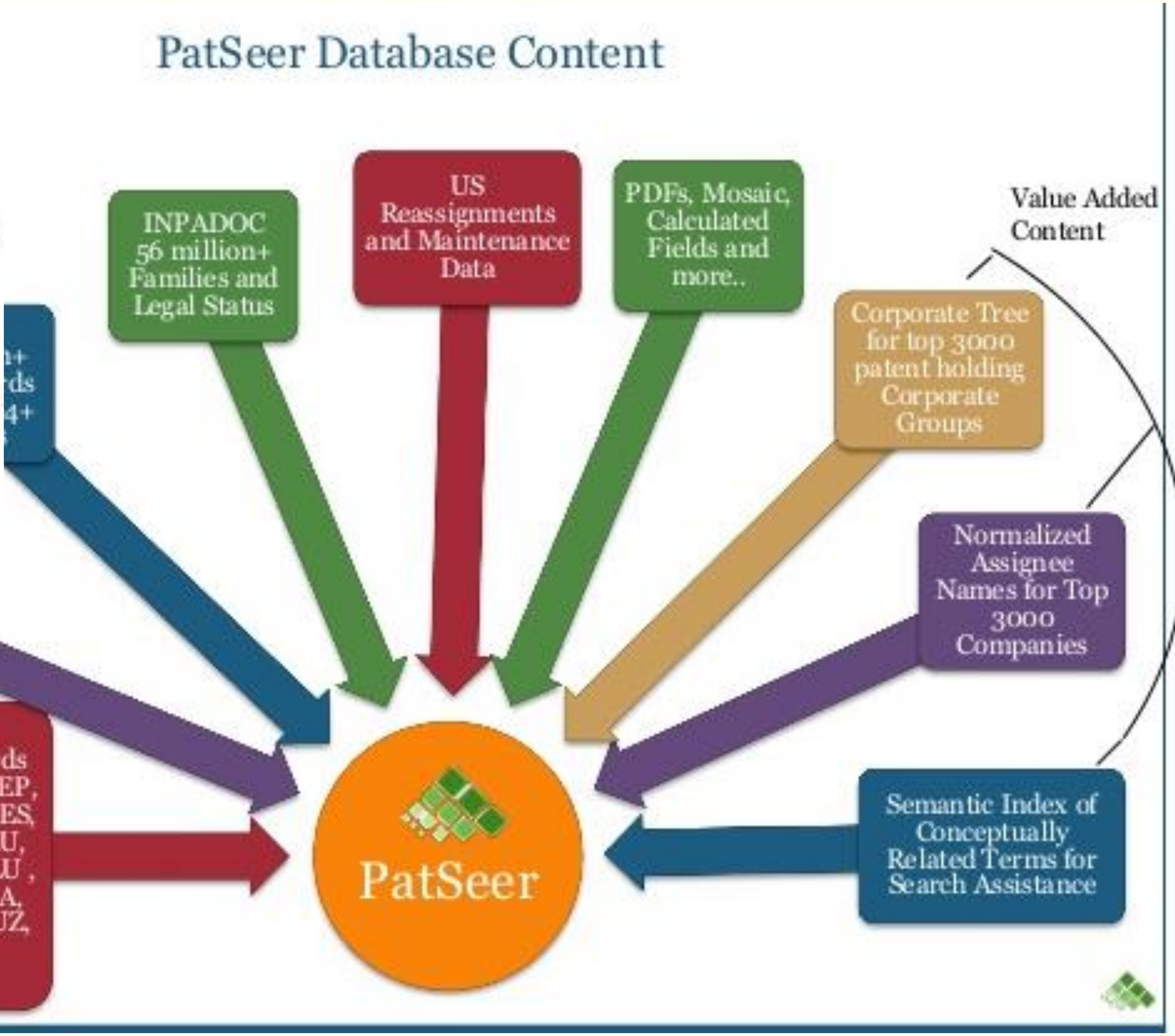
Executive branch:
chief of state: President Pranab MUKHERJEE (since 22 July 2012); Vice President Mohammad Hamid ANSARI (since 11 August 2007)
head of government: Prime Minister Narendra MODI (since 26 May 2014)

Patent Analysis



Searching in PatSeer

- Easy to use search forms to suit different types of users
- Search full-text in Original Language (incl. Non Latin Text) and English. Legal Status Search enhancements include date range, Event, Event Country (Incl. Designated Country Code events for EP,WO)
- Fully featured Search Syntax with 191+ search fields
- No compromise on search techniques –Proximity, complex Boolean with proximity, command line searching, Search Scripting, wildcards, left & middle truncation, hit count cutoff, Natural Language Search
- Integrated Multi-lingual Stemmer that supports stemming across English, German, French, Spanish, Russian and Swedish language content



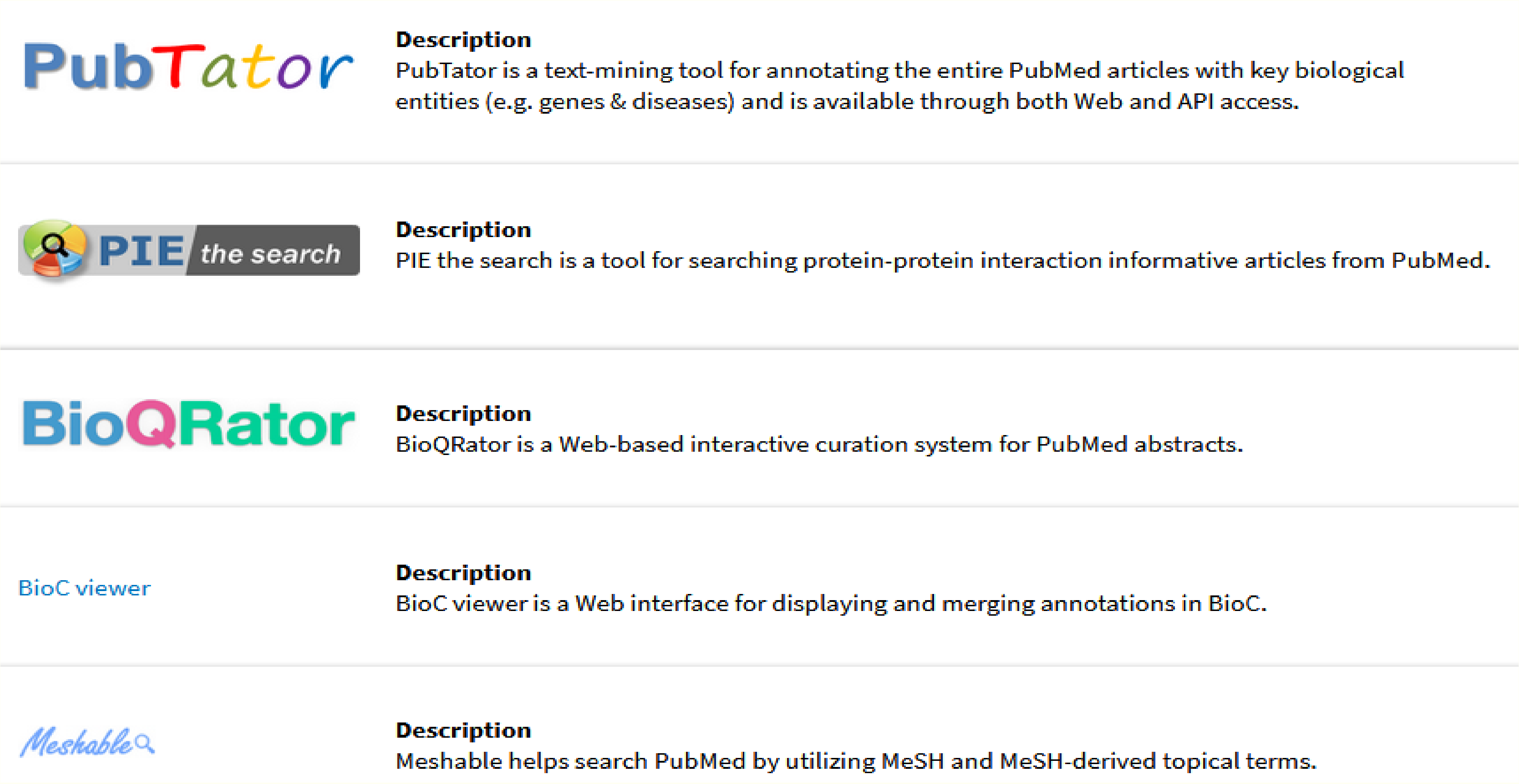
PatSeer Database Content

n+
ds
4+
INPADOC 56 million+ Families and Legal Status
US Reassignments and Maintenance Data
PDFs, Mosaic, Calculated Fields and more..
Value Added Content
Corporate Tree for top 3000 patent holding Corporate Groups
JP, KR, CN, FR, DE, DK, FI, RU, BE, NL, LU, TW
Normalized Assignee Names for Top 3000 Companies
61 million+ Full Text Records from 43 Patent Offices (US,EP, WO, DE, FR, GB,JP,KR,CN,ES, CA,CH,AT,AU,IN,BR,TH,RU, PH, SE , NO,DK,FI,BE,NL,LU , MX,AP, CO, DD, EA, IL, MA, MC, OA, TW, TJ, KG, AM, UZ, MD, GE and PT)
PatSeer
Semantic Index of Conceptually Related Terms for Search Assistance

PatSeer Pro Edition

A new era in web based professional patent analysis begins…

Medical Text Mining



PubTator

Description
PubTator is a text-mining tool for annotating the entire PubMed articles with key biological entities (e.g. genes & diseases) and is available through both Web and API access.

PIE the search

Description
PIE the search is a tool for searching protein-protein interaction informative articles from PubMed.

BioQRator

Description
BioQRator is a Web-based interactive curation system for PubMed abstracts.

BioC viewer

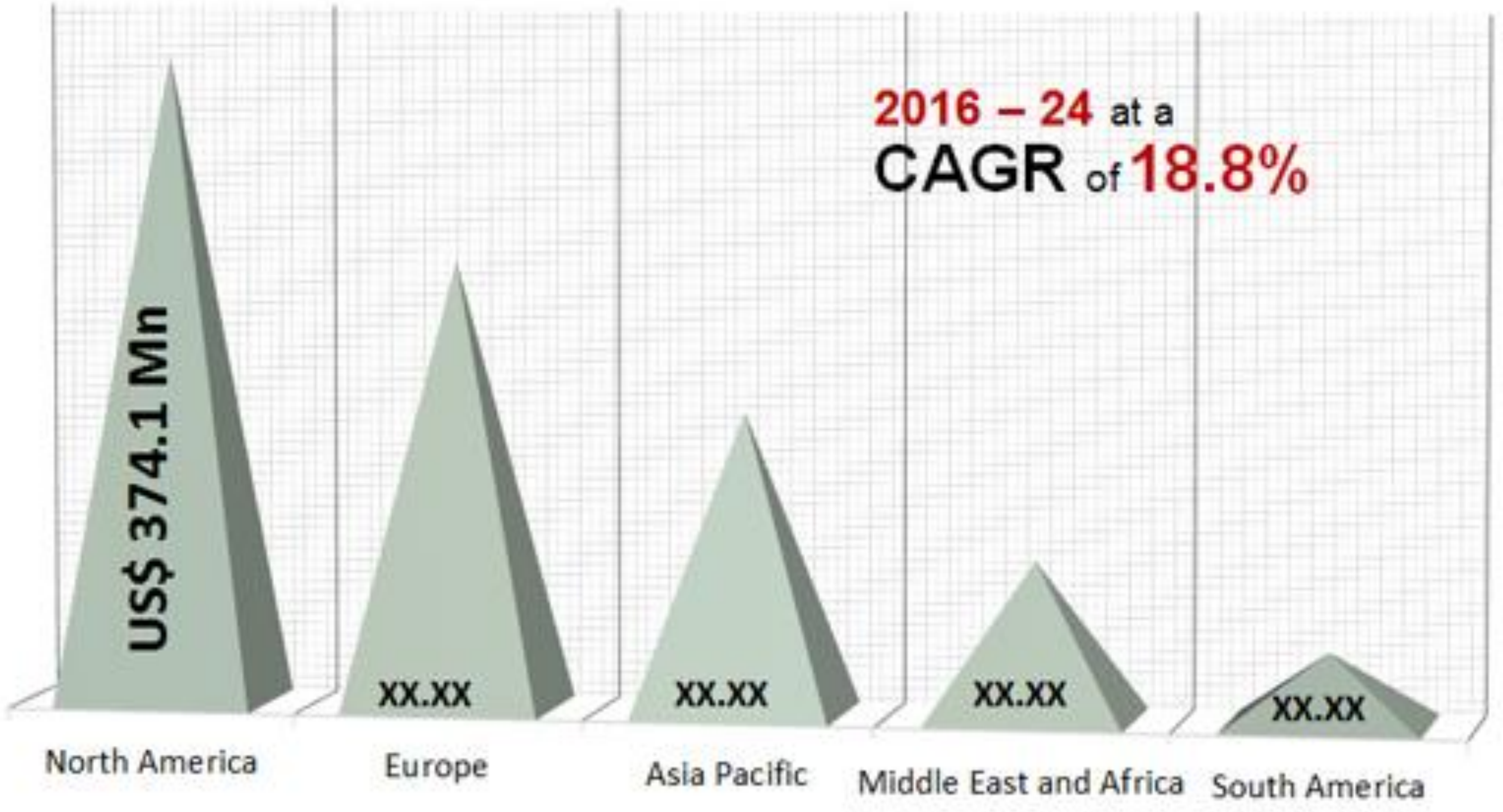
Description
BioC viewer is a Web interface for displaying and merging annotations in BioC.

Meshable

Description
Meshable helps search PubMed by utilizing MeSH and MeSH-derived topical terms.

Global Healthcare Natural Language Processing (NLP) Market

By Geography, 2015 (US$ Mn)



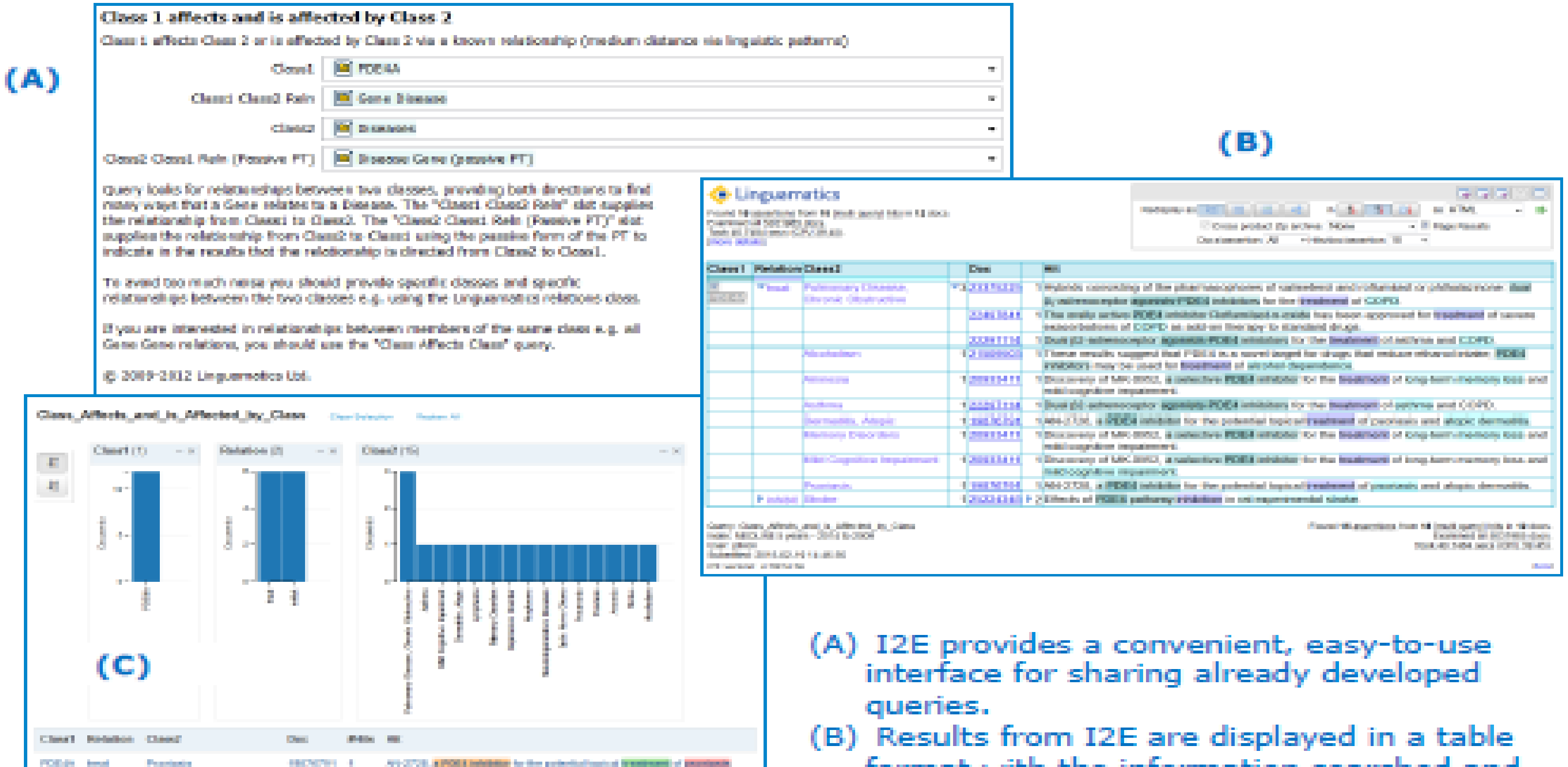
2016 – 24 at a CAGR of 18.8%

Drug Discovery

Linguamatics

Early to Late Stage Drug Discovery

Using text mining we can discover potential new drug repurposing targets or MOAs. Utilizing queries that come standard with I2E we are easily able to use publically available ontologies to identify new targets and then display them in a cluster table format or in a bar chart.



Class 1 affects and is affected by Class 2

Class 1 affects Class 2 or is affected by Class 2 via a known relationship (medium distance via linguistic patterns)

Class1	PDE4A
Class1 Class2 Reln	Gene Disease
Class2	Diseases
Class2 Class1 Reln (Passive PT)	Disease Gene (passive PT)

Query looks for relationships between two classes, providing both directions to find many ways that a Gene relates to a Disease. The "Class1 Class2 Reln" slot supplies the relationship from Class1 to Class2. The "Class2 Class1 Reln (Passive PT)" slot supplies the relationship from Class2 to Class1 using the passive form of the PT to indicate in the results that the relationship is directed from Class2 to Class1.

To avoid too much noise you should provide specific classes and specific relationships between the two classes e.g. using the Linguamatics relations class.

If you are interested in relationships between members of the same class e.g. all Gene Gene relations, you should use the "Class Affects Class" query.

© 2009-2012 Linguamatics Ltd.

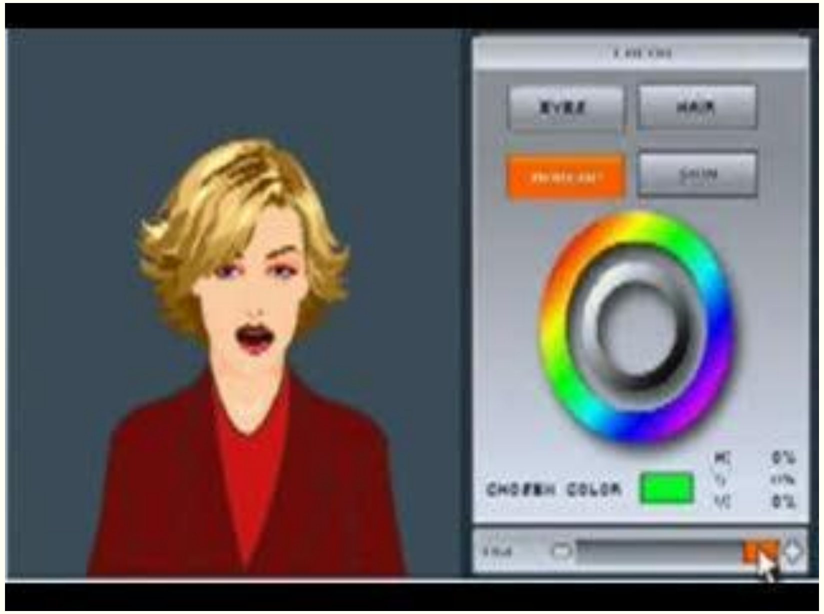
Class1	Relation	Class2	Doc	Hit
[illegible]	treat	Pulmonary Disease, Chronic Obstructive	[illegible]	Hybrids consisting of the pharmacophores of salmeterol and roflumilast or phthalazinone: dual β2-adrenoceptor agonists-PDE4 inhibitors for the treatment of COPD.
			[illegible]	The orally active PDE4 inhibitor Roflumilast N-oxide has been approved for treatment of severe exacerbations of COPD as add-on therapy to standard drugs.
			[illegible]	Dual β2-adrenoceptor agonists-PDE4 inhibitors for the treatment of asthma and COPD.
		Alcoholism	[illegible]	These results suggest that PDE4 is a novel target for drugs that reduce ethanol intake. PDE4 inhibitors may be used for treatment of alcohol dependence.
		Amnesia	[illegible]	Discovery of MK-0952, a selective PDE4 inhibitor for the treatment of long-term memory loss and mild cognitive impairment.
		Asthma	[illegible]	Dual β2-adrenoceptor agonists-PDE4 inhibitors for the treatment of asthma and COPD.
		Dermatitis, Atopic	[illegible]	AN-2728, a PDE4 inhibitor for the potential topical treatment of psoriasis and atopic dermatitis.
		Memory Disorders	[illegible]	Discovery of MK-0952, a selective PDE4 inhibitor for the treatment of long-term memory loss and mild cognitive impairment.
		Mild Cognitive Impairment	[illegible]	Discovery of MK-0952, a selective PDE4 inhibitor for the treatment of long-term memory loss and mild cognitive impairment.
		Psoriasis	[illegible]	AN-2728, a PDE4 inhibitor for the potential topical treatment of psoriasis and atopic dermatitis.
	inhibit	Stroke	[illegible]	Effects of PDE4 pathway inhibition in rat experimental stroke.

(A) I2E provides a convenient, easy-to-use interface for sharing already developed queries.

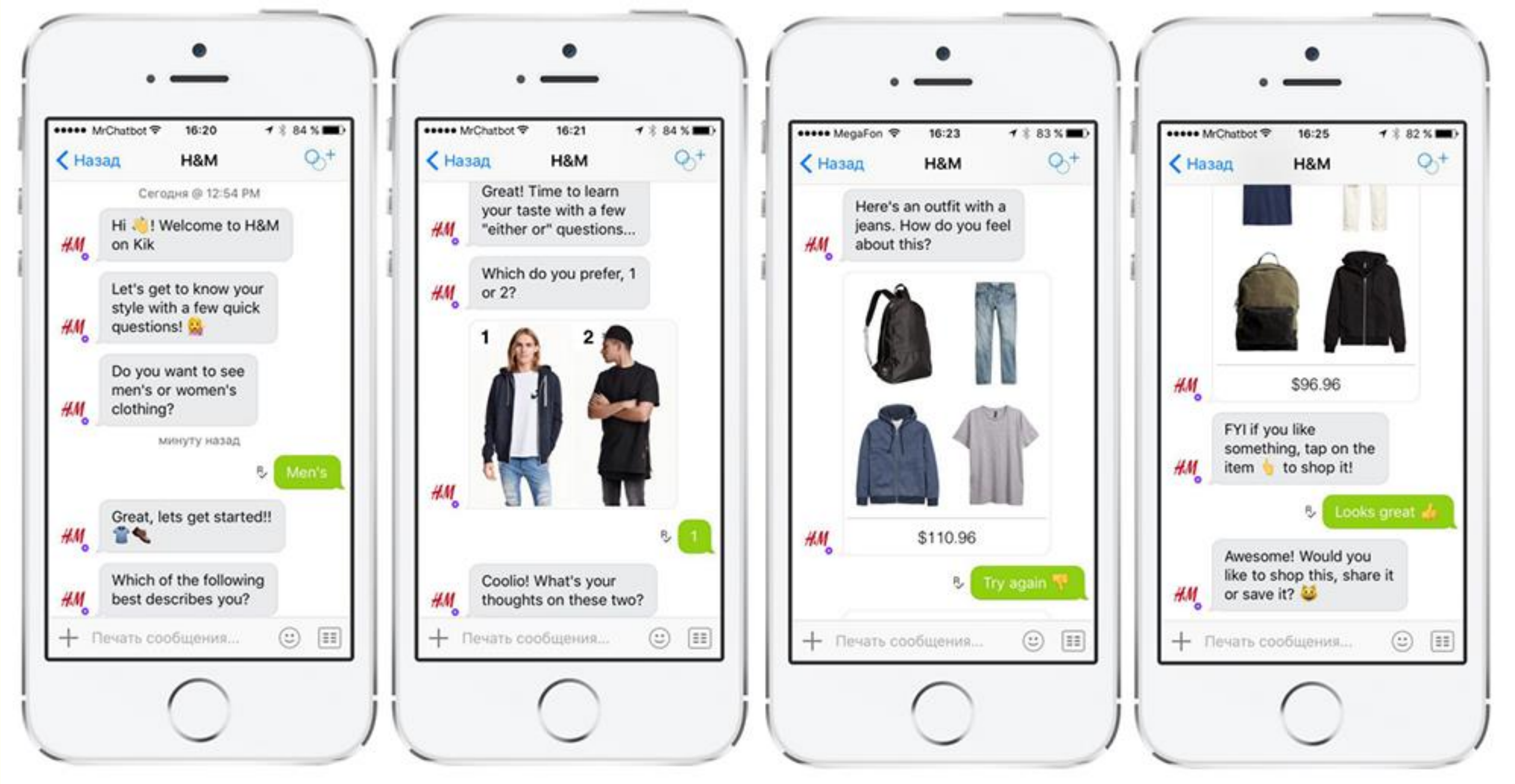
(B) Results from I2E are displayed in a table format with the information searched and

Chatbot

- E-commerce: AINI



- a chatterbot integrated with 3D animated agent character
- Improve customer services
- Reduce customer reliance on human operator



Сегодня @ 12:54 PM

Hi 👋! Welcome to H&M on Kik

Let's get to know your style with a few quick questions! 💁

Do you want to see men's or women's clothing?

минуту назад

Men's

Great, lets get started!! 👕👞

Which of the following best describes you?

Great! Time to learn your taste with a few "either or" questions...

Which do you prefer, 1 or 2?

1

Coolio! What's your thoughts on these two?

Here's an outfit with a jeans. How do you feel about this?

$110.96

Try again 👎

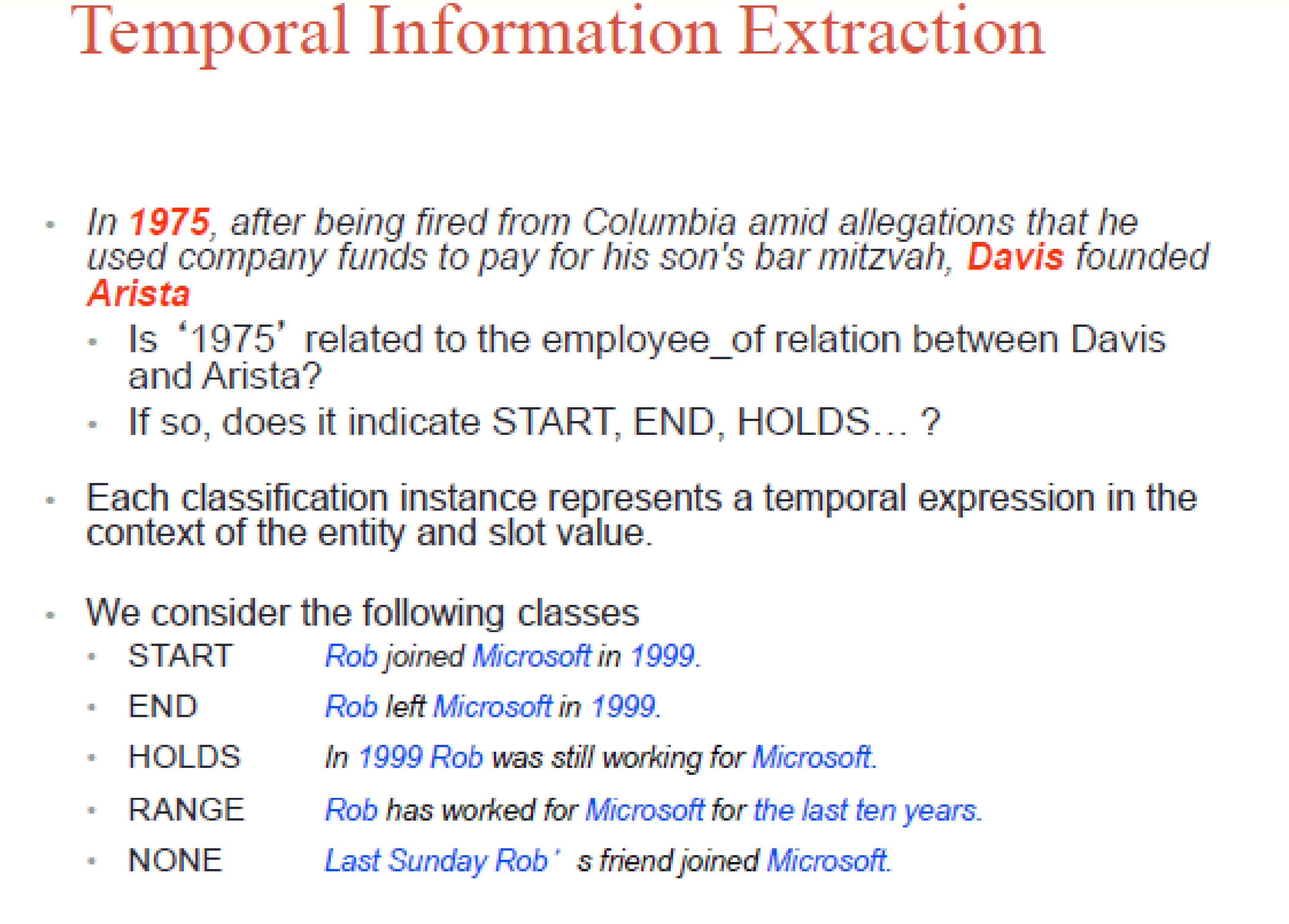
$96.96

FYI if you like something, tap on the item 👆 to shop it!

Looks great 👍

Awesome! Would you like to shop this, share it or save it? 😸

r



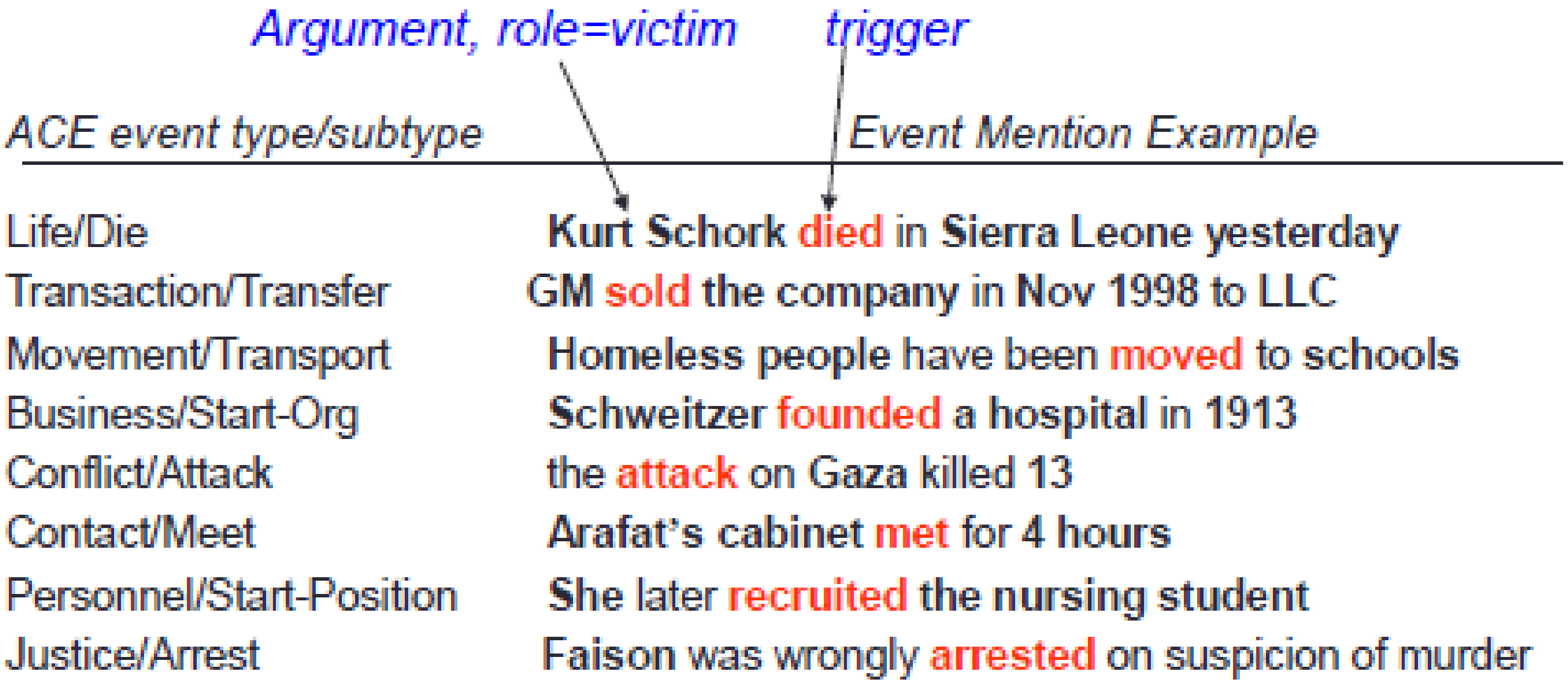
Temporal Information Extraction

- *In **1975**, after being fired from Columbia amid allegations that he used company funds to pay for his son's bar mitzvah, **Davis** founded **Arista***
  - Is ‘1975’ related to the employee_of relation between Davis and Arista?
  - If so, does it indicate START, END, HOLDS… ?
- Each classification instance represents a temporal expression in the context of the entity and slot value.
- We consider the following classes

START	*Rob joined Microsoft in 1999.*
END	*Rob left Microsoft in 1999.*
HOLDS	*In 1999 Rob was still working for Microsoft.*
RANGE	*Rob has worked for Microsoft for the last ten years.*
NONE	*Last Sunday Rob' s friend joined Microsoft.*

Event Extraction

- An event is specific occurrence that implies a change of states
- **event trigger**: the main word which most clearly expresses an event occurrence
- **event arguments**: the mentions that are involved in an event (participants)
- **event mention**: a phrase or sentence within which an event is described, including trigger and arguments
- ACE defined 8 types of events, with 33 subtypes

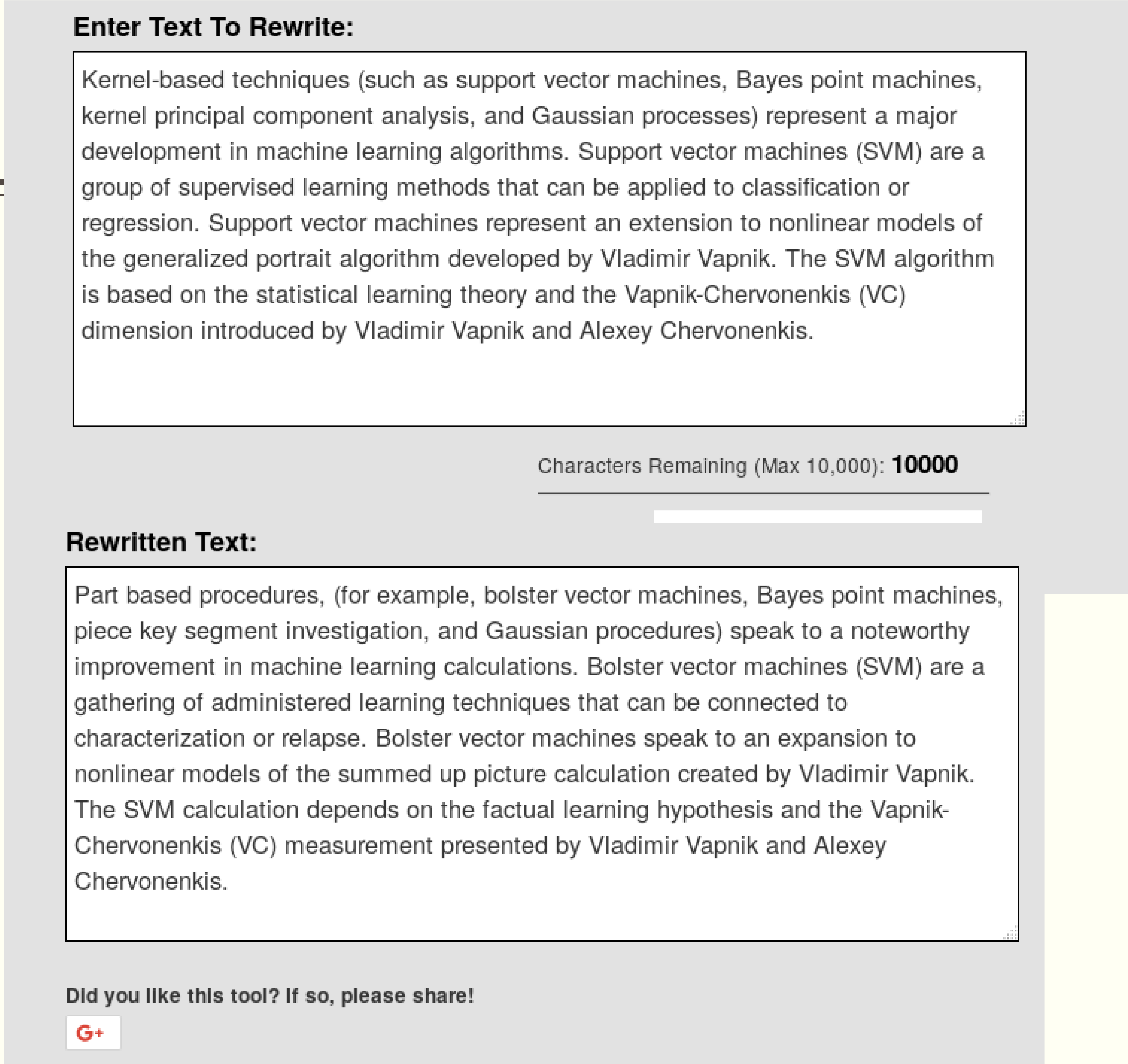


ACE event type/subtype	*Event Mention Example*
Life/Die	Kurt Schork died in Sierra Leone yesterday
Transaction/Transfer	GM sold the company in Nov 1998 to LLC
Movement/Transport	Homeless people have been moved to schools
Business/Start-Org	Schweitzer founded a hospital in 1913
Conflict/Attack	the attack on Gaza killed 13
Contact/Meet	Arafat's cabinet met for 4 hours
Personnel/Start-Position	She later recruited the nursing student
Justice/Arrest	Faison was wrongly arrested on suspicion of murder

Spin Bot

Spinbot

Article Spinning, Text Rewriting, Content Creation Tool.



Enter Text To Rewrite:

Kernel-based techniques (such as support vector machines, Bayes point machines, kernel principal component analysis, and Gaussian processes) represent a major development in machine learning algorithms. Support vector machines (SVM) are a group of supervised learning methods that can be applied to classification or regression. Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik. The SVM algorithm is based on the statistical learning theory and the Vapnik-Chervonenkis (VC) dimension introduced by Vladimir Vapnik and Alexey Chervonenkis.

Characters Remaining (Max 10,000): **10000**

Rewritten Text:

Part based procedures, (for example, bolster vector machines, Bayes point machines, piece key segment investigation, and Gaussian procedures) speak to a noteworthy improvement in machine learning calculations. Bolster vector machines (SVM) are a gathering of administered learning techniques that can be connected to characterization or relapse. Bolster vector machines speak to an expansion to nonlinear models of the summed up picture calculation created by Vladimir Vapnik. The SVM calculation depends on the factual learning hypothesis and the Vapnik-Chervonenkis (VC) measurement presented by Vladimir Vapnik and Alexey Chervonenkis.

Did you like this tool? If so, please share!

- https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-863j-natural-language-and-the-computer-representation-of-knowledge-spring-2003/lecture-notes/

- http://cs224d.stanford.edu/syllabus.html

- https://github.com/oxford-cs-deepnlp-2017/lectures

- https://www.technologyreview.com/s/538616/google-deepmind-teaches-artificial-intelligence-machines-to-read/

- https://www.fastcompany.com/3026423/why-google-is-investing-in-deep-learning

AMBIGUITY IN NATURAL LANGUAGES

Writing systems without word boundaries

- do not have explicit, systematic visible markers to distinguish the ending of one word and the beginning of another.
  - Burmese
  - Chinese
  - Japanese etc.

SVM, maximum entropy , CRF, EM

Sentence Boundary Markers

- In some contexts period is not sentence boundary marker
  - 27.5, etc., google.com, Mr.

  - The President lives in Washington D.C. He likes that place.
  - *Is K.H. Smith here?*
  - I *bought the apples, pears, lemons, etc. Did you eat them?*

  - CRF, HMM, Maximum Entropy, SMO, Rule based Systems

Ambiguity at POS level

The word “Like” can play different roles in a given sentence.

- Verb
- Noun
- Adjective
- Conjunction
- Adverb
- Preposition
- Interjection

- Project
  - Noun
  - Verb

- Can
  - Noun
  - Verb

CRF, HMM, Maximum Entropy, SMO, Rule based Systems

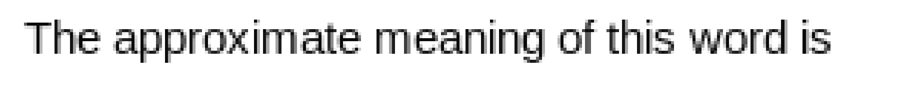
Differences in Morphology

English	Indian Languages (Dravidian Languages)
Weakly Inflectional	Agglutinative
A word may contain one or two root words	A word may contain more than two root words
Morpho-phonology is trivial when compared to Dravidian languages	Morpho-phonology is very complex because of sandhi rules

Morphological Ambiguity

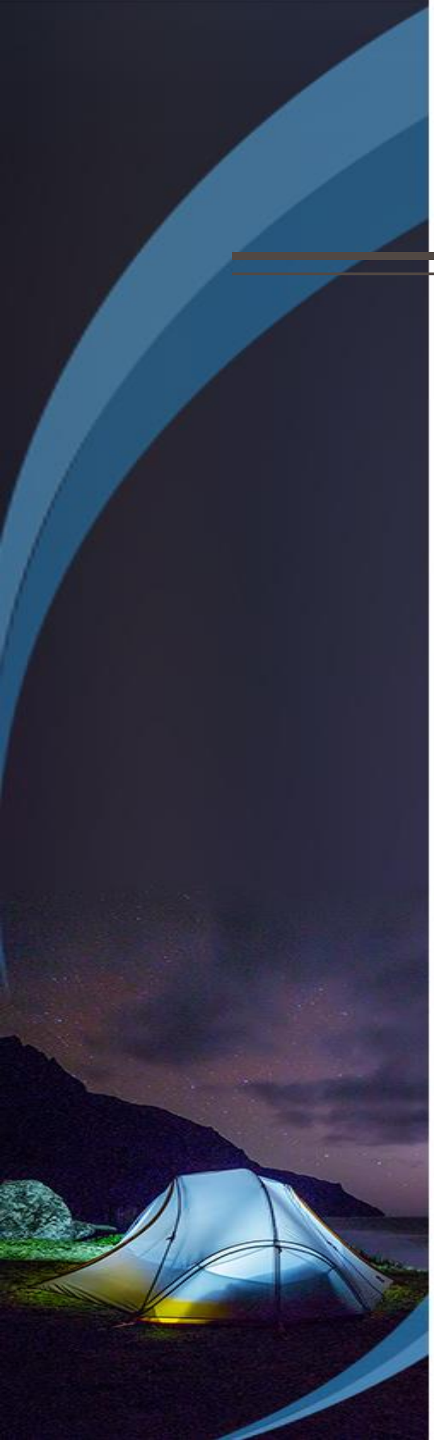
the world’s longest word

निरन्तरान्धकारितादिगन्तरकन्दलदमन्दसुधारसबिन्दुसान्द्रतरघनाघनवृन्दसन्देहकरस्यन्दमानमकरन्दबिन्दुबन्धुरतरमाकन्दतरुकुलतल्पकल्पमृदुलसिकताजालजटिलमूलतलमरुवकमिलदलघुलघुलयकलितरमणीयपानीयशालिकाबालिकाकरारविन्दगलन्तिकागलदेलालवङ्गपाटलघनसारकस्तूरिकातिसौरभमेदुरलघुतरमधुरशीतलतरसलिलधारानिराकरिष्णुतदीयविमलविलोचनमयूखरेखापसारितपिपासायासपथिकलोकान्


The approximate meaning of this word is

- In it, the distress, caused by thirst, to travellers, was alleviated by clusters of rays of the bright eyes of the girls; the rays that were shaming the currents of light, sweet and cold water charged with the strong fragrance of cardamom, clove, saffron, camphor and musk and flowing out of the pitchers (held in) the lotus-like hands of maidens (seated in) the beautiful water-sheds, made of the thick roots of vetiver mixed with marjoram, (and built near) the foot, covered with heaps of couch-like soft sand, of the clusters of newly sprouting mango trees, which constantly darkened the intermediate space of the quarters, and which looked all the more charming on account of the trickling drops of the floral juice, which thus caused the delusion of a row of thick rainy clouds, densely filled with abundant nectar

http://www.hitxp.com/articles/literature/world-longest-word-language-guinness-record/



Morphology

Application of extensive Sandhi changes sometimes results in telescoping of several words into long strings.

English Sentence: Do you say that there is no hot water?’

Telugu Sentence: vEdinILLu levu aNtavu A?

After Sandhi: vENNILLEvaNtAvA (one word)

- Reference: P. Bhaskara Rao, “Telugu” , Concise Encyclopedia of Languages of the world, Elsevier, pp 1055-1060.

Syntax

- Syntax concerns the proper ordering of words and its affect on meaning.
    - The dog bit the boy.
    - The boy bit the dog.
    - * Bit boy dog the the.
    - Colorless green ideas sleep furiously.
    - The rat killed the snake and swallowed it.

Ambiguity at Syntax Level

- Ambiguities compound to generate enormous numbers of possible interpretations.
- In English, a sentence ending in *n* prepositional phrases has *over* many syntactic interpretations (cf. Catalan numbers).

$$C_n = (2n)!/[(n+1)!n!]$$

  - "I saw the man with the telescope": 2 parses
  - "I saw the man on the hill with the telescope.": 5 parses
  - "I saw the man on the hill in Texas with the telescope": 14 parses
  - "I saw the man on the hill in Texas with the telescope at noon.": 42 parses
  - "I saw the man on the hill in Texas with the telescope at noon on Monday" 132 parses

Garden path Sentences

- The prime number few.
- Fat people eat accumulates.
- The cotton clothing is usually made of grows in Mississippi.
- Until the police arrest the drug dealers control the street.
- The man who hunts ducks out on weekends.
- When Fred eats food gets thrown.
- Mary gave the child the dog bit a bandaid.
- The girl told the story cried.
- I convinced her children are noisy.
- The horse raced past the barn fell.
- I know the words to that song about the queen don't rhyme.
- She told me a little white lie will come back to haunt me.
- The dog that I had really loved bones.
- That Jill is never here hurts.
- The man who whistles tunes pianos.
- The old man the boat.
- The raft floated down the river sank.
- We painted the wall with cracks.

http://www.fun-with-words.com/ambiguous_garden_path.html

Different Sentence Structures

- SOV : Korean, Mongolian, Turkish, the Indo-Aryan languages and the Dravidian languages "She bread ate"
- SVO : English, Bulgarian, Macedonian, Serbo-Croatian, Chinese "She ate bread."
- VSO : Classical Arabic, and Hawaiian. "Ate she bread"
- VOS : Fijian and Malagasy. "Ate bread she."
- OVS : Hixkaryana. "Bread ate she."
- OSV : Xavante and Warao. "Bread she ate.“
- No Strict Word Order: Latin, Greek, Persian, Romanian, Assyrian, Turkish and Finnish

Idioms

- kick the bucket
  - To die

- Beat around the bush
  - Avoid saying what you mean, usually because it is uncomfortable

- Break a leg
  - Good luck

- Pull yourself together
  - Calm down

- The early bird gets the worm
  - The first people who arrive will get the best stuff

Word Sense Ambiguity -- Polysemy

- the condition of a person or thing, as with respect to circumstances or attributes: a state of health.
- the condition of matter with respect to structure, form, constitution, phase, or the like: water in a gaseous state.
- status, rank, or position in life; station: He dresses in a manner befitting his state.
- the style of living befitting a person of wealth and high rank: to travel in state.
- a particular condition of mind or feeling: to be in an excited state.
- an abnormally tense, nervous, or perturbed condition: He's been in a state since hearing about his brother's death.
- a politically unified people occupying a definite territory; nation.

Named Entity Recognition

- Identifies and classifies strings of characters representing proper nouns

  [PER Neil A. Armstrong] , the 38-year-old civilian commander, radioed to earth and the mission control room here: "**[LOC Houston]** , **[ORG Tranquility]** Base here; the Eagle has landed."

- Useful for filtering documents
  - "I need to find news articles about organizations in which Bill Gates might be involved…"
- Disambiguate tokens: "Chicago" (team) vs. "Chicago" (city)
- Source of abstract features
  - E.g. "Verbs that appear with entities that are Organizations"
  - E.g. "Documents that have a high proportion of Organizations"

Coreference Resolution

- Identify all phrases that refer to each entity of interest – i.e., group mentions of concepts

  [Neil A. Armstrong] , **[the 38-year-old civilian commander]**, radioed to **[earth]**. **[He]** said the famous words, "**[the Eagle]** has landed"."

- The Named Entity recognizer only gets us part-way…
- …if we ask, "what actions did Neil Armstrong perform?", we will miss many instances (e.g. "He said…")
- Coreference resolver abstracts over different ways of referring to the same person
  - Useful in information extraction