Natural Language Processing

Complex AI Systems

Asmaa Elbadrawy
PhD, Lecturer
IFT Program, ASU

Arizona State University

Humans use language to represent knowledge (ideas) and communicate them to others.

Human language is called Natural Language (English, Spanish, etc.) as opposed to computer languages, which are called Formal Languages (first order logic, programming, etc.)

Why Computers need to process Natural Languages

- Communicate with humans
- Build KB’s from human writings (learn)
- Advance scientific understanding of language and how humans use then.

Natural Language Processing (NLP) has many applications.

Classifying news articles and web pages to facilitate searching.

Features in document classification

Category: psychology	Category: nutrition	Category: sports
Depression	Calories	Reps
Anxiety	Protein	Tension
Syndrome	Diet	Overload
Mental	Fat	Exercise
Thought	Carbs	Abs

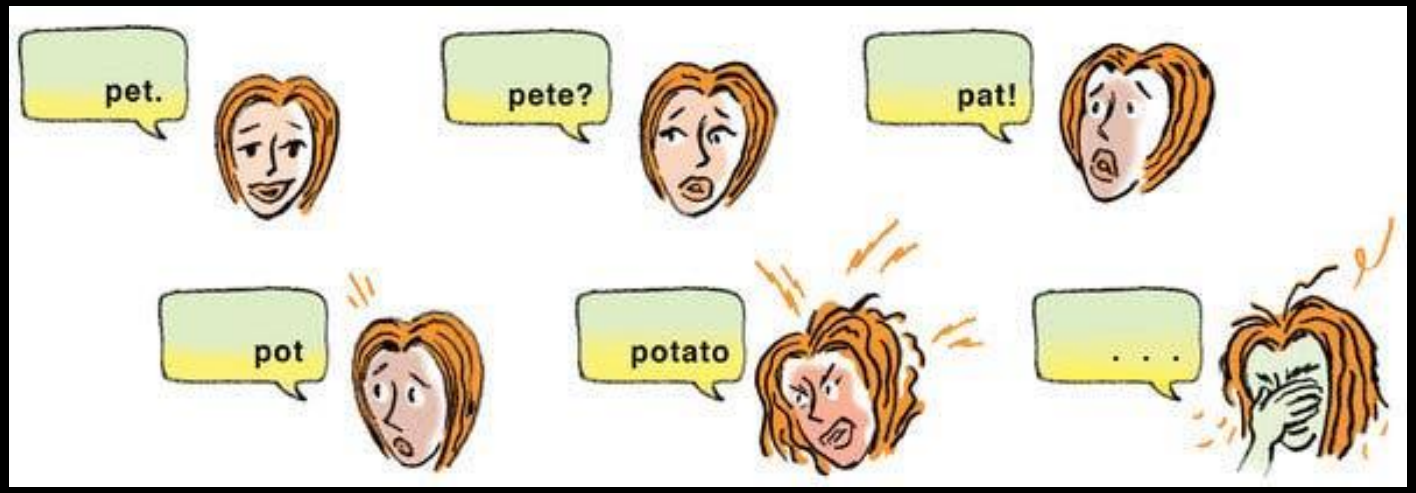
altexsoft
software r&d engineering

Source: https://www.altexsoft.com/blog/document-classific ation/

Auto-Translation.

Source: https://www.nytimes.com/2013/05/02/technology/personaltech/the-utility-and-drawbacks-of-translation-apps.html

Suggesting spelling & grammar corrections.



Source: https://www.nytimes.com/2015/01/09/style/when-autocorrect-goes-wrong-and-so-so-right.html

Suggesting spelling & grammar corrections.

We need a language model that can tell which word is likely to come next in a sentence.

Language Models:
Bag-of-Words

Represent a set of documents using the collection (bag) of words contained in all of them.

• Document represented as a set of words	Class Label
1. Stocks rallied on Monday, with major indexes gaining 1% as optimism persisted over the first quarter earnings season.	Business
2. Heavy rain continued to pound much of the east coast on Monday, with flood warnings issued in New York City and other locations.	Weather

- **The words appearing in each document represent its features!**
- **Build a dataset in which each document has a class label.**
- **Train a model (DT, Deep Learning, etc.) to classify a document based on its words.**

Language Models:
N-Grams

While some words can be common among different subjects, some phrases (that consist of n consecutive words) are more specific to one subject.

Quarter

This is a common word in economics and in sports!

We can use a sequence of 4 words instead of a single word

- **First quarter earnings report**

  More specific to business

- **Fourth quarter touchdown passes**

  More specific to sports

This 4-word sequence is called an n-gram (4-gram).

The n-gram model can tell us if a phrase/short sentence is likely a correct English phrase.

- **A black cat.**

  Is likely proper English because it appears in many training examples.

- **Black cat a.**

  Is likely not proper English because it appears in zero training examples.

Language Models:
Part-of-Speech Tagging

A native English speaker can directly say that “a black cat” is sound because it follows a familiar pattern (article-adjective-noun) but “Black cat a” does not follow any familiar pattern.

- **A black cat.**
- **Black cat a.**

If we can tag words with their POS, then define sound sentence patterns, this can help us identify proper sentences in a more generic way without having to have an example of each possible sentence in our data corpus.

- **A black cat.**
- **Black cat a.**

This is similar to building a KB of language rules. More sophisticated language models consider the grammar rules.

POS Tags

Tag	Word	Description	Tag	Word	Description
CC	*and*	Coordinating conjunction	PRP$	*your*	Possessive pronoun
CD	*three*	Cardinal number	RB	*quickly*	Adverb
DT	*the*	Determiner	RBR	*quicker*	Adverb, comparative
EX	*there*	Existential there	RBS	*quickest*	Adverb, superlative
FW	*per se*	Foreign word	RP	*off*	Particle
IN	*of*	Preposition	SYM	+	Symbol
JJ	*purple*	Adjective	TO	*to*	to
JJR	*better*	Adjective, comparative	UH	*eureka*	Interjection
JJS	*best*	Adjective, superlative	VB	*talk*	Verb, base form
LS	*1*	List item marker	VBD	*talked*	Verb, past tense
MD	*should*	Modal	VBG	*talking*	Verb, gerund
NN	*kitten*	Noun, singular or mass	VBN	*talked*	Verb, past participle
NNS	*kittens*	Noun, plural	VBP	*talk*	Verb, non-3rd-sing
NNP	*Ali*	Proper noun, singular	VBZ	*talks*	Verb, 3rd-sing
NNPS	*Fords*	Proper noun, plural	WDT	*which*	Wh-determiner
PDT	*all*	Predeterminer	WP	*who*	Wh-pronoun
POS	*'s*	Possessive ending	WP$	*whose*	Possessive wh-pronoun
PRP	*you*	Personal pronoun	WRB	*where*	Wh-adverb
$	$	Dollar sign	#	#	Pound sign
“	‘	Left quote	”	’	Right quote
(	[	Left parenthesis	)	]	Right parenthesis
,	,	Comma	.	!	Sentence end
:	;	Mid-sentence punctuation			

Fig 23.1, Russell & Norvig’s Textbook

Considering POS and familiar sentence patterns is similar to building a KB of language rules. More sophisticated language models consider grammar rules.

This approach can facilitate tasks such as auto translation.

English <adjective then noun>
Spanish <noun then adjective>