

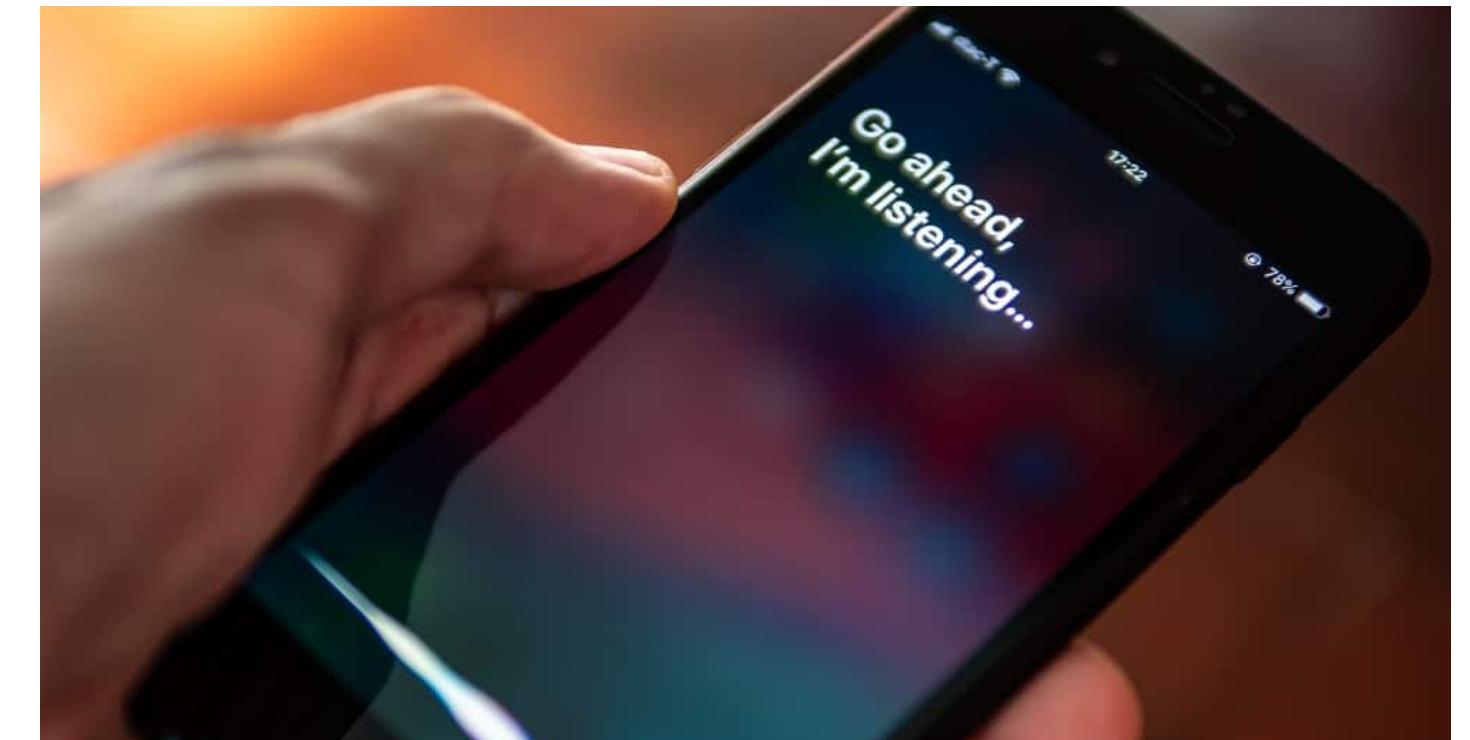
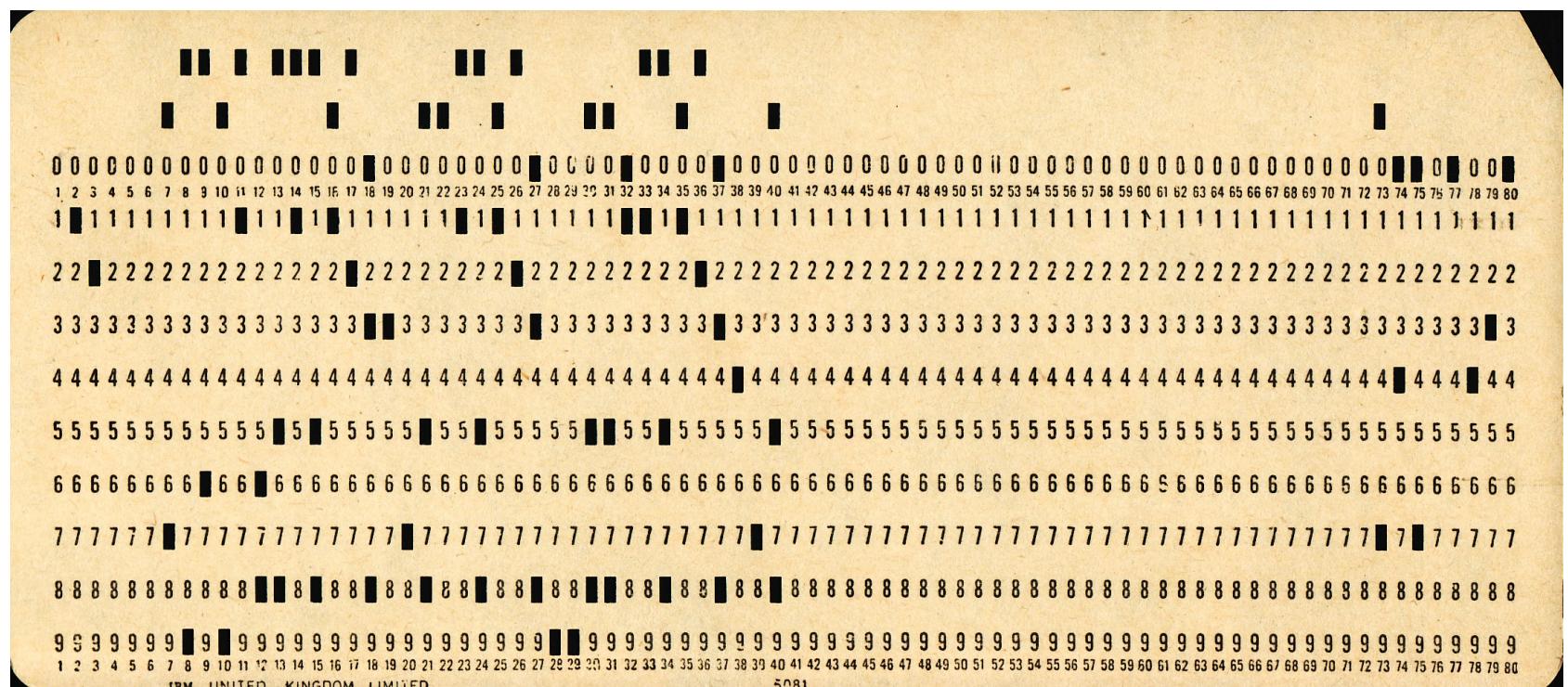
Processamento de Linguagem Natural

Introdução

Yuri Malheiros (yuri@ci.ufpb.br)

Introdução

- # • Comunicação com máquinas



Introdução

- Uma inteligência artificial pode interagir com um ser humano através de:
 - Reconhecimento de fala
 - Compreensão de linguagem
 - Geração de linguagem
 - Sintetização de fala
 - Recuperação de informação
 - Inferência
 - Etc.

Introdução

- Resolver esses problemas é o principal objetivo do processamento de linguagem natural
- Muitos avanços aconteceram nos últimos anos para solucionar esses problemas, mas muitos desafios ainda existem

Respondendo Perguntas

- 2011 o IBM Watson venceu o Jeopardy



- O que significa convergir?
- Em que ano Abraham Lincoln nasceu?
- Quantos estados existiam no Estados Unidos naquele ano?
- Qual a quantidade de seda exportada para Inglaterra pela China no século 18?
- O que os cientistas pensam sobre a clonagem humana?

Tradução

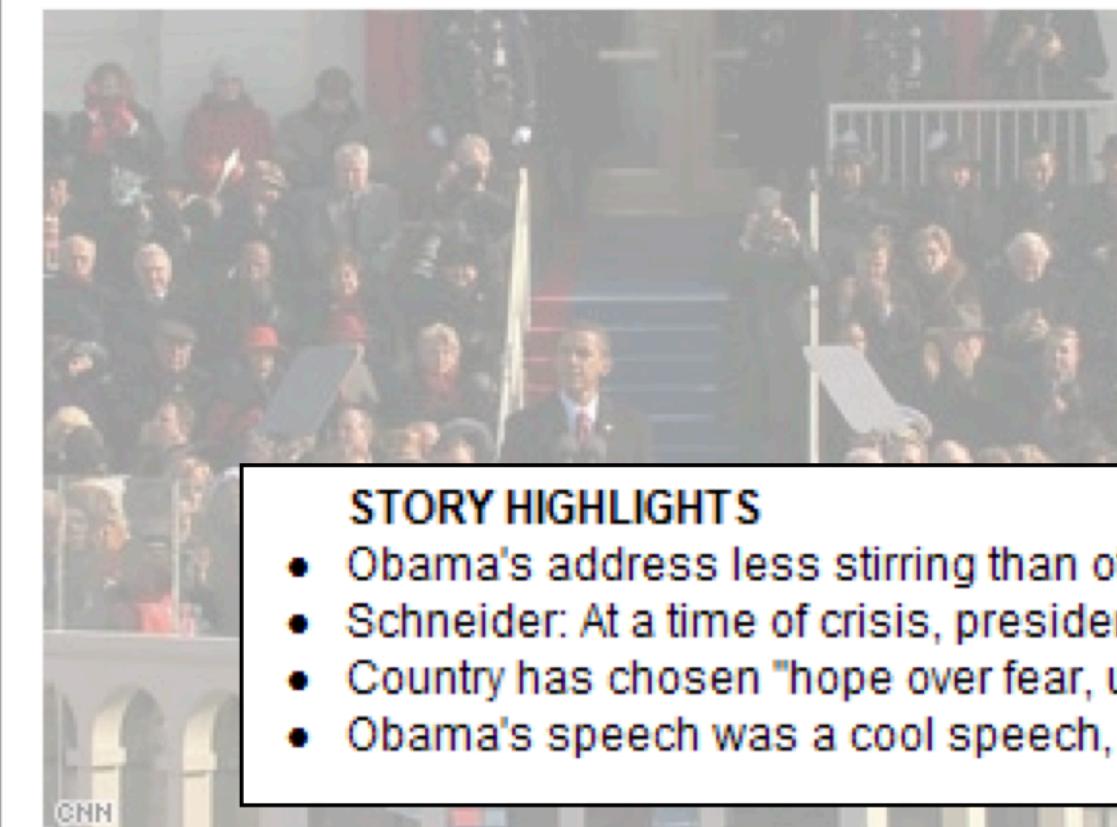
- O Google Translate conseguir traduzir hoje textos em 132 idiomas

Detect language	Danish	Hmong	Lingala	Portuguese	Tatar
Afrikaans	Dhivehi	Hungarian	Lithuanian	Punjabi	Telugu
Albanian	Dogri	Icelandic	Luganda	Quechua	Thai
Amharic	Dutch	Igbo	Luxembourgish	Romanian	Tigrinya
Arabic	English	Ilocano	Macedonian	Russian	Tsonga
Armenian	Esperanto	Indonesian	Maithili	Samoan	Turkish
Assamese	Estonian	Irish	Malagasy	Sanskrit	Turkmen
Aymara	Ewe	Italian	Malay	Scots Gaelic	Twi
Azerbaijani	Filipino	Japanese	Malayalam	Sepedi	Ukrainian
Bambara	Finnish	Javanese	Maltese	Serbian	Urdu
Basque	French	Kannada	Maori	Sesotho	Uyghur
Belarusian	Frisian	Kazakh	Marathi	Shona	Uzbek
Bengali	Galician	Khmer	Meiteilon (Manipuri)	Sindhi	Vietnamese
Bhojpuri	Georgian	Kinyarwanda	Mizo	Sinhala	Welsh
Bosnian	German	Konkani	Mongolian	Slovak	Xhosa
Bulgarian	Greek	Korean	Myanmar (Burmese)	Slovenian	Yiddish
Catalan	Guarani	Krio	Nepali	Somali	Yoruba
Cebuano	Gujarati	Kurdish (Kurmanji)	Norwegian	Spanish	Zulu
Chichewa	Haitian Creole	Kurdish (Sorani)	Odia (Oriya)	Sundanese	
Chinese	Hausa	Kyrgyz	Oromo	Swahili	
Corsican	Hawaiian	Lao	Pashto	Swedish	
Croatian	Hebrew	Latin	Persian	Tajik	
Czech	Hindi	Latvian	Polish	Tamil	

Sumarização

- Condensar documentos

WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.



CNN
President Obama renewed his call for a massive plan to stimulate economic growth.
[more photos »](#)

STORY HIGHLIGHTS

- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must begin aid in

his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

Obama, too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

Some presidents used their inaugural address to set out a bold agenda.

Classificação de Documentos



Tecnologias

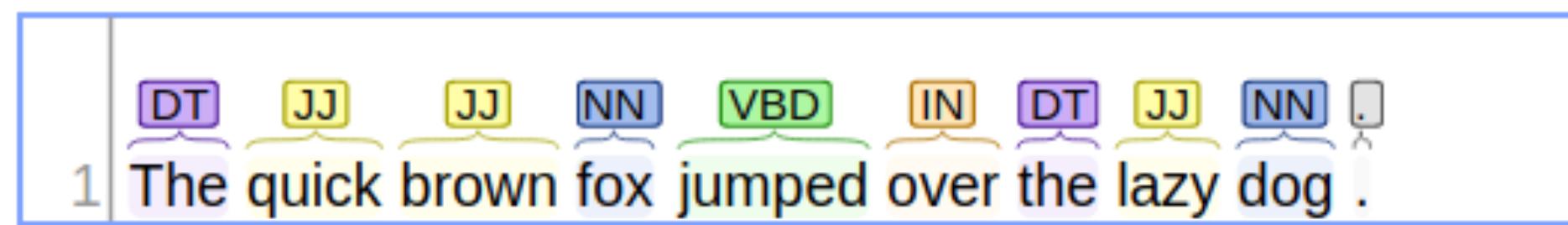
- Modelo de linguagem
- Part-of-speech tagging
- Parser sintático
- Reconhecimento de entidades nomeadas
- Resolução de correferência
- Desambiguação

Modelo de Linguagem

- Modelos que atribuem probabilidades a palavras ou sequências de palavras
 - Probabilidade de uma palavra: $P(w_1)$
 - Probabilidade de uma sequência de palavras: $P(w_1, w_2, w_3, \dots, w_n)$
 - Probabilidade da próxima palavra: $P(w_n|w_1, w_2, \dots, w_{n-1})$

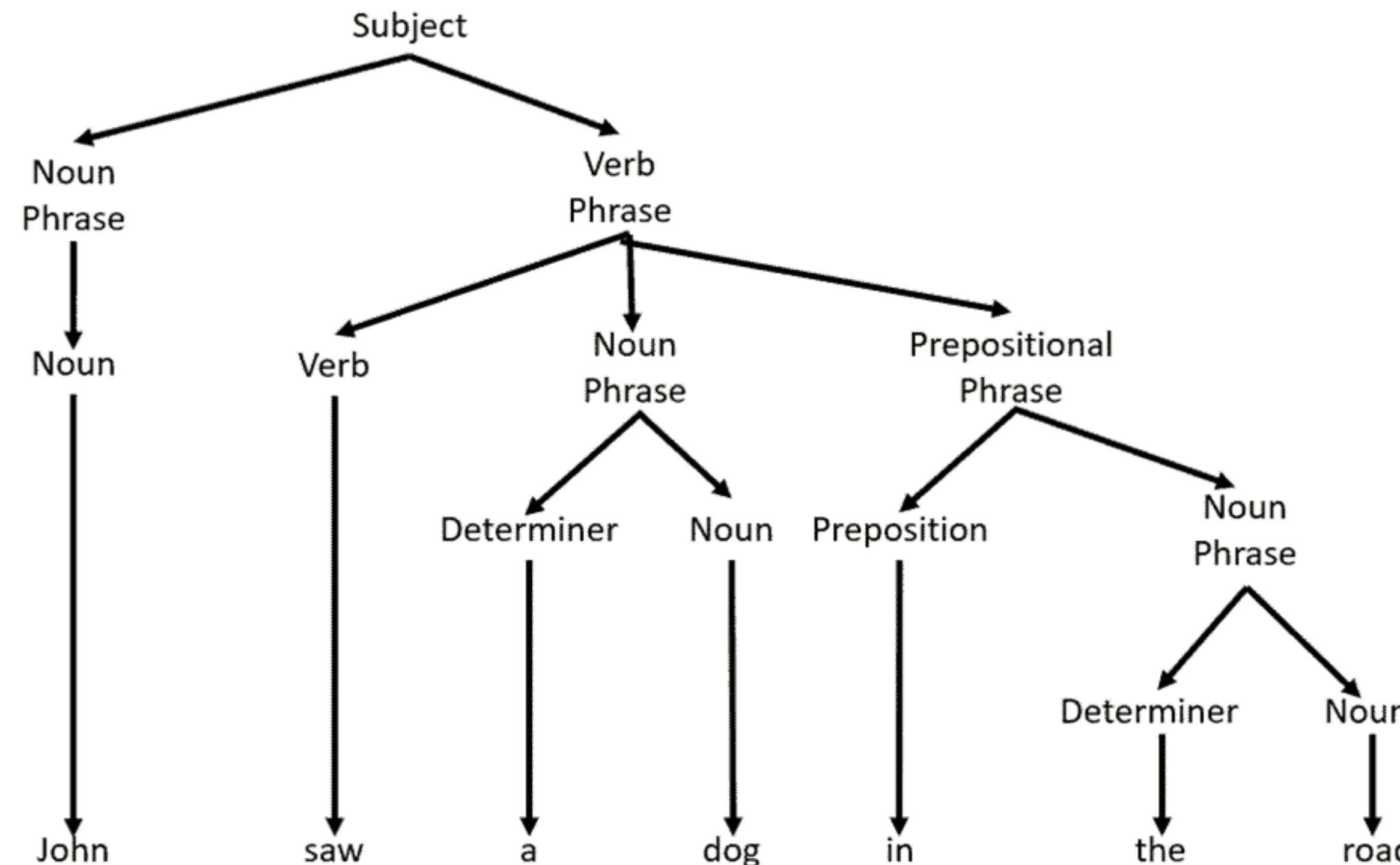
Part-of-Speech tagging

Part-of-Speech:



Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	<i>and, but, or</i>	PDT	predeterminer	<i>all, both</i>	VBP	verb non-3sg present	<i>eat</i>
CD	cardinal number	<i>one, two</i>	POS	possessive ending	's	VBZ	verb 3sg pres	<i>eats</i>
DT	determiner	<i>a, the</i>	PRP	personal pronoun	<i>I, you, he</i>	WDT	wh-determ.	<i>which, that</i>
EX	existential 'there'	<i>there</i>	PRP\$	possess. pronoun	<i>your, one's</i>	WP	wh-pronoun	<i>what, who</i>
FW	foreign word	<i>mea culpa</i>	RB	adverb	<i>quickly</i>	WP\$	wh-possess.	<i>whose</i>
IN	preposition/subordin-conj	<i>of, in, by</i>	RBR	comparative adverb	<i>faster</i>	WRB	wh-adverb	<i>how, where</i>
JJ	adjective	<i>yellow</i>	RBS	superlatv. adverb	<i>fastest</i>	\$	dollar sign	\$
JJR	comparative adj	<i>bigger</i>	RP	particle	<i>up, off</i>	#	pound sign	#
JJS	superlative adj	<i>wildest</i>	SYM	symbol	<i>+, %, &</i>	"	left quote	' or "
LS	list item marker	<i>1, 2, One</i>	TO	"to"	<i>to</i>	"	right quote	' or "
MD	modal	<i>can, should</i>	UH	interjection	<i>ah, oops</i>	(left paren	[, (, {, <
NN	sing or mass noun	<i>llama</i>	VB	verb base form	<i>eat</i>)	right paren],), }, >
NNS	noun, plural	<i>llamas</i>	VBD	verb past tense	<i>ate</i>	,	comma	,
NNP	proper noun, sing.	<i>IBM</i>	VBG	verb gerund	<i>eating</i>	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	<i>Carolinas</i>	VBN	verb past part.	<i>eaten</i>	:	sent-mid punc	: ; ... --

Parser Sintático

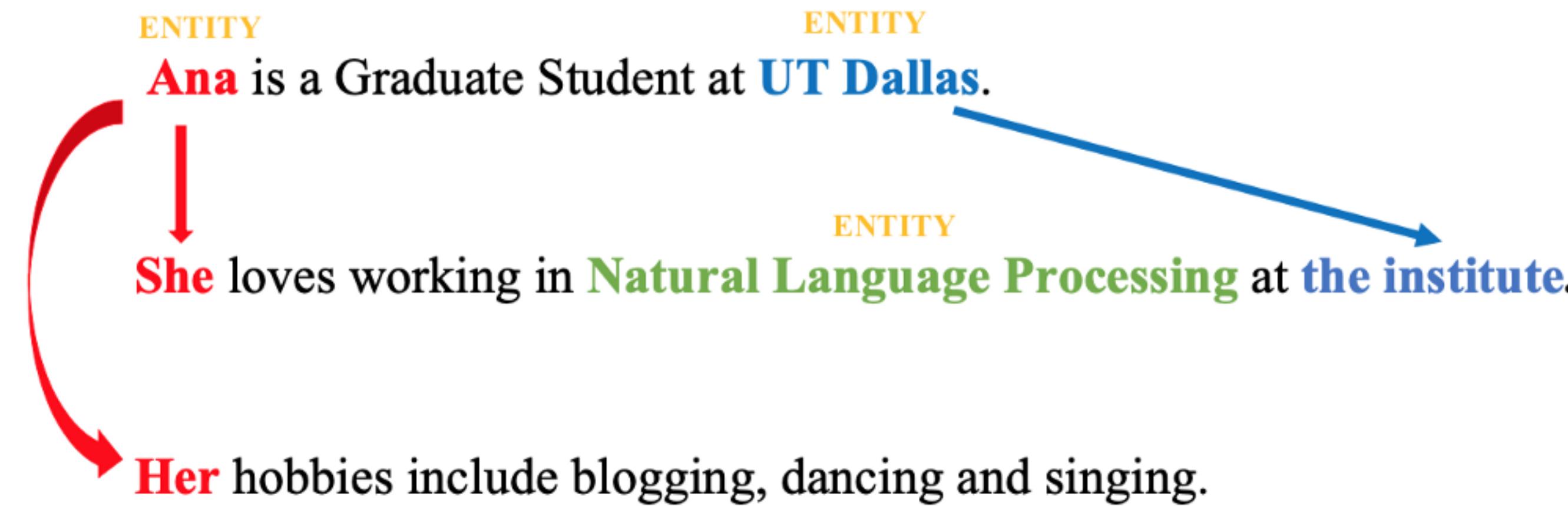


Reconhecimento de Entidades Nomeadas

Person p Loc l Org o Event e Date d Other z

Barack Hussein Obama II * (born August 4, 1961 *) is an American * attorney and politician who served as the 44th President of the United States * from January 20, 2009 *, to January 20, 2017 *. A member of the Democratic Party *, he was the first African American * to serve as president. He was previously a United States Senator * from Illinois * and a member of the Illinois State Senate *.

Resolução de Correferência



Desambiguação

- Sentido: banco (instituição financeira ou lugar para sentar?)
- Part-of-speech: jogo (substantivo ou verbo?)
- Estrutura sintática: eu posso ver um homem com um telescópio enrolado em um papel

Incertezas

- Como lidar com essas incertezas?
 - Retornar todas as possibilidades
 - Retornar as possibilidades mais prováveis (modelos probabilísticos)
 - Essa abordagem só é boa se as probabilidades forem precisas
 - Como calculamos essas probabilidades?

Corpora

- Um corpus é uma coleção de texto
 - Podem ser anotados com informações extras ou não
- Exemplos:
 - Penn Treebank: 1 milhão de palavras do Wall Street Journal anotadas
 - Canadian Hansards: 10 milhões de pares de sentenças em Inglês/Francês
 - Reviews de produtos
 - Documentos
 - Postagens em redes sociais

PLN Estatística

- Como várias outras partes da Inteligência Artificial, o PLN usa muitos métodos estatísticos
 - Precisamos de muitos dados sobre um determinado contexto
 - Soluções podem ser aprendidas a partir desses dados (aprendizagem de máquina)

Esparsidade

- A linguagem natural é esparsa
- A frequência de uma palavra é inversamente proporcional a sua posição em um ranking de frequência (Zipf's law)
- A palavra mais frequente ocorre aproximadamente
 - Duas vezes mais que a segunda palavra mais frequente
 - Três vezes mais que a terceira palavra mais frequente
 - ...
- No Brown Corpus, 135 palavras ocupam metade dos dados

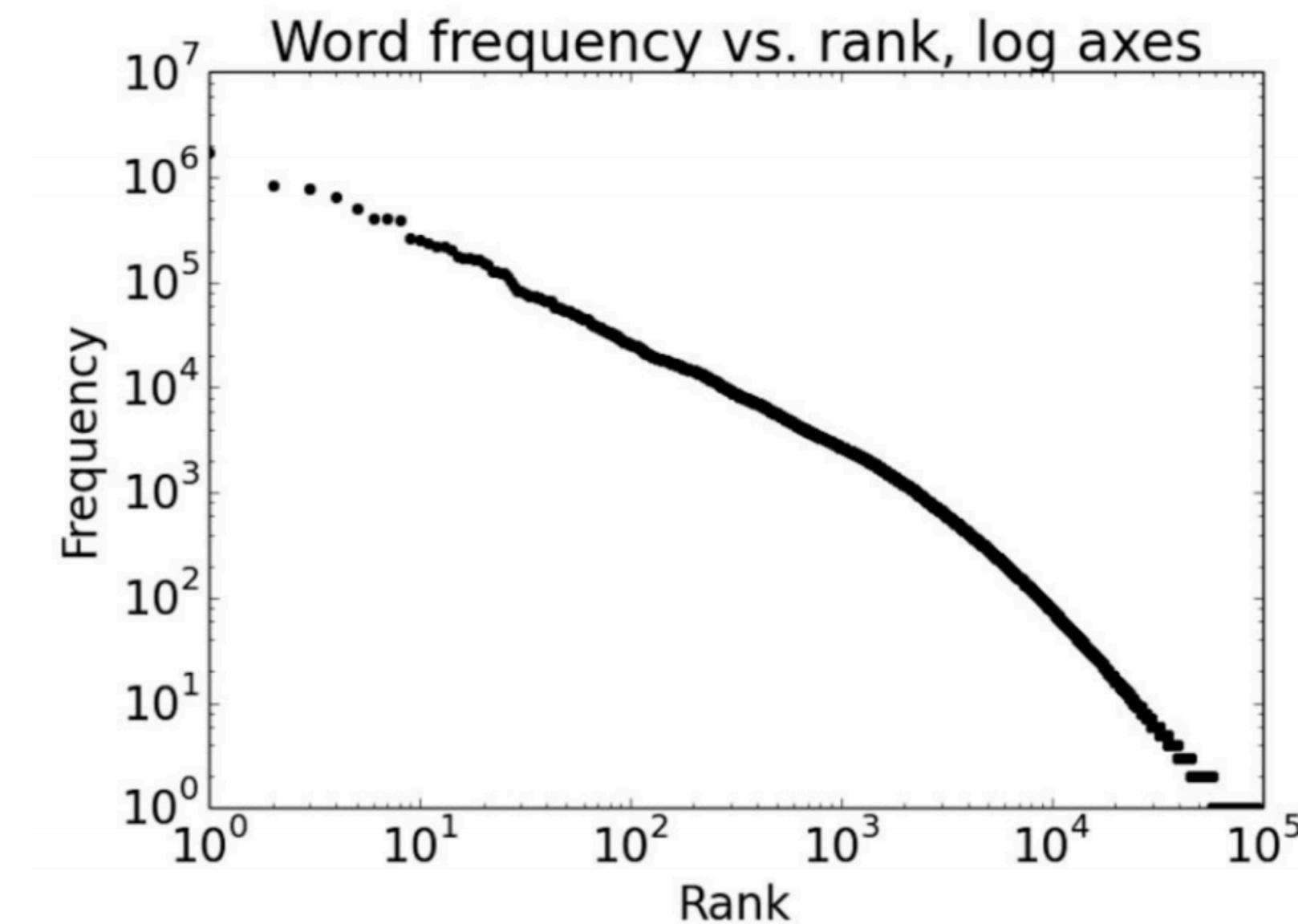
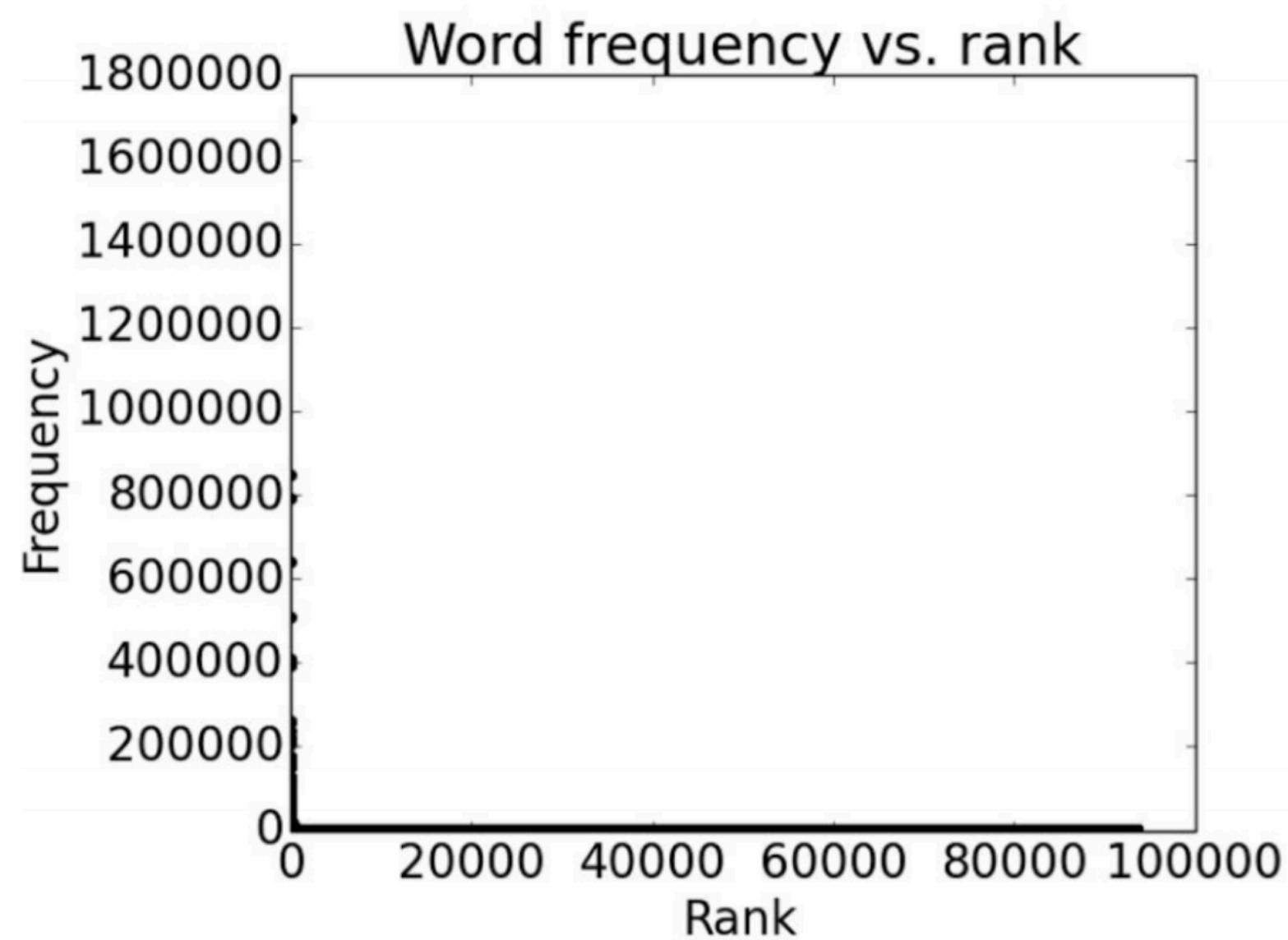
Esparsidade

- No livro Tom Sawyer, temos as seguintes frequências:

Word	Count f	rank r	fr
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	410	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
two	104	100	10400
turned	51	200	10200
comes	16	500	8000
family	8	1000	8000
brushed	4	2000	8000
Could	2	4000	8000
Applausive	1	8000	8000

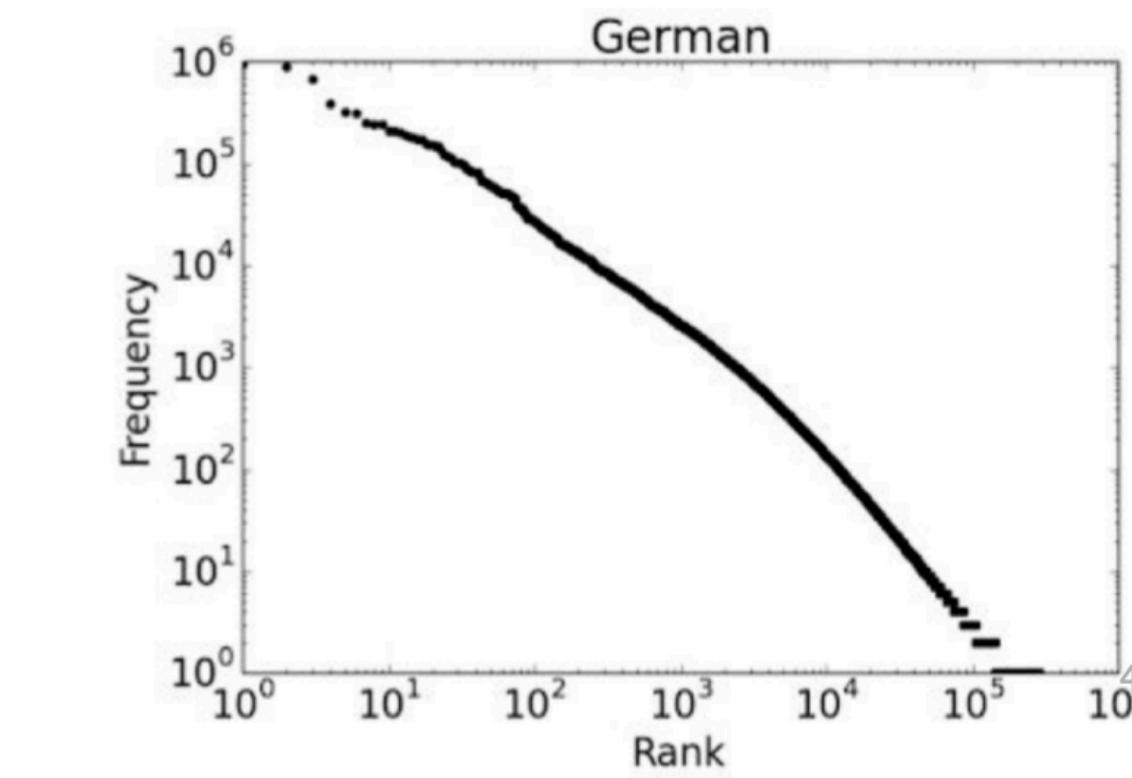
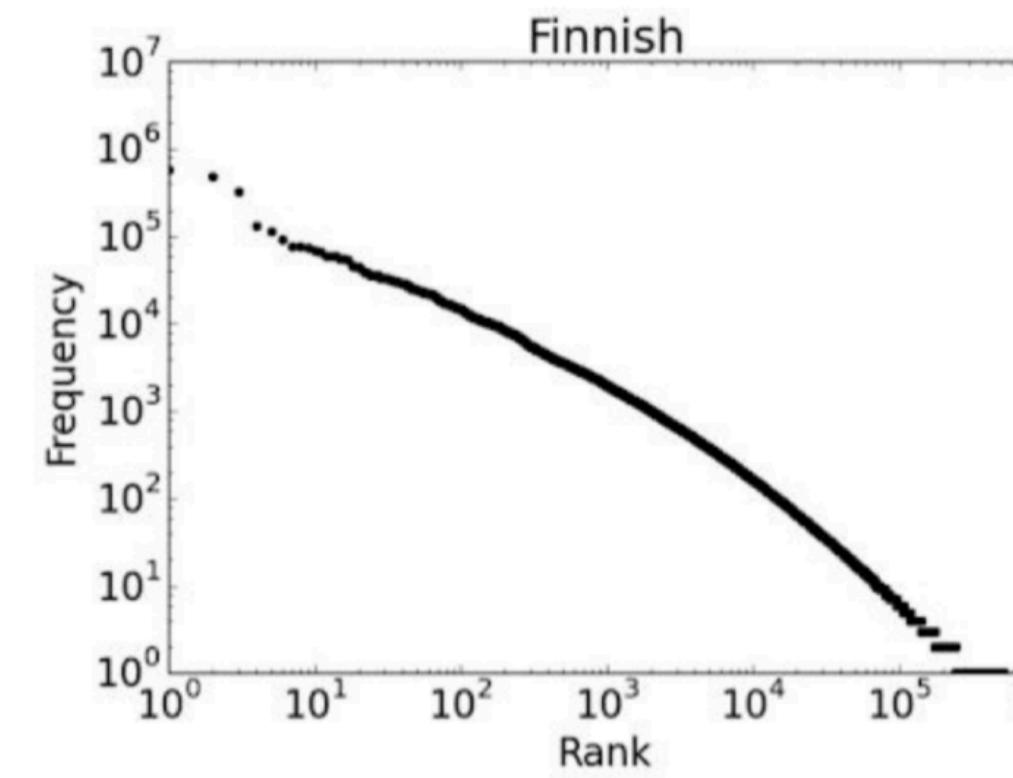
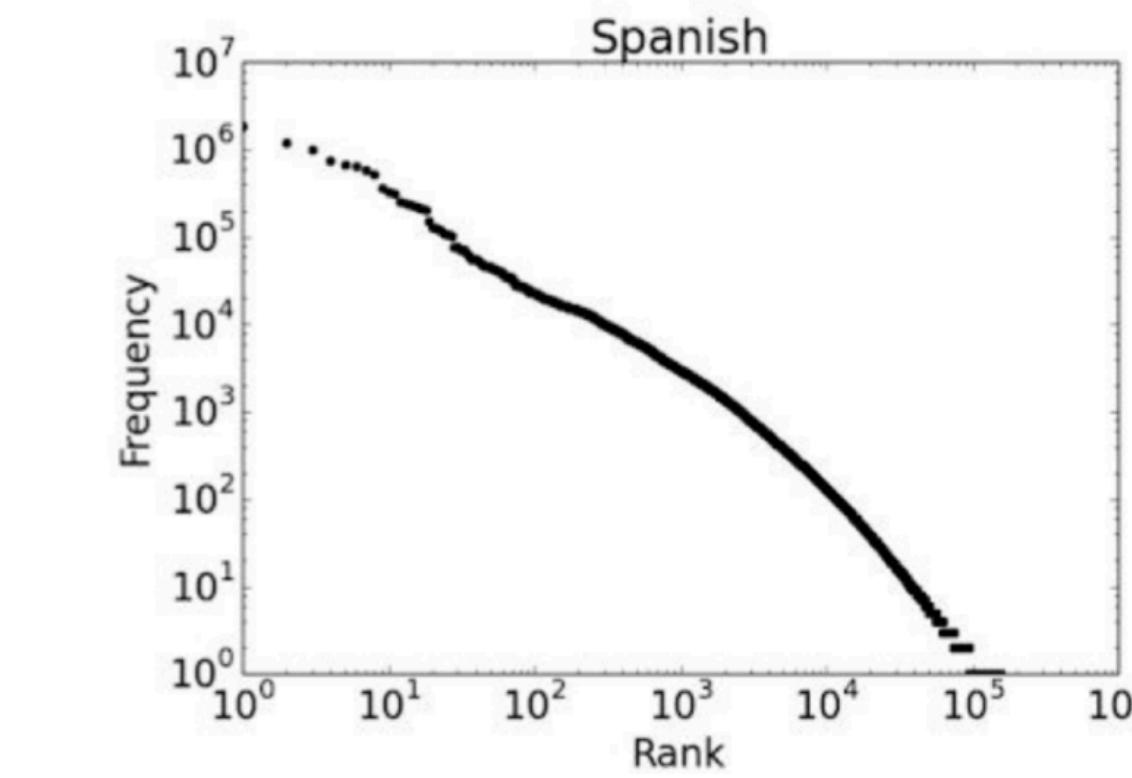
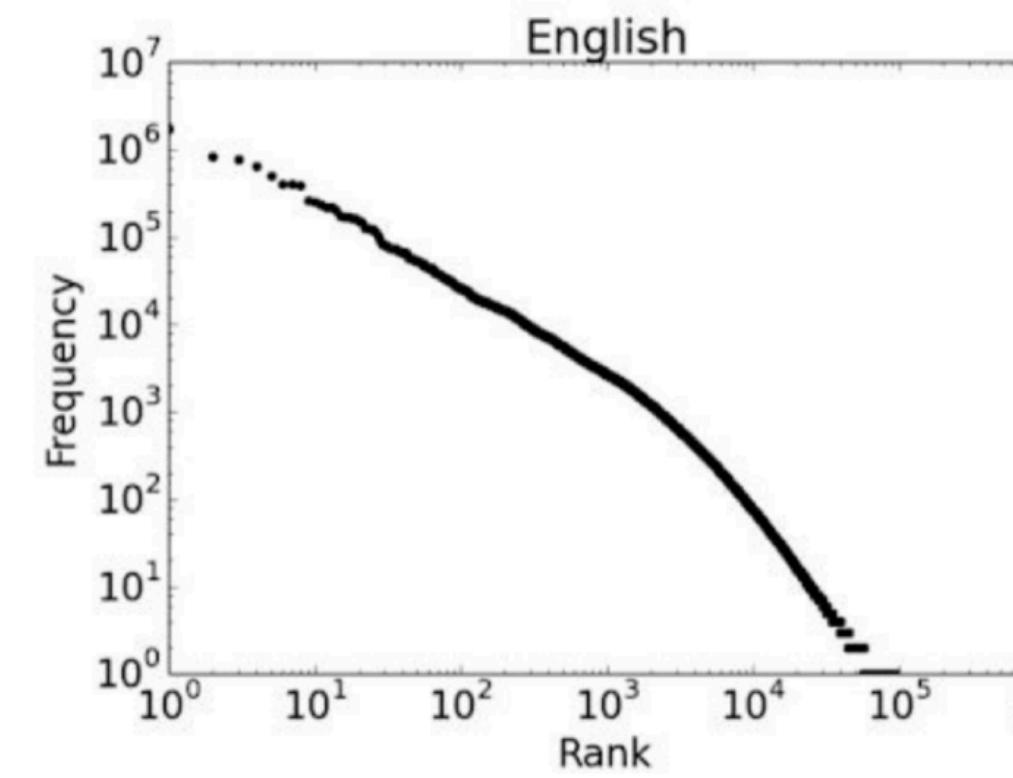
Esparsidade

- Gráfico ranking x frequênciа



Esparsidade

- Essa tendência aparece em outros idiomas também:



Expressividade

- Uma mesma forma pode ter vários sentidos (ambiguidade), mas diferentes formas podem ter o mesmo sentido:
 - Algumas crianças apareceram - Poucas crianças visitaram
 - Ela deu o livro para João - Ela deu para João o livro
 - A janela ainda continua aberta? - Por favor, feche a janela

Fatores Que Estão Mudando o PLN

- Aumento do poder computacional
- Aumento da disponibilidade de dados
 - Web e Internet
 - Redes Sociais
- Avanços na aprendizagem de máquina
- Avanços na compreensão da linguagem em um contexto social