CS 4650/7650: Natural Language Processing

Jacob Eisenstein

Lecture 8: Finite-state architectures

September 11, 2014

Michael Jordan on Reddit AMA

I'd use the billion dollars to build a NASA-size program focusing on natural language processing (NLP), in all of its glory (semantics, pragmatics, etc).
Intellectually I think that NLP is fascinating, allowing us to focus on highly-structured inference problems, on issues that go to the core of "what is thought" but remain eminently practical, and on a technology that surely would make the world a better place.

`http://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan/`

Homework

Swahili

1.	atanipenda	‘s/he will like me’
2.	atakupenda	‘s/he will like you’
3.	atampenda	‘s/he will like him/her’
4.	atatupenda	‘s/he will like us’
5.	atawapenda	‘s/he will like them’
6.	nitakupenda	‘I will like you’
7.	nitampenda	‘I will like him/her’
8.	nitawapenda	‘I will like them’
9.	utanipenda	‘you will like me’
10.	utampenda	‘you will like him/her’
11.	tutampenda	‘we will like him/her’
12.	watampenda	‘ they will like him/her’
13.	wametulipa	‘they have paid us’
14.	tulikulipa	‘we paid you’
15.	atanipiga	‘s/he will beat me’
16.	atakupiga	‘s/he will beat you’
17.	atampiga	‘s/he will beat him/her’
18.	ananipiga	‘s/he is beating me’
19.	anakupiga	‘s/he is beating you’
20.	anampiga	‘s/he is beating him/her’
21.	amekupiga	‘s/he has beaten you’
22.	amenipiga	‘s/he has beaten me’
23.	amempiga	‘s/he has beaten him/her’
24.	alinipiga	‘s/he beat me’
25.	alikupiga	‘s/he beat you’
26.	alimpiga	‘s/he beat him/her’
27.	atakusumbua	‘s/he will annoy you’
28.	unamsumbua	‘you are annoying him/her’

Homework

Protesters took to the streets of Barcelona on Thursday to demand that the Spanish government allow its Catalonia region to vote for independence.

Homework

Protesters took to the streets of Barcelona on Thursday to demand that the Spanish government allow its Catalonia region to vote for independence.

- Protest+er+s

Homework

Protesters took to the streets of Barcelona on Thursday to demand that the Spanish government allow its Catalonia region to vote for independence.

- Protest+er+s
- street+s

Homework

Protesters took to the streets of Barcelona on Thursday to demand that the Spanish government allow its Catalonia region to vote for independence.

- Protest+er+s
- street+s
- govern+ment

Homework

Protesters took to the streets of Barcelona on Thursday to demand that the Spanish government allow its Catalonia region to vote for independence.

- Protest+er+s
- street+s
- govern+ment
- in+depend+ence

Homework

Protesters took to the streets of Barcelona on Thursday to demand that the Spanish government allow its Catalonia region to vote for independence.

- Protest+er+s
- street+s
- govern+ment
- in+depend+ence
- journalist $\rightarrow$ journal+ist $\stackrel{?}{\rightarrow}$ jour+nal+ist

Weighted finite-state acceptors for NLP

- Edit distance
- Derivational morphology
- Language models

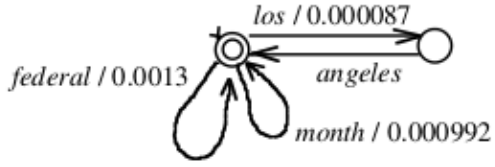
Some semirings

Name	$\mathbb{K}$	$\oplus$	$\otimes$	$\bar{0}$	$\bar{1}$	Applications
Boolean	$\{0, 1\}$	$\vee$	$\wedge$	0	1	identical to an unweighted FSA
Probability	$\mathbb{R}_+$	$+$	$\times$	0	1	sum of probabilities of all paths
Log-probability	$\mathbb{R} \cup -\infty \cup \infty$	$\oplus_{\log}$	$+$	$-\infty$	0	log marginal probability
Tropical	$\mathbb{R} \cup -\infty \cup \infty$	min	$+$	∞	0	best single path

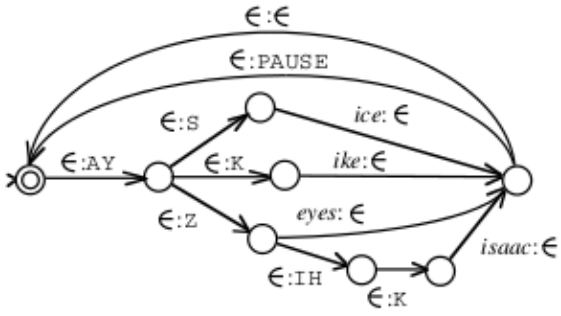
Finite-state transducers for NLP

- Edit distance
- Simple translation
- Transliteration
- Stemming
- Morphological analysis
- Sequence labeling

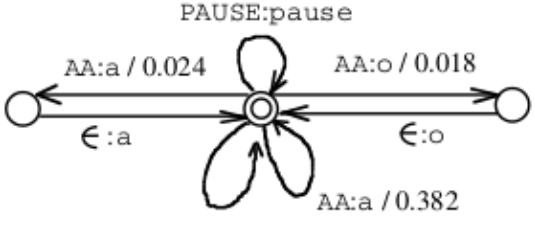
Transliteration



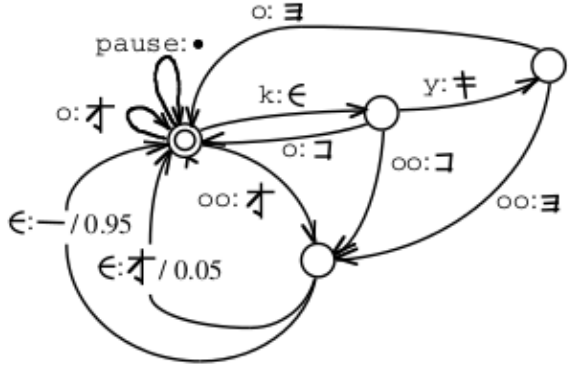
a WFSA A: Produce an English phrase



b WFST B: Convert English phrase to English sounds

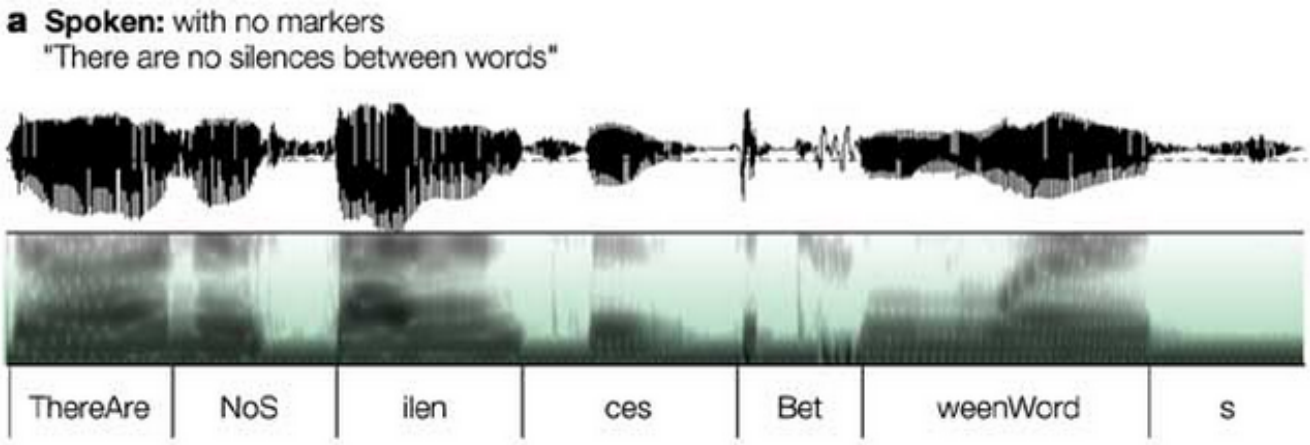


c WFST C: Convert English sounds to Japanese sounds



d WFST D: Convert Japanese sounds to Katakana

Segmentation



original, un-segmented text

再往远些看，随着汉字识别和语音识别技术的发展，
中文计算机用户将跨越语言差异的鸿沟，
在录入上走向中西文求同的道路。

separated word entities after segmentation

再 往 远 些 看 ， 随着 汉字 识别 和 语音 识别 技术 的 发展 ，
中文 计算机 用户 将 跨越 语言 差异 的 鸿沟 ，
在 录入 上 走 向 中 西文 求 同 的 道路 。