



Introduction to Arabic NLP

Sakhar Alkhereyf

October 31, 2020

Agenda

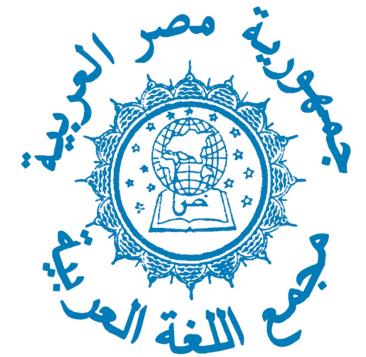
- Natural Languages
- Arabic Language
- NLP: Overview
- NLP: Applications
- NLP: Tools
- Arabic NLP: Overview
- Arabic NLP: Datasets
- Arabic NLP: Tools

Natural Languages

- Any language that has evolved naturally in humans
 - Arabic, English, French ... etc
- Constructed Languages (AKA: artificial languages)
 - programming languages (Python, Java, C ... etc)
- Natural Languages have evolved naturally
 - Not designed
 - No previous grammar or rules
 - Inherently ambiguous
- All human language varieties in the world are natural

Natural Language Regulations

- Some have more published prescriptions (standardization/regulation)
 - Modern Standard Arabic (MSA): Academies of Arabic Language
 - French: Académie Française (the French Academy)
- Prescriptions include:
 - Dictionaries: Lexicon
 - Orthography: Council for German Orthography
 - Grammar
- In real world, people do not follow language regulations



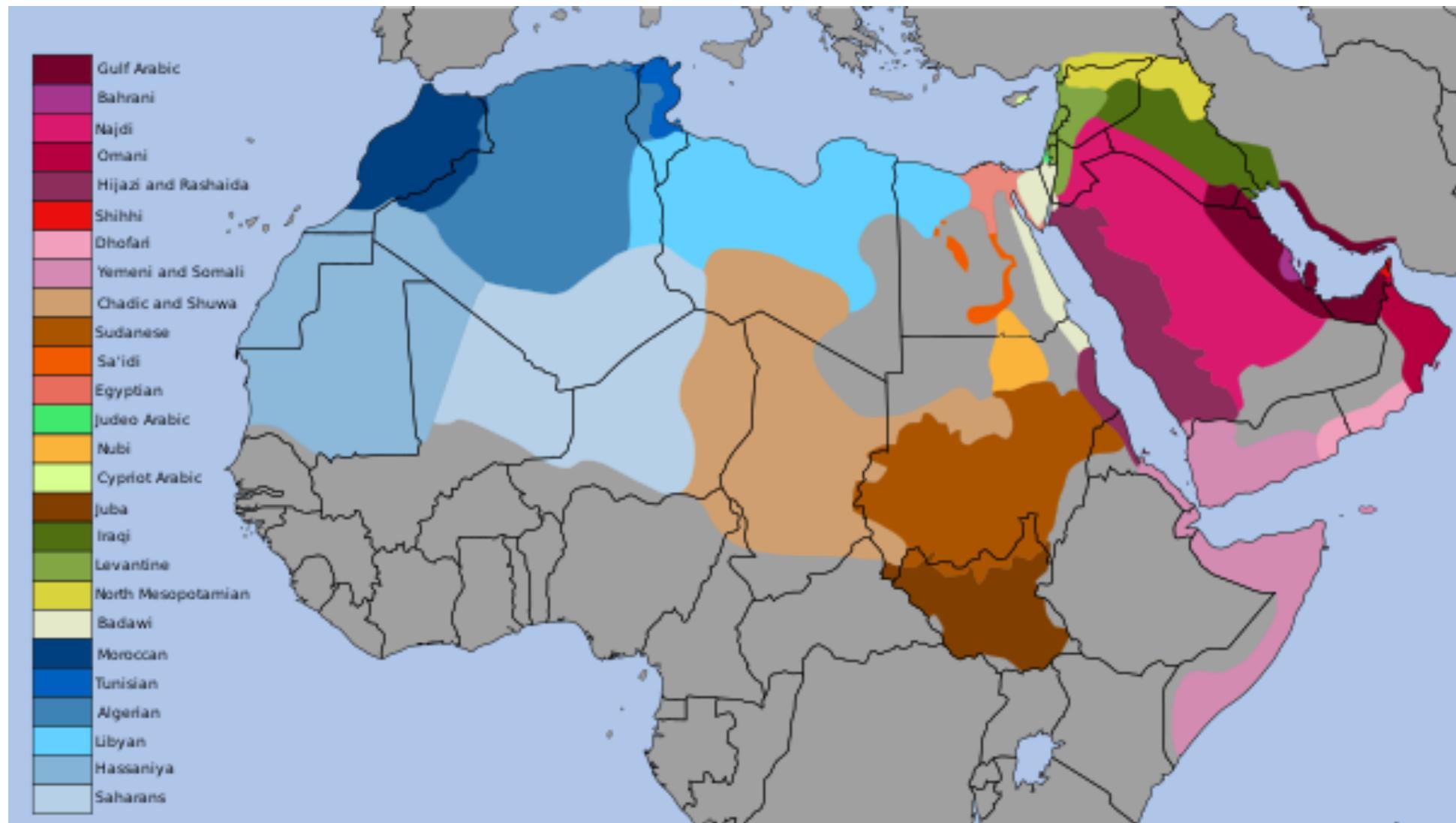
Arabic

العربية

- 6th most spoken language > 360 million speakers
- Exhibits “Diglossia”
 - Two (or more) varieties of the same language are used
 - Modern Standard Arabic (Fusha) & Dialects
 - Also Classical Arabic (CA)
- Dialects are different than MSA
 - Different lexicon
 - Different morphology
 - Different grammar

MSA vs Arabic Dialects

- MSA is mainly written, well regulated
 - No one's native language
 - Taught at school, used in formal situations
- Dialects are mainly spoken, no formal standards
 - Native language for Arabic speakers
 - Dominant in social media
- Dialects are mutually intelligible

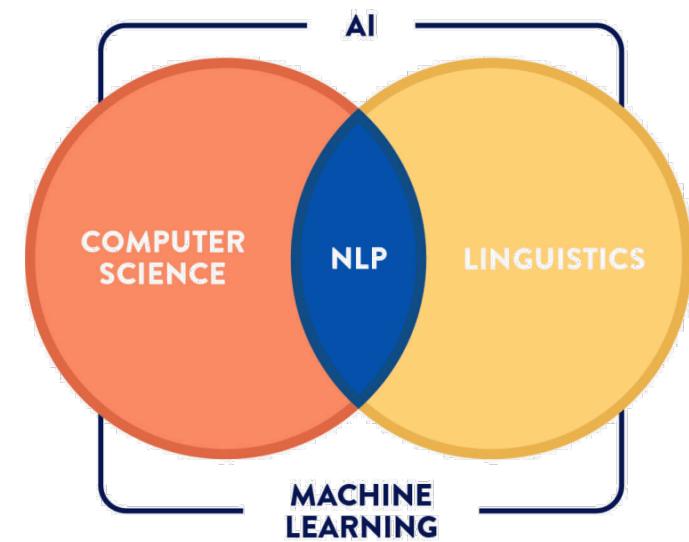


Map distributed under a CC-BY 3.0 license from Wikipedia.

Natural Language Processing (NLP)

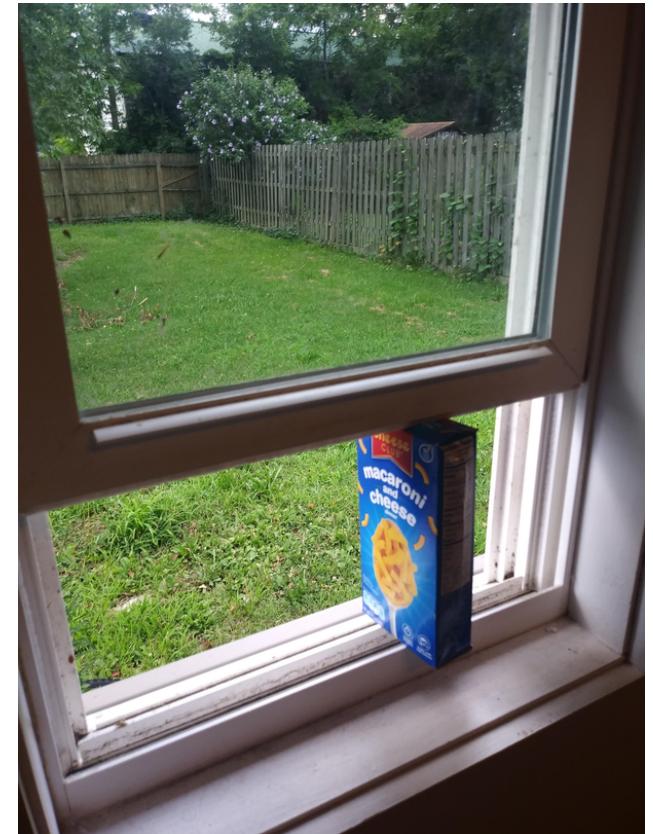


- AKA: Computational Linguistics
- Design and study of language technologies
 - Spelling correction, Machine Translation, Text Analytics ...
- Interdisciplinary field
 - Computer Science (AI), linguistics, sociology
- Most of NLP problems are AI-hard and still unsolved



Why NLP?

- Most of communications unstructured
- Human language is hard to model
 - Ambiguity: bank (river or financial?)
 - Illiteral meaning: sarcasm
 - “Now Mac supports Windows”
- Challenges
 - Scale
 - Language variation
 - Unknown representation



NLP Technologies

Applications

- Sentiment Analysis
- Machine Translation
- Question Answering
- Text Classification

Core Technologies

- Language Modeling
- Part-of-Speech Tagging
- Named-Entity Recognition
- Syntactic Parsing

Machine Translation

The screenshot shows the Google Translate web interface. At the top, there's a navigation bar with a menu icon, the "Google Translate" logo, and a user profile icon with a letter "S". Below the bar, there are two tabs: "Text" (selected) and "Documents". The main area has four language selection dropdowns: "DETECT LANGUAGE" (grayed out), "ENGLISH" (grayed out), "ARABIC" (selected, highlighted in blue), and "ITALIAN" (grayed out). Between the first and second language pairs is a double-headed arrow icon. Between the third and fourth language pairs is a downward arrow icon. The left column contains the source text in Arabic and its corresponding machine-generated English translation. The right column contains the target text in English and its corresponding machine-generated Italian translation. A green dot is located at the bottom center of the page.

Google Translate

≡ S

Text Documents

DETECT LANGUAGE ENGLISH ARABIC ITALIAN

ARABIC ENGLISH ITALIAN

وصل إلى مطار الملك عبدالعزيز في جدة الليلة، محترف الاتحاد المصري أحمد حجازي، وكانت نادي إدارة الاتحاد قد وقعت مع المدافع المصري قبل نهاية فترة الانتقالات الصيفية قادماً من وست بروميتش الإنجليزي حتى نهاية الموسم الحالي.

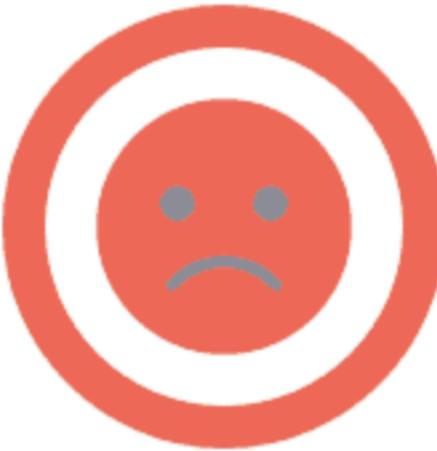
The Egyptian Al-Ittihad professional Ahmed Hegazy arrived at King Abdulaziz Airport in Jeddah tonight, and the Al-Ittihad management club had signed with the Egyptian defender before the end of the summer transfer period coming from West Bromwich Albion, England, until the end of the current season.

wasal 'ilaa matar almalik eibdaleziz fi jidat alliylati, muhtarak alaitihad almisrii 'ahmad hijazi, wakanat nadi 'iidarat alaitihad qad waqaeat mae almudafie almisrii qabl nihayat fтрат alaintiqalat alsayfiat qadmaan min wst birwmitsh 'albiun al'iinjalizii hataa nihayat almawsim alhali.

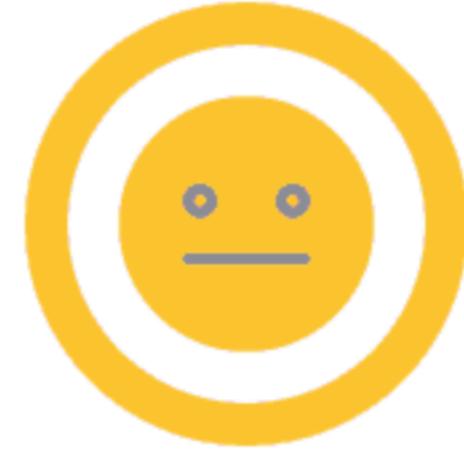
Sentiment Analysis



I am happy

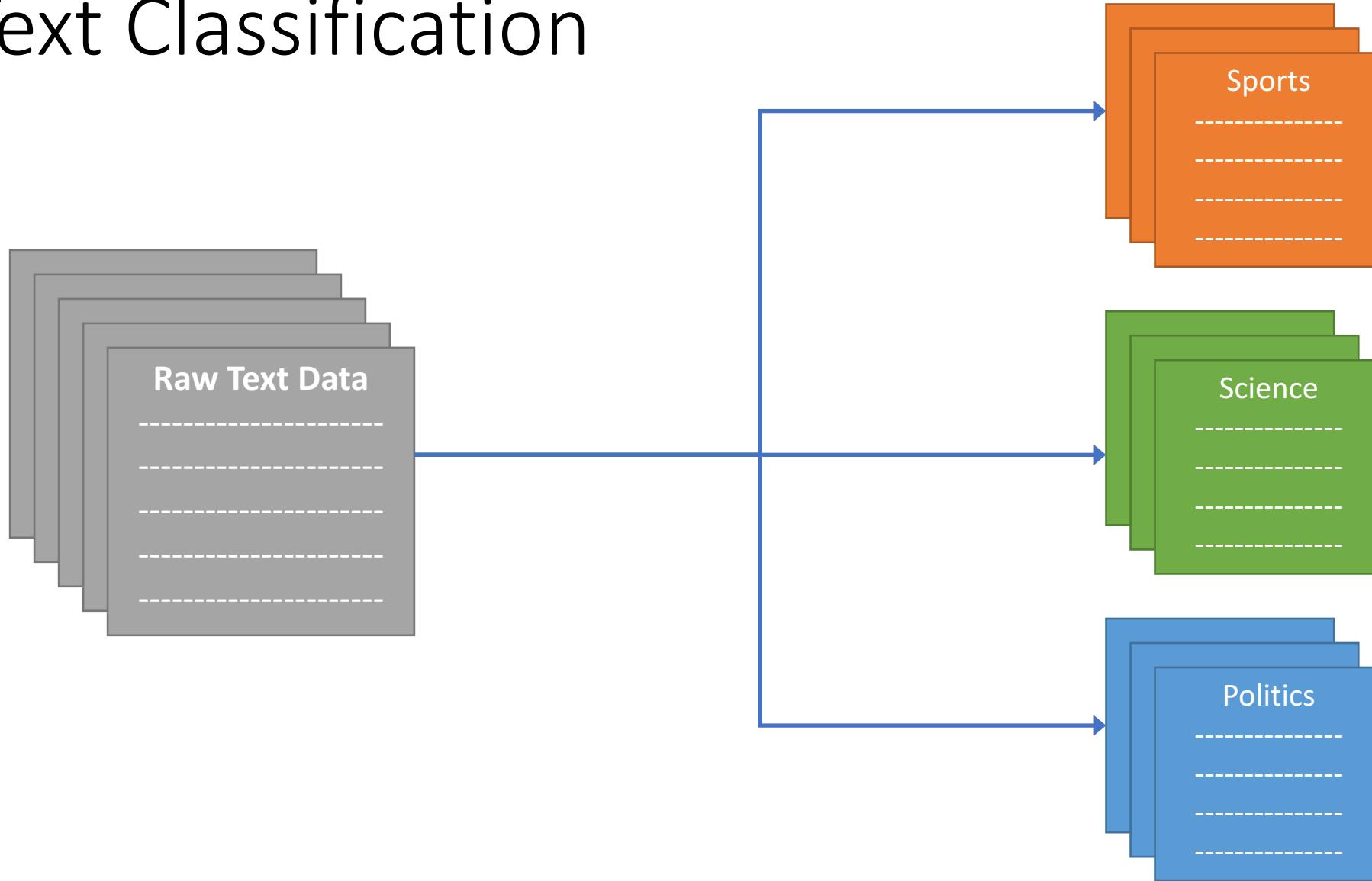


I am sad



I am Sakhar

Text Classification

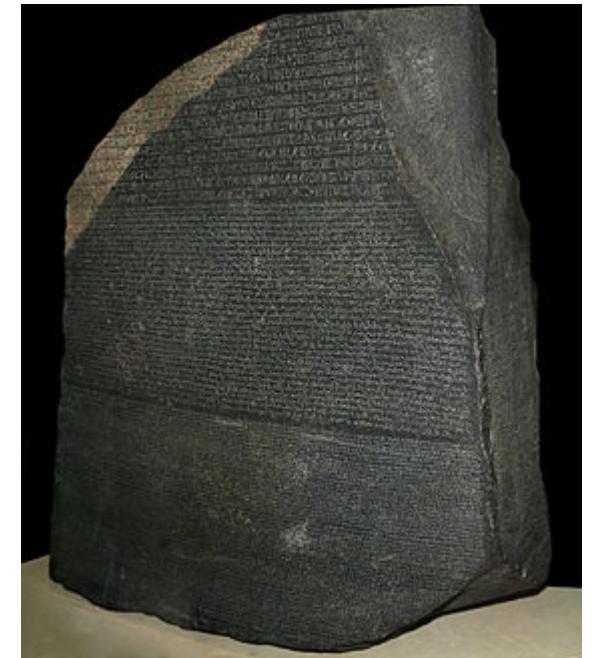


NLP Terminologies

- Corpus
- Tokenization
- Stemming
- Lemmatization
- Part-of-speech (POS) tagging
- Named-entity recognition (NER)

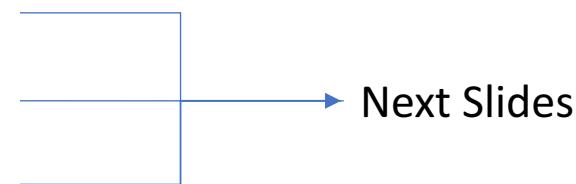
Corpus

- Literally Latin for *body*
- In linguistics and NLP refers to a collection of texts
- Can be in single/multiple languages
- *Genre*: news articles, social media ...
- Examples: Yelp reviews, Tweets, Wikipedia articles



NLP Core Technologies

- Used in the pipeline for other *downstream tasks*
 - Machine Translation
 - Sentiment Analysis
 - Text Classification
- Examples
 - Tokenization & Segmentation: splitting text into smaller units
 - Lemmatization & Stemming: *studies, studying, studied* → **Study**
 - POS tagging
 - NER
 - Parsing



Next Slides

Part-of-speech tags

- AKA: lexical categories, morphological classes, lexical tags
- Coarse-grained
 - Noun, verb, adjective, ...

Plays[VERB] well[ADVERB] with[PREPOSITION] others[NOUN]

- Fine-grained
 - Nouns: noun-proper-singular, noun-proper-plural, noun-common-mass, ...
 - Verbs: verb-past, verb-present-3rd, verb-base, ...
 - Adjectives: adjective-simple, adjective-comparative, ...

Plays[VBZ] well[RB] with[IN] others[NNS]

Named-Entity Recognition

- Sub task in information extraction
- Classify named-entities into categories
 - persons, organizations, locations ...

Aramco **ORG** 's headquarters is located in **Dhahran GPE**, between the **two CARDINAL** cities **Dammam GPE** and **Al-Khobar GPE** in the **Eastern NORP** province of **Saudi Arabia GPE** on the coast of **the Arabian Gulf LOC**. **Amin Hassan Al-Nasser PERSON** is the President.

Tokenization + POS + NER & Lemmatization

اَكْدُ دُكْتُور بَدْرَان عَمَر مُدِير جَامِعَة مَلِك
سُعُود رِيَاض اَنْجَامِعَة بَذَلْ اَقْصَى جَهْد اِسْتِيعَاب
اَكْبَر عَدَد مُمْكِن مِن طَالِب طَالِب .

verb nominal particle proper noun

Lemmatized:

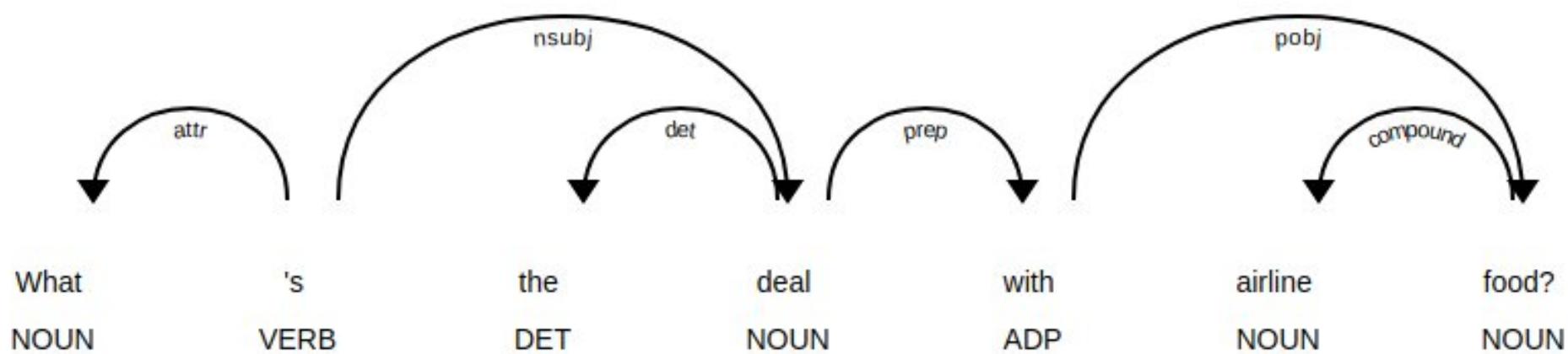
أَكْدُ دُكْتُور بَدْرَان عَمَر مُدِير جَامِعَة مَلِك
سُعُود رِيَاض أَنْجَامِعَة بَذَلْ أَقْصَى جَهْد اِسْتِيعَاب
أَكْبَر عَدَد مُمْكِن مِن طَالِب طَالِب .

verb nominal particle proper noun

•Chunking (Shallow Parsing)

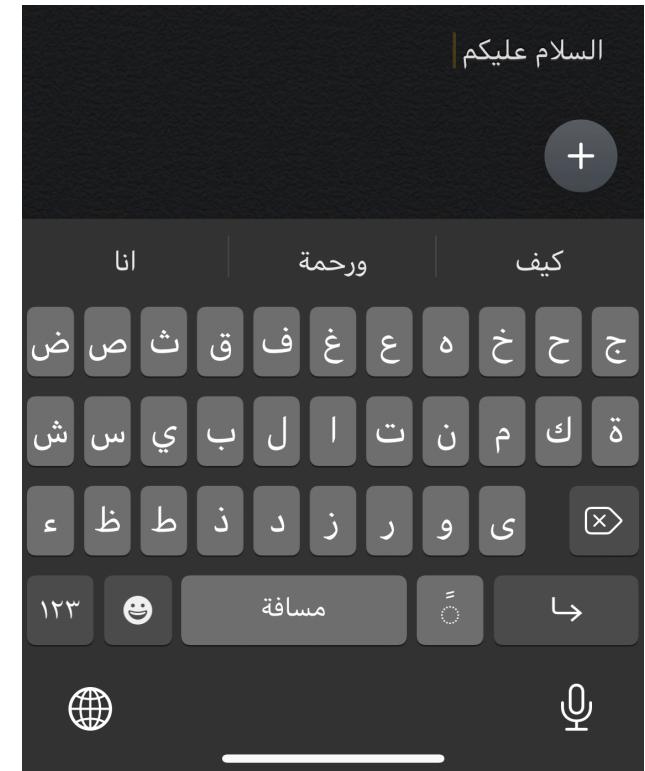


Parsing



Language Modeling

- LM predicts the probability of a sequence of words
 - $P(\text{"the cat is small"}) > P(\text{"small the cat is"})$
- Given a sequence of text predict next
 - Riyadh is the capital of
 - Sequence: letters, words, sentences ...
- Used to improve other NLP tasks
 - Speech Recognition, Machine Translations, OCR, information retrieval ...
 - “Wreck a nice beach” vs “recognize speech”

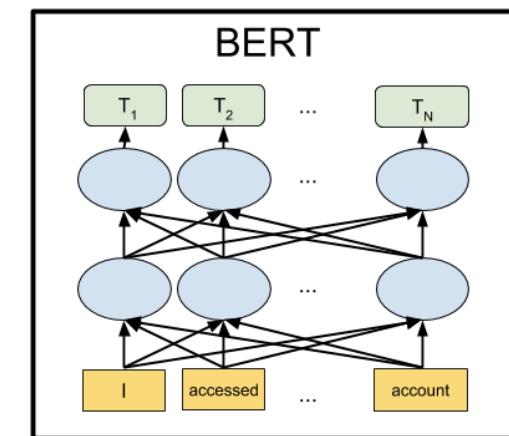
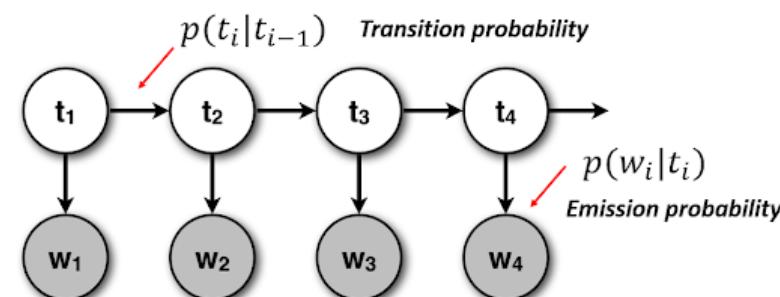


NLP Approaches

Rule-Based NLP	Classical ML/Statistical NLP	Neural Networks/Deep Learning
Linguists write explicit rules	Humans extract features, machines learn rules	Machines learn the rules and features

Input: an adjective and the noun it modifies
Output: inflect adjective agreed with the noun it modifies

Noun-adjective agreement Rule:
If Noun.number = broken_plural
Then
begin
 synthesize_noun(Noun.gender,Adj.stem)
 synthesize_noun(Noun.definiteness,
 Adj.stem)
end
else
begin
 synthesize_noun(Noun.number,Adj.stem)
 synthesize_noun(Noun.gender,Adj.stem)
 synthesize_noun(Noun.definiteness,
 Adj.stem)
end
End Rule



NLP Approaches

Rule-Based NLP	Classical ML/Statistical NLP	Neural Networks/Deep Learning
Linguists write explicit rules	Humans extract features, machines learn rules	Machines learn the rules and features
<ul style="list-style-type: none">• Easy to understand the rules• Limited or no data is needed• Useful for limited domains	<ul style="list-style-type: none">• Less computational resources• Less Data	<ul style="list-style-type: none">• Capture semantics• Allow “transfer learning”• Auto learn representation
<ul style="list-style-type: none">• Time consuming, hard to scale• Perform poorly when rules aren't followed	<ul style="list-style-type: none">• No semantic capturing• Feature engineering is expensive• Doesn't generalize to other tasks	<ul style="list-style-type: none">• Require huge amount of data• expensive computational resources

NLP Approaches

- Machine learning approaches are dominant in NLP
 - Deep learning became popular in last years
- However, all approaches still in use
- Rule-based methods might be best if rules are limited
 - Example: extract articles from legal documents



NLP Computational Tools

spaCy



PL	Python	Python	Java
Supports Arabic	No	Partially	Partially

NLP Application Domains

- Healthcare
 - Medical document analytics
- Finance (FinTech)
 - Financial sentiment analysis
- Marketing
 - Product sentiment analysis
- Law
 - Legal document analytics

Arabic NLP

- Harder than English NLP
- Less resources
 - Datasets
 - Tools
- Less research

Arabic NLP Challenges

- Orthographic (spelling) inconsistency
- Morphological complexity
- Dialectal Variation
 - Phonological variations
 - Example: pronunciation of the letter ق: /q/, /g/ ...
 - Lexical, Morphological ...

Orthographic (Spelling) Inconsistency

- No orthographic standards for dialects

Arabic Orthography	Arabic Transliteration	Frequency
مبِيقولهاش	<i>mbyqwlhAš</i>	≈ 26,000
ما بِيقولهاش	<i>mA byqwlhAš</i>	≈ 13,000
ما بِقلهاش، مبِقولهاش، مبِقلهاش، ما بِقلهاش، ما بِيقولهاش	<i>mAbqlhAš, mbqwlhAš, mbqlhAš, mA bqlhAš, mAbiyqwlhAš</i>	≤ 10,000
ما بِقولهاش، ما بِقولهاش، مبِقلهاش، ما بِقلهاش	<i>mAbqwlhAš, mA bqwlhAš, mbyqlhAš, mA byqlhAš</i>	≤ 1,000
مبِئلهاش، ما بِيئولهاش، ما بِيئلهاش، ما بِيؤلهاش	<i>mbŷlhAš, mAbŷwlhAš, mA byŷwlhAš, mAbyŵlhAš</i>	≤ 100
ما بِيؤلهاش، ما بِئلهاش، مبِئولهاش، ما بِيئلهاش، ما بِئولهاش، ما بِئلهاش، ما بِؤلهاش، مبِئولهاش، مبِئولهاش، ما بِئلهاش، مبِئلهاش	<i>mA byŵlhAš, mAbŷlhAš, mbyŷwlhAš, mA byŷlhAš, mAbŷwlhAš, mA bŷlhAš, mA bw̄lhAš, mbŷwlhAš, mbyŵlhAš, mA bw̄lhAš, mbw̄lhAš</i>	≤ 10

Morphological Complexity

- Arabic is morphologically rich
 - A core word has many inflected forms

قال، قالت، قلت، يقول، تقول، يقولون، نقول،
قاله، قالها، قالت، قالتها، سيقول، وسيقول،
و سنقول، و سنقولها

و سـنـقـولـهـا

و + سـ+نـ+قـولـ+هـا

ha+qwl+n+s+w

it+say+will+we+and

- English is **not** morphologically rich
 - number of inflected forms is small
 - Say, says, saying, said

Arabic Dialectal Variation



Morphological unit	MSA (Fusha)	Saudi (Riyadh)	Saudi (Jeddah)	Egyptian (Cairo)
Future Particle	swf س + سوف	b	ح	حـ + حـ
Progressive Particle	X	qAEd + قاعد جاـس	قـاعـد	b
Possessive Particle	X	Hq حق	Hq حق + tbE تـبع	btAE بـتـاع

الآن حين حين دلوقتي

Arabic NLP Common Preprocessing

Transliteration

- Transferring a word from the alphabet of one language to another
- In this context: Arabic to Latin script (Romanization)


السلام → AlslAm
- Useful for avoiding encoding issues (intermediate representation)
- Also, if human doesn't understand Arabic script

Buckwalter transliteration (2002)

- ASCII only transliteration scheme
- One-to-one representation
 - Each Arabic letter is represented by one ASCII letter and vice versa
 - Includes diacritics (harakat) and other characters

Arabic	إ	ي	و	ه	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض	ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	ا
Buckwalter	A	b	t	v	j	H	x	d	*	r	z	s	\$	S	D	T	Z	E	g	f	q	k	l	m	n	h	w	y	Y

Arabic	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ	ـ
Buckwalter	a	i	u	F	K	N	~	o	p	'	>	<	&	}		Y												

Dialect Identification

- Which dialect is this?
- Fine-grained vs coarse-grained
 - Country, Region, City, Sub-culture

Arabic NLP datasets and corpora: MADAR



- MADAR: Multi-Arabic Dialect Applications and Resources
- Under development
 - Parallel corpus of 25 cities from Arab world
 - Translation of Basic Traveling Expression Corpus (BTEC)
 - <https://camel.abudhabi.nyu.edu/madar/#Resources>

Standard Arabic	City	Dialectal
هناك ، أمام بيانات السائح تماما .	Jeddah	شوفه هناك، قدام مكتب المعلومات السياحية بالضبط.
لم اسمع بهذا العنوان من قبل بالقرب من هنا .	Jeddah	ماقد سمعت بدا العنوان هنا.
استمر في السير في هذا الطريق حتى تجد صيدلية .	Jeddah	امشي سيدا لين ما تلاقي الصيدلية.
كم تكلفة الإفطار ؟	Jeddah	بكم الفطور؟

Standard Arabic	City	Dialectal
هناك ، أمام بيانات السائح تماما .	Riyadh	هناك، بالضبط مقابل مكتب معلومات السياح.
لم اسمع بهذا العنوان من قبل بالقرب من هنا .	Riyadh	ما قد سمعت بهذا العنوان هنا.
استمر في السير في هذا الطريق حتى تجد صيدلية .	Riyadh	امش على طول لين تلقي صيدلية.
كم تكلفة الإفطار ؟	Riyadh	بكم الفطور؟

Arabic NLP datasets and corpora: Arap-tweet

- Released in 2018
- 2.4M Tweets corpus
 - 11 regions: Gulf, North/South Levant, ...
 - 16 countries
 - From 1,100 users on Twitter
- Annotated with age, gender, dialect

Dialect	Region
Moroccan	Morocco
Algerian	Algeria
Tunisian	Tunisia
Libyan	Libya
Egyptian	Egypt
Sudanese	Sudan
Lebanese	North Levant
Syrian	North Levant
Jordanian	South Levant
Palestinian	South Levant
Iraqi	Iraq
Qatari	Gulf
Kuwaiti	Gulf
Emirati	Gulf
Saudi	Gulf
Yemeni	Yemen

Arabic NLP datasets and corpora: AraBERT



- Arabic language model based on Google's BERT (2020)
- ~70M sentences (23GB) with ~3B words: Wikipedia and other corpora
- Two versions:
 - AraBERTv0.1: not segmented
 - AraBERTv1: segmented → المدرسات + ات + مدرس + ال
- <https://github.com/aub-mind/arabert>
- Multi-dialect-Arabic-BERT
 - Weights initialized using AraBERT weights (not from scratch)
 - Trained on 10M Arabic tweets
 - <https://github.com/mawdoo3/Multi-dialect-Arabic-BERT>

Arabic NLP tools

- MADAMIRA
- CAMeL Tools
- AraNet: A Deep Learning Toolkit for Arabic Social Media



MADAMIRA (2014)



- SOTA Arabic morphological analysis (2014)
 - Currently supports Standard Arabic (MSA) and Egyptian Arabic
 - Extensions: Saudi, Jordanian, Iraqi, Moroccan, ...
 - Hybrid approach: rule-based and machine learning components
- Written in Java
 - Stand-alone
 - Server-mode: XML format
 - API
- Licensed
- Demo: <https://camel.abudhabi.nyu.edu/madamira/>

CAMeL Tools (2020)



- Collection of open-source Arabic NLP tools
- Pre-processing: transliteration, normalization
- NER
- Sentiment analysis
 - pre-trained BERT models (AraBERT and others) fine-tuned on sentiment datasets
- Dialect identification
- https://github.com/CAMEL-Lab/camel_tools

AraNet (2020)

- A Deep Learning Toolkit for Arabic Social Media
- Set of pre-trained BERT models fine-tuned with
 - Age & Gender
 - Dialect identification
 - Emotion
 - Irony
 - Sentiment
- <https://github.com/UBC-NLP/aranet>

Conclusions

- Arabic NLP is harder than English
 - Less resources, less research
 - Dialectal variation
 - Morphological complexity
- Good news: a lot of research in recent years
 - More datasets and tools
- No best solution for all applications
 - May need to develop new solutions for some problems

Thanks!

Questions?