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

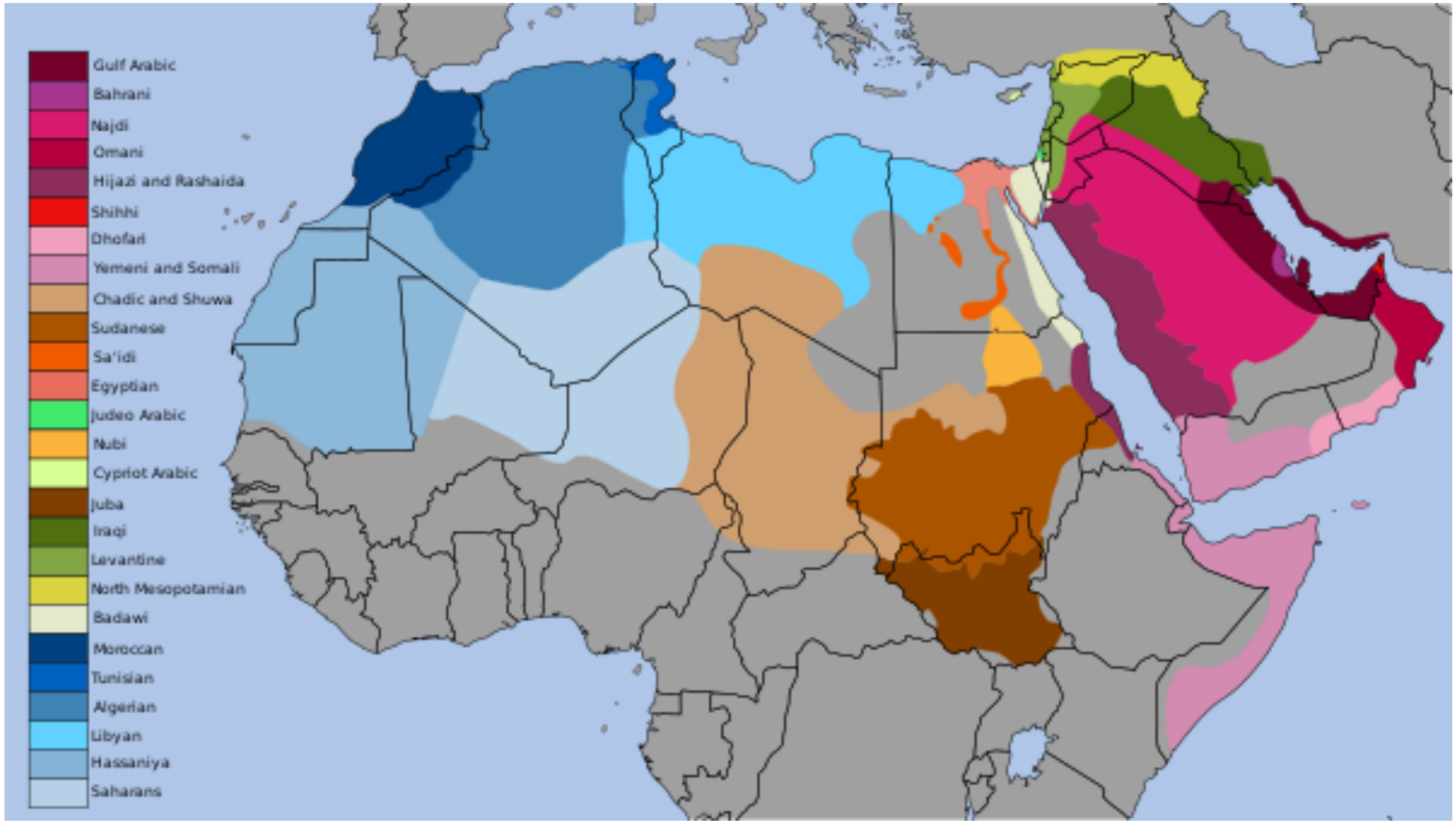
- Exhibits “Diglossia”
  - Two (or more) varieties of the same language are used
  - Modern Standard Arabic (Fusha) & Dialects
    - Also Classical Arabic (CA)
- Dialects are different than MSA
  - Different lexicon
  - Different morphology
  - Different grammar

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

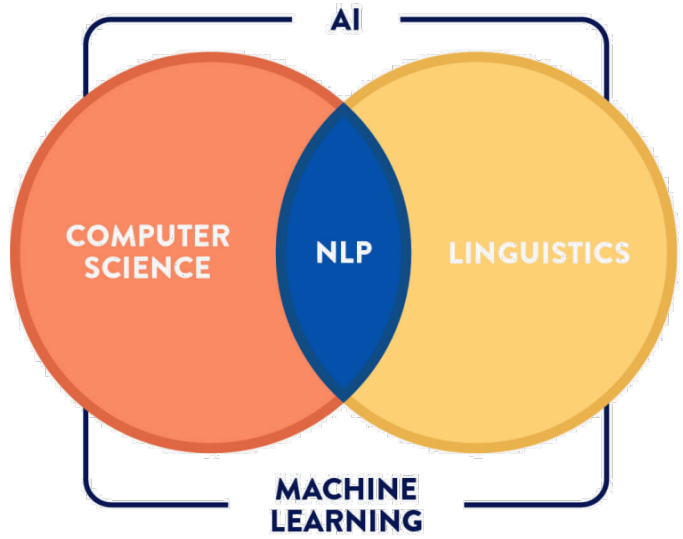


Natural Language Processing (NLP)

- 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

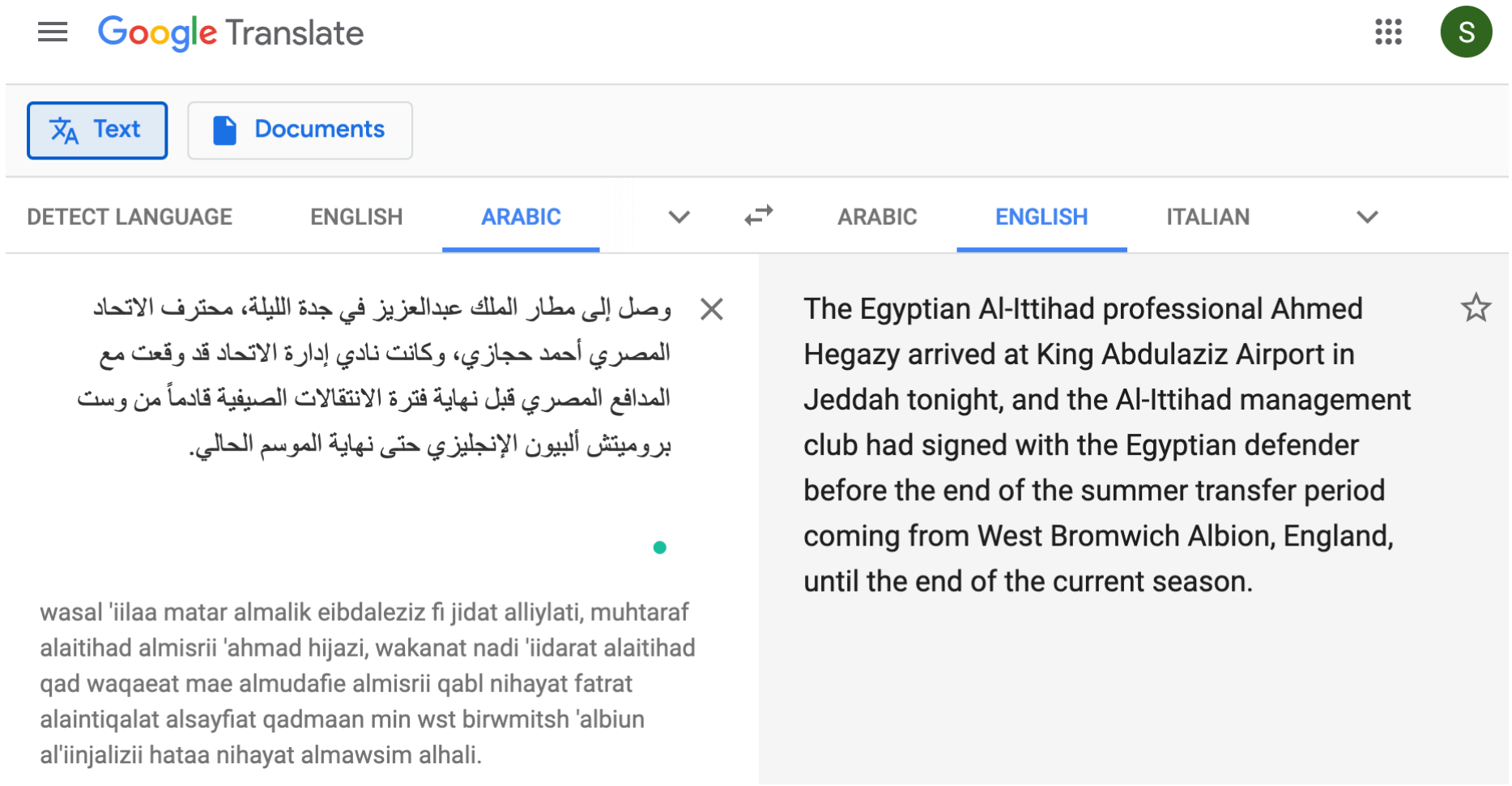
Applications

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

Core Technologies

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

Machine Translation



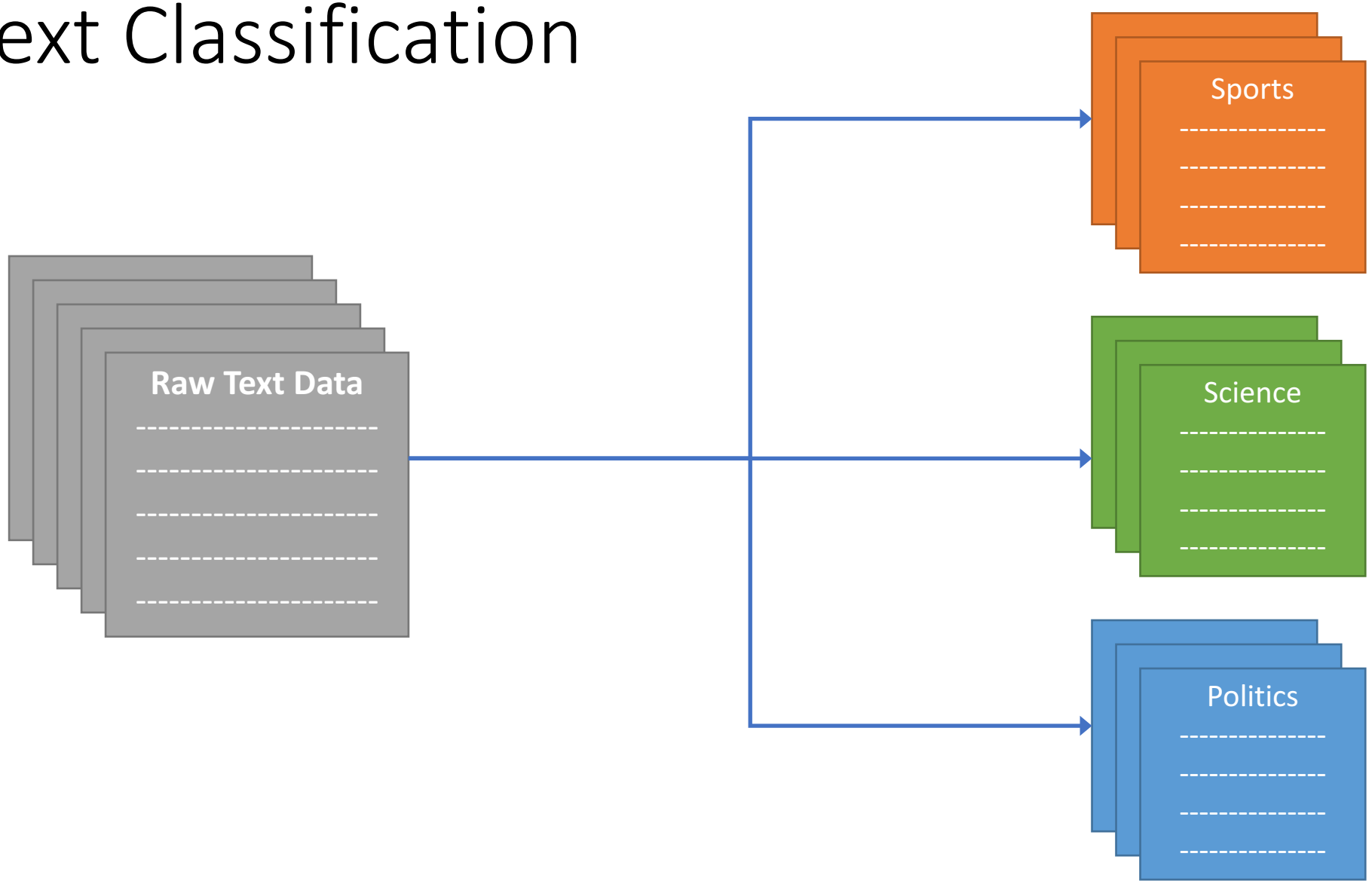
Sentiment Analysis

I am happy

I am sad

I am Sakhar

Text Classification

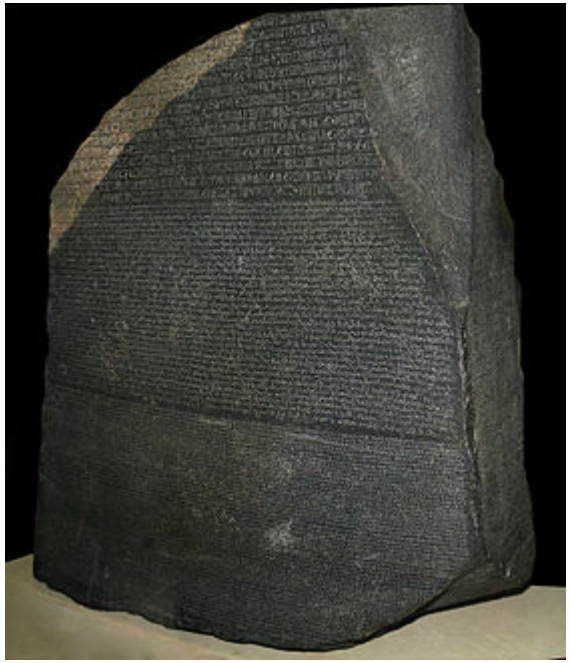


NLP Terminologies

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

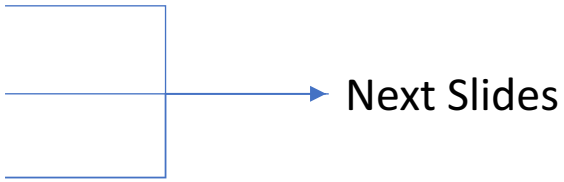
Corpus

- 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



Part-of-speech tags

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

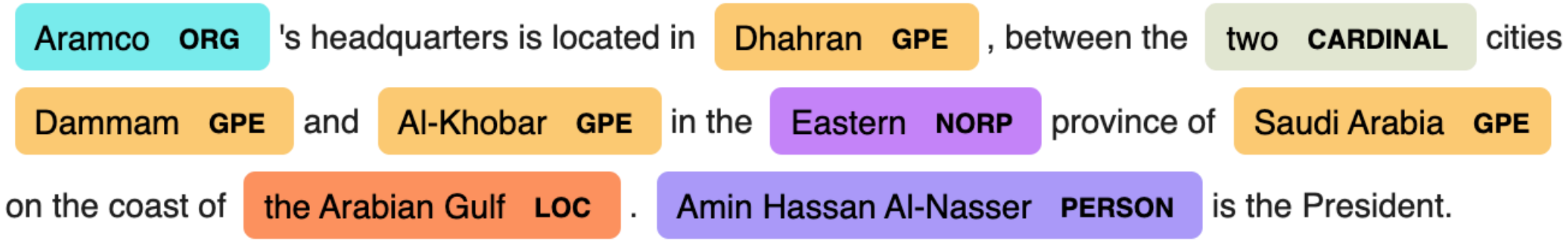
Plays[VERB] well[ADVERB] with[PREPOSITION] others[NOUN]

- 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, ...

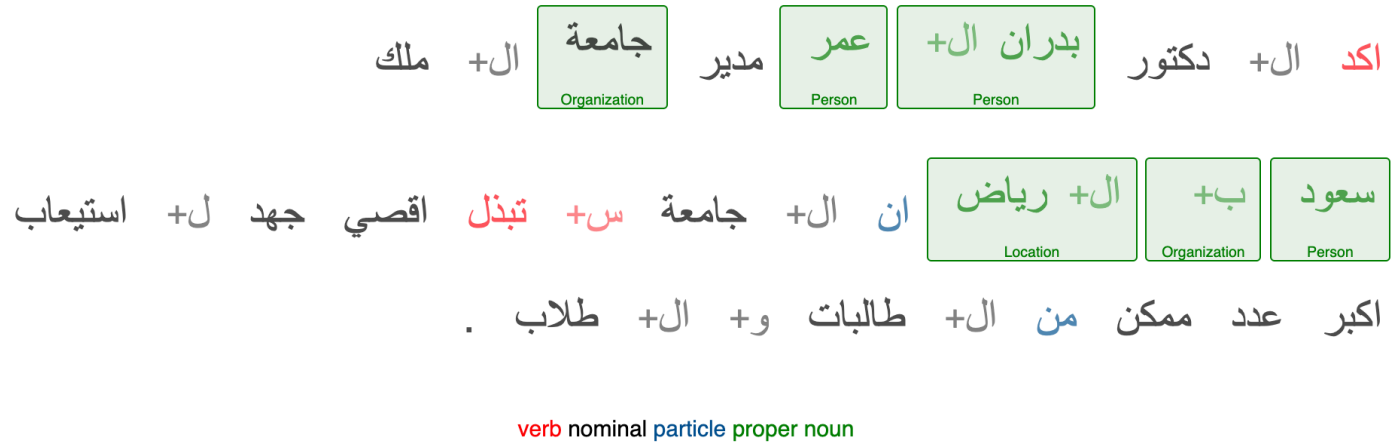
Plays[VBZ] well[RB] with[IN] others[NNS]

Named-Entity Recognition

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



Tokenization + POS + NER & Lemmatization



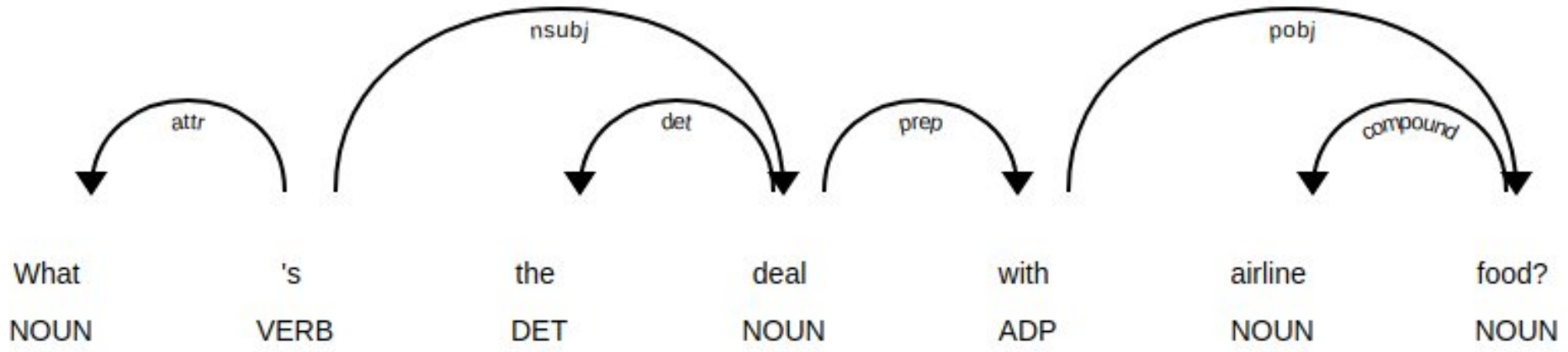
Lemmatized:

أَكَّد دُكْتُور بَدْران عُمَر مُدِير جامِعَة مَلِك
سُعُود رِياض أَنَّ جامِعَة بَذَل أَقْصَى جَهْد اِسْتِيعاب
أَكْبَر عَدَد مُمْكِن مِن طالِب طالِب .

verb nominal particle proper noun

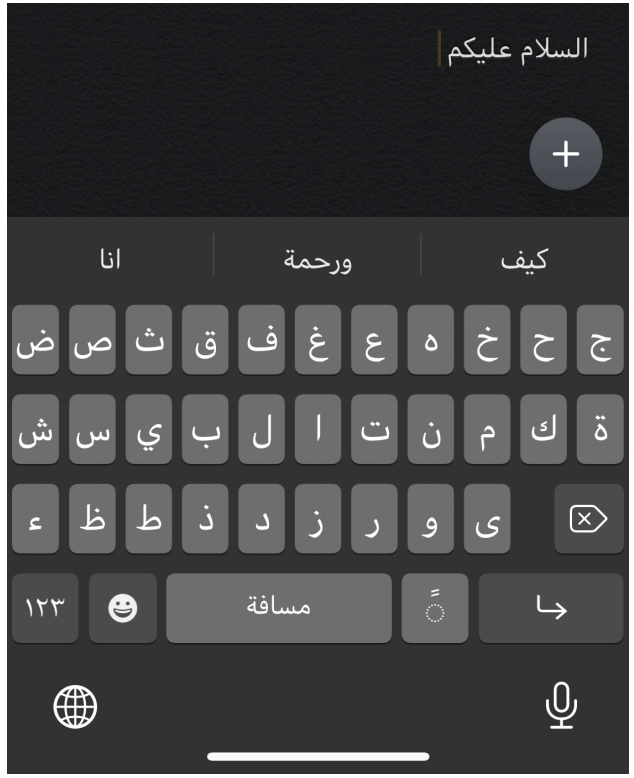
- Chunking (Shallow Parsing)

Parsing


attr
nsubj
det
prep
pobj
compound
What
's
the
deal
with
airline
food?
NOUN
VERB
DET
NOUN
ADP
NOUN
NOUN

Language Modeling

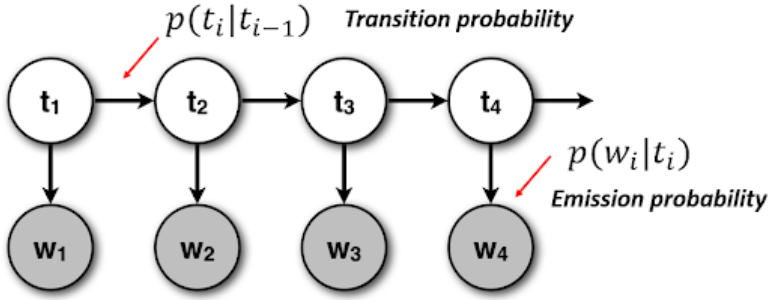
- LM predicts the probability of a sequence of words
  - P(“the cat is small”) > P(“small the cat is”)
- Given a sequence of text predict next
  - Riyadh is the capital of …..
  - Sequence: letters, words, sentences …
- Used to improve other NLP tasks
  - Speech Recognition, Machine Translations, OCR,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

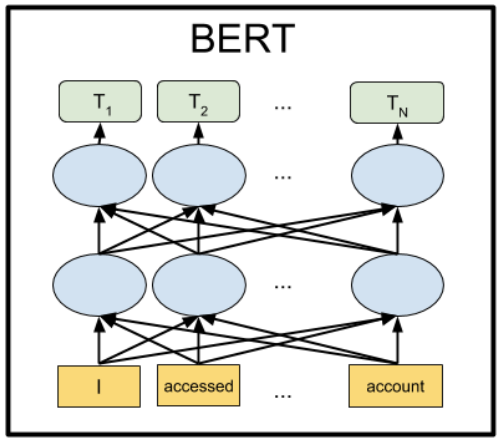


NLP Approaches

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
```



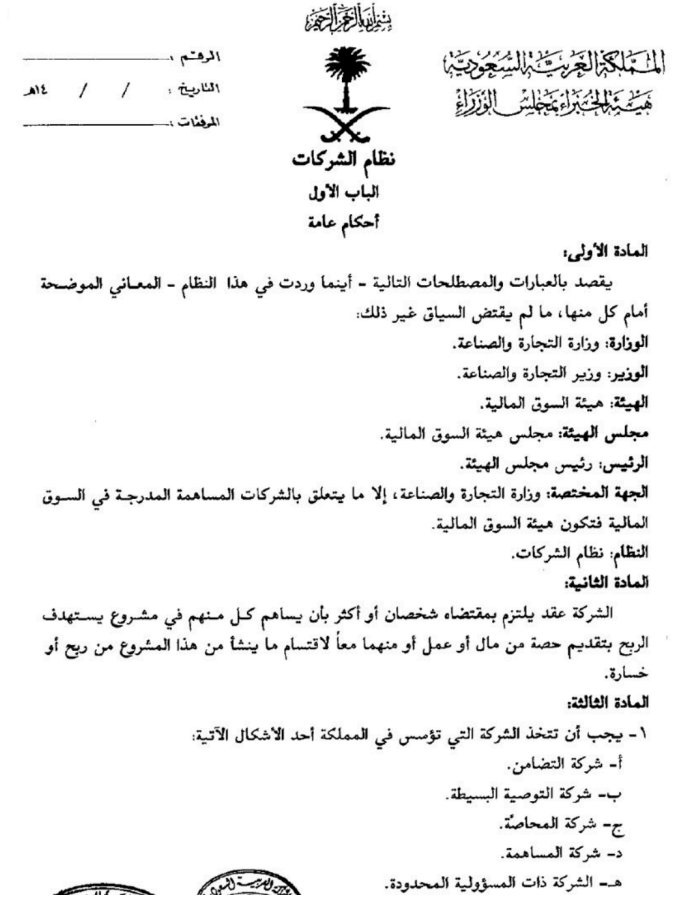


NLP Approaches

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
• Easy to understand the rules • Limited or no data is needed • Useful for limited domains	• Less computational resources • Less Data	• Capture semantics • Allow “transfer learning” • Auto learn representation
• Time consuming, hard to scale • Perform poorly when rules aren’t followed	• No semantic capturing • Feature engineering is expensive • Doesn’t generalize to other tasks	• 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



بسم الله الرحمن الرحيم

المملكة العربية السعودية
هيئة الخبراء بمجلس الوزراء

الرقم: ـــــــــــ
التاريخ: / / ١٤هـ
المرفقات: ـــــــــــ

نظام الشركات

الباب الأول

أحكام عامة

المادة الأولى:

يقصد بالعبارات والمصطلحات التالية - أينما وردت في هذا النظام - المعاني الموضحة أمام كل منها، ما لم يقتض السياق غير ذلك:

الوزارة: وزارة التجارة والصناعة.

الوزير: وزير التجارة والصناعة.

الهيئة: هيئة السوق المالية.

مجلس الهيئة: مجلس هيئة السوق المالية.

الرئيس: رئيس مجلس الهيئة.

الجهة المختصة: وزارة التجارة والصناعة، إلا ما يتعلق بالشركات المساهمة المدرجة في السوق المالية فتكون هيئة السوق المالية.

النظام: نظام الشركات.

المادة الثانية:

الشركة عقد يلتزم بمقتضاه شخصان أو أكثر بأن يساهم كل منهم في مشروع يستهدف الربح بتقديم حصة من مال أو عمل أو منهما معاً لاقتسام ما ينشأ من هذا المشروع من ربح أو خسارة.

المادة الثالثة:

١- يجب أن تتخذ الشركة التي تؤسس في المملكة أحد الأشكال الآتية:

أ- شركة التضامن.

ب- شركة التوصية البسيطة.

ج- شركة المحاصّة.

د- شركة المساهمة.

هـ- الشركة ذات المسؤولية المحدودة.

NLP Computational Tools

	spaCy	NLTK	CoreNLP
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
مبيقولهاش	*mbyqwlhAš*	≈ 26,000
ما بيقولهاش	*mA byqwlhAš*	≈ 13,000
مابقلهاش، مبقولهاش، مبقلهاش، ما بقلهاش، مابيقولهاش	*mAbqlhAš, mbqwlhAš, mbqlhAš, mA bqlhAš, mAbyqwlhAš*	≤ 10,000
مابقولهاش، ما بقولهاش، مبيقلهاش، ما بيقلهاش	*mAbqwlhAš, mA bqwlhAš, mbyqlhAš, mA byqlhAš*	≤ 1,000
مبئلهاش، مابيئولهاش، ما بيئولهاش، مابيؤلهاش	*mbŷlhAš, mAbyŷwlhAš, mA byŷwlhAš, mAbyŵlhAš*	≤ 100
ما بيؤلهاش، مابئلهاش، مبيئولهاش، ما بيئلهاش، مابئولهاش، ما بئلهاش، ما بؤلهاش، مبئولهاش، مبيؤلهاش، مابؤلهاش، مبؤلهاش	*mA byŵlhAš, mAbŷlhAš, mbyŷwlhAš, mA byŷlhAš, mAbŷwlhAš, mA bŷlhAš, mA bŵlhAš, mbŷwlhAš, mbyŵlhAš, mAbŵlhAš, mbŵlhAš*	≤ 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، وسنقول، **وسنقولها** ....

وسنقولها

و+س+ن+قول+ها

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 سوف + s س	b ب	H ح	h هـ + H ح
Progressive Particle	X	qAEd قاعد + jAls جالس	qAEd قاعد	b ب
Possessive Particle	X	Hq حق	Hq حق + tbE تبع	btAE بتاع

الآن الحين دحين دلوقتي

Arabic NLP Common Preprocessing

- Remove diacritics
  - Fathah ـَ damma ـُ kasra ـِ
- Normalize inconsistently typed characters
  - ة (ta marbota) → ـهـ h
  - Drop hamza أ إ آ → ا
- Remove repeated characters and Kashida (tadweel)
  - هلاااا → هلا
  - هـــلا → هلا

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 vise versa
  - Includes diacritics (harakat) and other characters

Arabic	ا	ب	ت	ث	ج	ح	خ	د	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	ف	ق	ك	ل	م	ن	ه	و	ي	ى
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	َ	ِ	ُ	ً	ٍ	ٌ	ّ	ْ	ة	ء	أ	إ	ؤ	ئ	آ	ى
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 تلقى 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 → ال + مدرس + ات
- 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)

MADAMIRA مداميرا
Arabic Analyzer المحلل الآلي للغة العربية

- 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?