Arabic Language Understanding Tools

Mustafa Jarrar
SinaLab, Birzeit University
Palestine

Artificial Intelligence

Opportunity to invest in **Language** as **industry** and **sustainable R&D**

The 4th Revolution

Opportunity for developing countries to compete and impact

We build tools and **resources** for NLU

Lexical Resources at SinaLab - Birzeit University

Lexicographic Database

150 lexicons
Largest Arabic lexicographic database

Arabic Ontology/ Wordnet

Formal Arabic Wordnet
with ontologically clean content

Annotated Corpora

كراسات
Currasat
Arabic Dialects Corpora
مدونة اللهجات العربية

Dialects,
NER, WSD, synonyms
Intents, hate
....

NLP library

APIs
Linguistic Data, synonyms, Nested NER, intents, ...

Synonyms 90s% WSD 84% NER 90s% Intent 88.4% Offensive 88.4% .....

Big Linguistic Data Graph

https://ontology.birzeit.edu

developed in

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

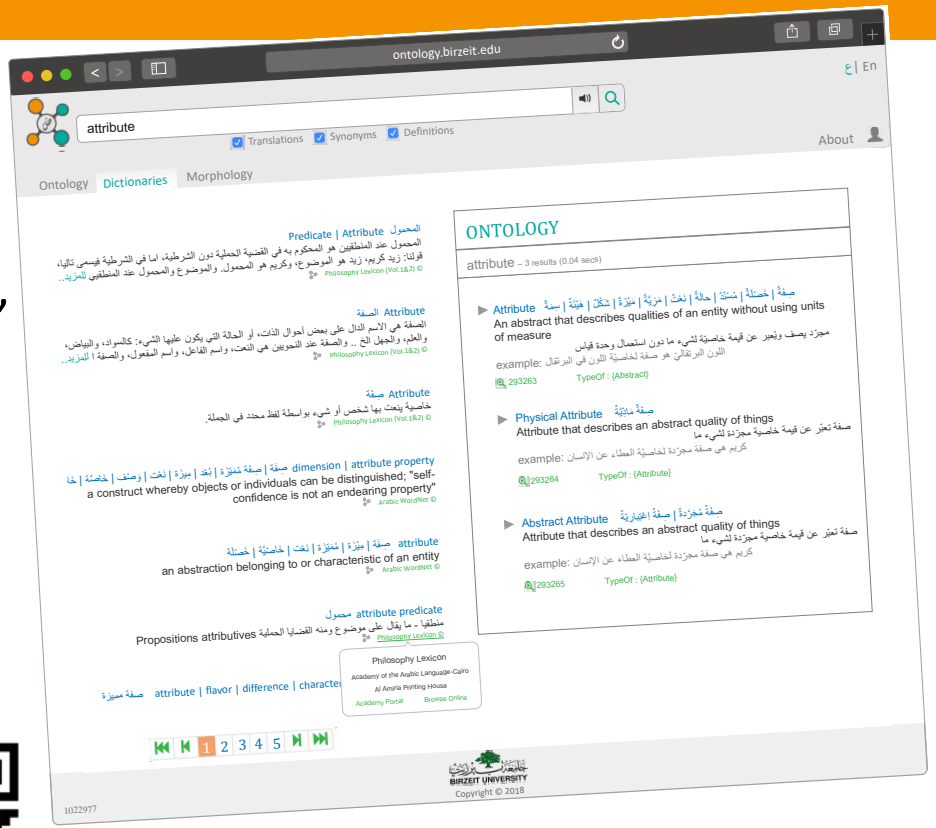
Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) — حوسبة المعاجم
- + **Arabic Ontology** — الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** — كراسات مدونة العاميات
- + **Arabic Synonyms** — استخراج مترادفات
- + **Named Entity Recognition (Wojood)** — وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** — سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** — تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** — خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

The Lexicographic Database

- **The largest lexicographic Arabic database (150 Lexicons)**
- **Contains most lexicon types**: glossaries, thesauri, bi/trilingual dictionaries, morph datasets, **Arabic Ontology**, and more.
- **Covers most domains:** science, technology, law, business, art, philosophy, ...
- Digitized over 10 years.



https://ontology.birzeit.edu

Some Statistics

150 Lexicons

Category	Lexical Concepts	Lexical entries	Synsets	Translations pairs	Glosses	Semantic relations
Total (Millions)	1.1 M	2.4 M	1.8 M	1.5 M	0.7 M	0.5 M
Sub Counts		1,100K Arabic 1,100K English 200K French 3K Others 1,300K Single-word 1,000K Multi-word	800K Arabic 800K English 200K French 50K Others	1,000K English-Arabic 300K English-French 200K French-Arabic	400K Arabic 300K English 1K Others	170K Sub-super links 29K Part-of links 260K Has-Domain links 30K Other links

- **API accessible** for NLP applications.

Mustafa Jarrar, Hamzeh Amayreh: **An Arabic-Multilingual Database with a Lexicographic Search Engine**. NLDB 2019. Pages(234--246), LNCS 11608, Springer. 2019.

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) حوسبة المعاجم
- + **Arabic Ontology** الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** كراسات مدونة العاميات
- + **Arabic Synonyms** استخراج مترادفات
- + **Named Entity Recognition (Wojood)** وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

BIRZEIT UNIVERSITY

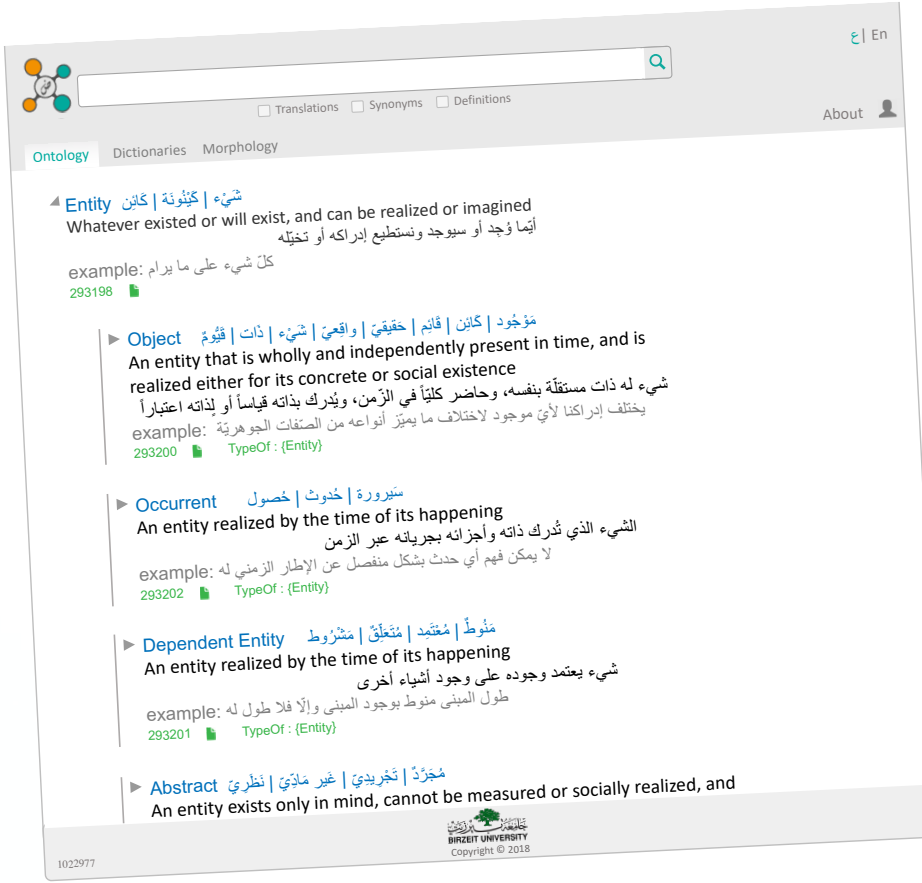
Arabic Ontology

- Characterization of the intended meaning (i.e., concepts) that the Arabic terms convey.

- Formal Arabic Wordnet - with ontologically-clean content.

➢ Linked with **PWN**, **Wikidata**, BFO, DOLCE

➢ Linked with many lexicons



https://ontology.birzeit.edu/concept/293198

Arabic Ontology

- Current size so far (but the numbers are dynamic)

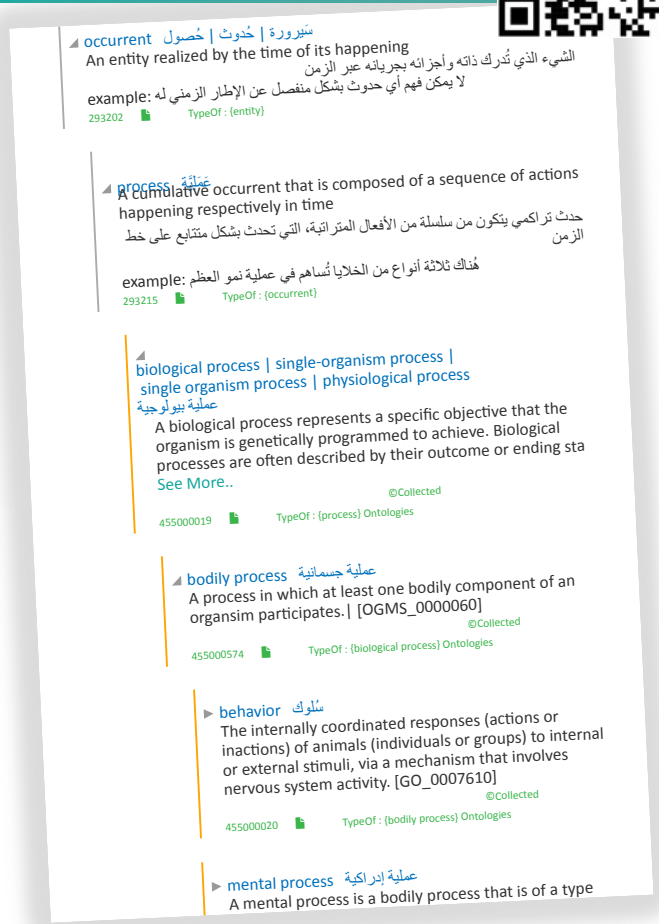
  1800 fully-done concepts (mostly top levels)

  17K partially investigated (ready for NLP applications)

  Some branches are elaborated, other not yet.

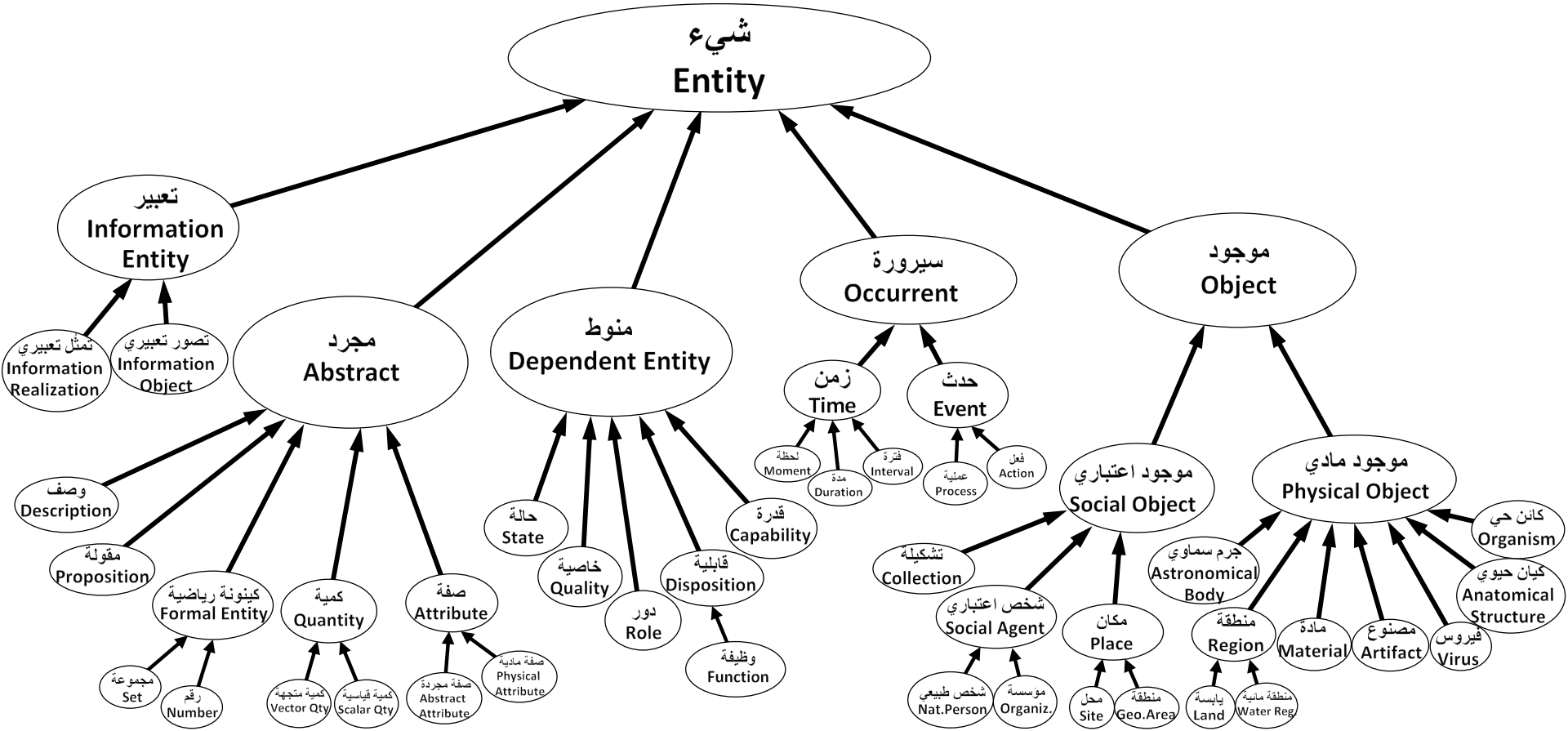
- English labels are not our target - provided for readability and communication.

- Methodology: Built top-down and bottom-up at the same time.



https://ontology.birzeit.edu/concept/293198

Top Levels of the Arabic Ontology



Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) — حوسبة المعاجم
- + **Arabic Ontology** — الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** — كراسات – مدونة اللهجات العربية
- + **Arabic Synonyms** — استخراج مترادفات
- + **Named Entity Recognition (Wojood)** — وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** — سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** — تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** — خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

Arabic Dialects Corpora

حوسبة اللهجات العربية

Currasat كراسات

Arabic Dialects Corpora

مدونة اللهجات العربية

Seven morphologically-annotated Arabic dialect corpora (**1.35 million** tokens)

Curras2 كُراس: Palestinian dialect corpus (56K tokens)

Baladi بَلَدي : Lebanese dialect corpus (10K tokens)

Nabra نَبرة : Syrian dialect corpus (60K tokens)

Lisan لِسان: Yemeni, Iraqi, Libyan, and Sudanese dialects corpora

Yemeni (1.2 million), Iraqi (46K tokens), Libyan(52K tokens), Sudanese(53K tokens)

DPPA

AMERICAN UNIVERSITY OF BEIRUT

جامعة بيرزيت BIRZEIT UNIVERSITY

- El Haff, K., Jarrar, M., Hammouda, T., Zaraket, F., (2022). **Curras + Baladi: Towards a Levantine Corpus**. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.
- Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: **Nâbra: Syrian Arabic Dialects with Morphological Annotations.** In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
- Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: **Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations.** The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(-). IEEE. Egypt. 2023

Try and download

كراسات
Currasat
Arabic Dialects Corpora
مدونة اللهجات العربية

هيك

Search

ع | EN

✔ Word ✔ Stem ✔ Lemma ✔ Gloss
◉ Whole Word ○ Substring
✔ Palestinian ✔ Lebanese ✔ Iraqi ✔ Libyan ✔ Sudanese ✔ Yemeni

About Publications

313 results (4.6 secs)

Gloss	Lemma	POS	Suffix	Stem	Prefix	Word	Context
this way, this, thus	هيك / هكذا	ضمير اشارة		هيك	و/عطف	وهيك	اللي اكلناه , عزمونا الاسبوع الماضي , عشان نردلهم العزيمة , وهيك بدك تحطيلهم ترترسي
this way, this, thus	هيك / هكذا	ضمير اشارة		هيك	و/عطف	وهيك	وهي قاطعا بحقلة قمح ، صارت عجيبة خلت القمحات يطول حشيش وهيك قدرت تتخبى بيناتن ، ومن هون صرنا نعمل القمحية بإيامنا .
this way, this, thus	هيك / هكذا	ضمير اشارة		هيك	و/عطف	وهيك	بيغنوا : " بلاطة فوق بلاطة صاحبة البيت ضراطة ! " وهيك منكون وصلنا لآخر جولتنا بعالم عيد البربارة .
inon+like_that;thus	هيك	تعجب		هيك	ب/حرف جر	بهيك	يمشون وياه الناس ولاكن بيهم شي اشلون اندز هم يمتحنون بهيج مكان ياوزير ياعار
inon+like_that;thus	هيك	تعجب		هيج	ب/حرف جر	بهيج	يجوز ياخذج يريد يخرع بيج جهاله
this way, like this, this, thus	هيك	ضمير اشارة		هيك		هيك	وين يعرف يسولف ويدبر هذا الحجي اكيد واحد كايله احجي هيج
inon+like_that;thus	هيك	تعجب		هيج	ب/حرف جر	بهيج	ولله حيرتو العالم ولعبتو بيها طوبه كل بيج اينزل كلام اليعجبه وينشر

BIRZEIT UNIVERSITY

In cooperation with: AMERICAN UNIVERSITY OF BEIRUT United Nations

https://sina.birzeit.edu/currasat/

Try and download

كراسات Currasat
Arabic Dialects Corpora
مدونة اللهجات العربية

write | Search

ع | EN

✔ Word ✔ Stem ✔ Lemma ✔ Gloss
◉ Whole Word ○ Substring
✔ Palestinian ✔ Lebanese ✔ Iraqi ✔ Libyan ✔ Sudanese ✔ Yemeni

About Publications

313 results (4.6 secs)

Gloss	Lemma	POS	Suffix	Stem	Prefix	Word	Context
be written/be fated/be destined | write	كتَبَ	فعل مضارع مفرد، مذكر، متكلم		كتب	و/عطف+ا/للمضارع المتكلم المفرد	واكتب	وراح اكون هدية واكتب وصية
be written/be fated/be destined | write	كتَبَ	فعل مضارع مفرد، مذكر، مخاطب		كتب	و/عطف+ب/أداة مضارعة+ت/للمضارع المخاطب المذكر المفرد	وبتكتب	وتقعد على الفيس وبتكتب
be written/be fated/be destined | write	كتَبَ	فعل ماضي مفرد، مذكر، مخاطب	ت/للماضي: فاعله مخاطب مذكر مفرد	كتب	و/عطف	وكتبت	وكتبت بوصيتك هو السبب
be written/be fated/be destined | write	كتَبَ	فعل مضارع مفرد، مذكر، غائب	ل/حرف جر+ها/ضمير متصل للغائب المؤنث المفرد	كتب	و/عطف+ي/للمضارع الغائب المذكر المفرد	ويكتبلها	بس شوفت انو قبل كم يوم كان يحطلها أغاني ويحكي عنها ويكتبلها . . . يا الله كيف
be written/be fated/be destined | write	كتَبَ	فعل مضارع مفرد، مذكر، غائب		كتب	ي/للمضارع الغائب المذكر المفرد	يكتب	الابن : (يكتب ستاتوس ع الفيس) لا تتحدثوا عن الوجع اكثر الاشياء وجعا
be written/be fated/be destined | write	كتَبَ	فعل مضارع مجهول مفرد، مؤنث، غائب		نكتب	ت/للمضارع الغائب المذكر المفرد	تتكتب	، وفرجيت كيف هاللغَه تطوّرت من أنجق لهجِه ما كان بِيسْوَا تِتْكتب للغَه عالميّه ما بِيسْوَا ما تِتْكتب .
be written/be fated/be destined | write	كتَبَ	فعل مضارع مجهول مفرد، مؤنث، غائب		نكتب	ت/للمضارع الغائب المذكر المفرد	تتكتب	أنجق لهجِه ما كان بِيسْوَا تِتْكتب للغَه عالميّه ما بِيسْوَا ما تِتْكتب .
scatter/sprinkle/write in prose	نثَرَ	فعل مضارع مفرد، مذكر، غائب		نثر	ب/أداة مضارعة	بنثر	وبنثر رماد
they (people) + write + [masc.pl.]	كَتَبَـ	فعل مضارع	وا/للمضارع: فاعله مذكر جمع	كُتُب	يَ/للمضارع الغائب المذكر المفرد	يَكْتُبوا	و الكل يتكلم عن السفن الجرف الأجنبية و كل الصيادين يبغوا يكتبوا

BIRZEIT UNIVERSITY

In cooperation with: AMERICAN UNIVERSITY OF BEIRUT United Nations

https://sina.birzeit.edu/currasat/

Annotation Tools

SinaLab

Tawseem Portal

بوابة سينا لتوسيم المدونات

Tymaa | Rol | دخول

Loggin 14:26:35 (118)

Stat Log

Lemmas | Annotations

بصير

اسم ☑ فعل ☑ وظيفية ☑

Count: 2

ب+صير صار صارَ
فعل مضارع مفرد مذكر غائب
ب/أداة مضارعة
become/begin to

ب+صير صار صارَ
فعل مضارع مفرد مذكر متكلم
ب/أداة مضارعة
become/begin to

Dialect	Context	Token	Prefix	Stem	Suffix	POS	Number	Gender	Person	DA Lemma	MSA Lemma	Gloss
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	شلون		شلون		أداة استفهام				شلُوْن	كَيْفَ 1	how
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	دا		دا		أداة استقبال				رَحْ | رَاحْ3	سَوْفَ 1	will/shall
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	تدخلي	ت/للمضارع المخاطب المؤنث المفرد	دخل	ي/للمضارع: فاعله مخاطب مؤنث مفرد	فعل مضارع	مفرد	مؤنث	مخاطب	دَخَلْ 1	دَخَلْ 1	enter
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	تسلمي	ت/للمضارع المخاطب المؤنث المفرد	سلم	ي/للمضارع: فاعله مخاطب مؤنث مفرد	فعل مضارع	مفرد	مؤنث	مخاطب	سَلَّم 1	سَلَّم 1	hand over/surren
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	عالنفسا	ع/حرف جر +ال/أداة تعريف	نفسا		اسم	مفرد	مؤنث		نَفْساء 1	نَفْساء 1	in childbed
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	..		..		علامة ترقيم				..	..	..
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	ما		ما		أداة نفي				ما 2	ما 2	not
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	بصير	ب/أداة مضارعة	صير		فعل مضارع	مفرد	مذكر	غائب	صارَ	صارَ	become/begin to
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	يالله	ي/أداة نداء	الله		اسم	مفرد	مذكر		الله	اَللهُ	to/for + Allah/God +
حلبية	شلون دا تدخلي تسلمي عالنفسا ... ما بصير يالله خيتو	خيتو		خيتو		اسم		مؤنث	-	أُخْت 1	أُخْت 1	sisters | sister/count

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

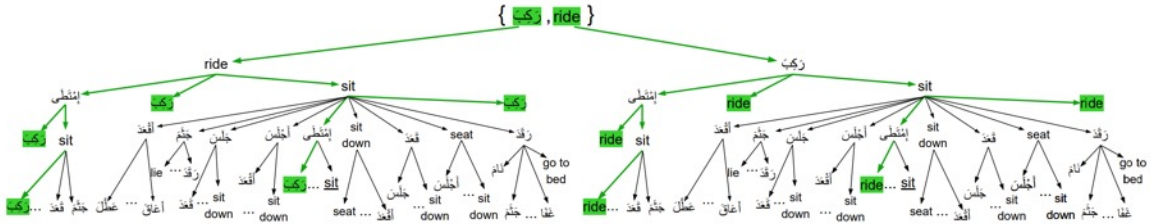
Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) حوسبة المعاجم
- + **Arabic Ontology** الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** كراسات مدونة العاميات
- + **Arabic Synonyms** استخراج مترادفات
- + **Named Entity Recognition (Wojood)** وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

Extracting Synonyms from lexicographic graphs

A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms

Sana Ghanem[1], Mustafa Jarrar[1], Radi Jarrar[1], Ibrahim Bounhas[2,3]
[1]Department of Computer Science, Birzeit University, Palestine
[2]LISI Laboratory of Computer Science for Industrial System, INSAT, Carthage University, Tunisia
[3]JARIR: Joint group for Artificial Reasoning and Information Retrieval, Tunisia
{swghanem, mjarrar, rjarrar}@birzeit.edu
ibrahim.bounhas@isd.uma.tn



A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms

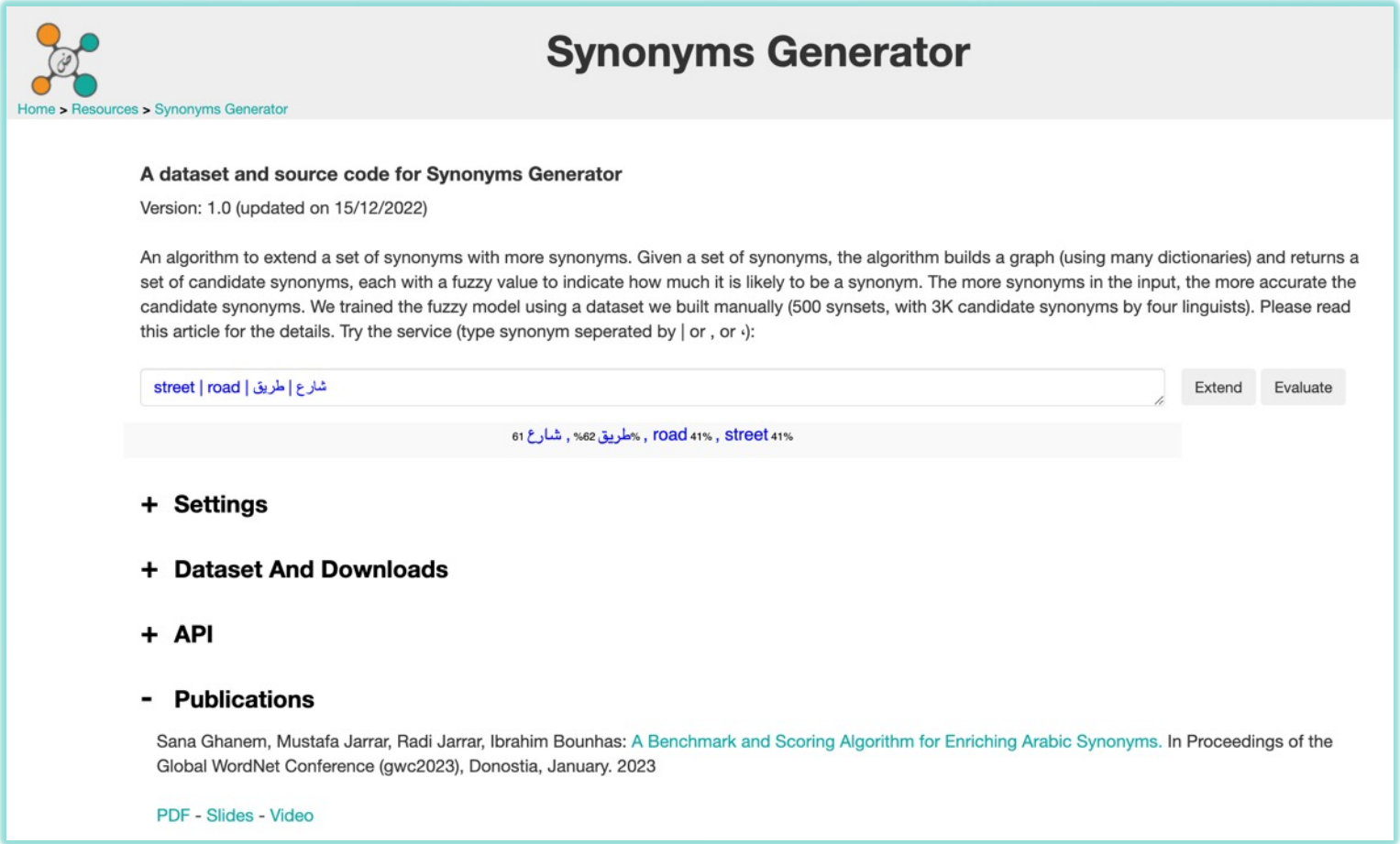
Dataset Construction

- Four Linguists (500 synsets, 3K synonyms in total).
- Each candidate synonyms was annotated with a fuzzy value by four different linguists.

تَحَالُف | حِلْف | اِتِّحَاد فِدْرَالِي | confederacy | confederation | federation		
a union of political organizations		
مُحَالَفَة	60 نفس الدلالة، الأسلوب ضعيف ، غير شائعة	
اِئْتِلاف	80 نفس الدلالة، الأسلوب صحيح ، شائعة الى حد قليل	
اِتِّحَاد	100 نفس الدلالة والاسلوب والشيوع	
جامِعَة	60 نفس الدلالة، الأسلوب ضعيف ، غير شائعة	

Score	Meaning
100	Same semantics, style, use
90	Same semantics, style, less used
80	Same semantics, style, rarely used
70	Same semantics, style, not used
60	Close semantics, weak style, uncommon
50	Close semantics, not exact purpose
40	Semantically related
30	Semantically related (somehow)
20	Semantically different
10	Semantically very different
0	Semantically unrelated

Demo



https://sina.birzeit.edu/synonyms/

Download: [Github.Synonyms]

A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms

Extend a synset with more synonyms above a given fuzzy value

A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms

Evaluate/**rank** synonyms in a given synset (wordnets, BERT's output, etc.)

street | road | شارع | طريق

Evaluate

61 شارع , %62 طريق% , road 41% , street 41%

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) حوسبة المعاجم
- + **Arabic Ontology** الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** كراسات مدونة العاميات
- + **Arabic Synonyms** استخراج مترادفات
- + **Named Entity Recognition (Wojood)** وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

Wojood

Nested Arabic Named Entity Corpus

Wojood

Nested Arabic Named Entity Corpus

Flat NER

Sami	works	at	Jimmy	Carter	Center
B-PERS	O	O	B-ORG	I-ORG	I-ORG

✓mature

Nested NER

Sami	works	at	Jimmy	Carter	Center
B-PERS	O	O	B-PERS	I-PERS	
			B-ORG	I-ORG	I-ORG

!

Jarrar, M., Khalilia, M., Ghanem, S. (2022). **Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT**. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France.

Wojood

❖ NER Corpus

Corpus	Nested?	size (tokens)	No. of entities	Entity classes	Which Arabic	Domain
Ontonotes5	No	300k	28k	18	MSA	News
ANERCorp	No	150k	11k	4	MSA	News
Canercorpus	No	258k	72k	14	Classic	Religion
AQMAR	No	74k	-	open	MSA	4 domains
Wojood Corpus	YES	550K	75K	21	MSA & Dialect	**Multi domains** Media, History, Culture, Health, Finance, ICT, Law, Elections, Politics, Migration, Terrorism

❖ NER BERT

- Multi-task learning (nested entities)
- 88.4% F1-score

Annotation Guidelines

21 entity classes

PERS People names
فيروز، عادل إمام، ابن احمد، الملك عبدالله، النبي محمد

NORP Group of people
العرب، المسحيين، سكان القدس، وزراء الخارجية العرب

OCC Occupation or professional title
رئيس جامعة بيرزيت، مدير بنك فلسطين، قائد الجيش

ORG Legal/social body
بنك القاهر، ريال مدريد، داعش، الجيش المصري،

GPE Geopolitical: country, city, state
ليبا، مدينة القدس، الجمهورية اللبنانية، روسيا الاتحادية

LOC Geographical location (non-GPE)
البحر الميت، قناة السويس، آسيا، الوطن العربي

FAC Places: landmark, road, building..
مطار صنعاء، سجن ابو غريب، المسجد الأقصى

EVENT Events of general interest
حرب 1973، القمة العربية 2005، عيد الفطر ، يوم الأرض

DATE Specific/relative date (>day)
13يونيو، 2019-2020، الفترة العثمانية

TIME Specific/relative time (<day)
الساعة ١٢، من الساعة 5 حتى 7 مساء، خلال ساعتين

LANGUAGE Human language or dialect
اللغة العربية، الفصحى، الدارجة المغربية، اللغة الفرنسية

WEBSITE Website or specific URL
موقع فيسبوك، يوتيوبschema.org ،

LAW Geographical location (non-GPE)
قانون الاستثمار ، المادة 114 من قانون العقوبات 2005

PRODUCT Vehicle, weapon, food, ...
مرسيدس سي١٨٠، ايفون ١٣، دبابة مركابا، تروفين

CARDINAL Numerals in digits/words
، 1.5 ، 30 ، 150صفر، اثنان ، أربعة وعشرون ، مليون

ORDINAL does not refer to a quantity
٣ كيلومتر، مئة قدم ، 3 طن ، 50 غرام ، 25 سم مربع

PERCENT Word/symbol refers to a percent
5بالمئة ، 10% ، 9 من كل الف

QUANTITY Value measured by units
٣ كيلومتر، مئة قدم ، 3 طن ، 50 غرام ، 25 سم مربع

UNIT Name/symbol of a unit
ميل ، كيلو ، كيلومتر ، إنش ، كيلوغرام ، هكتار ، مل

MONEY Monetary quantity, incl. currency
مئة وخمسون درهم اماراتي ، اثنان وثلاثون يورو ، 8 دولار

CURR Name/symbol of currency
دولار ، جنيه مصري ، دينار ، فرنك ، ريال سعودي€ ,

Corpus Collection

Source - Topics	Sentences	Tokens
Web Articles[1] (MSA) Health, Finance, ICT, Law, Elections, Politics, Migration and Terrorism	9,053	258,102
Archive[2] (MSA) History and Culture	12,271	227,020
Social Media[3] (Dialect) General topics	5,653	65,342
Total	26,977	550,464

1 un.org, hrw.org, msf.org, who.org, mipa.institute, elections.ps, sa.usembassy.gov, diplo- matie.ma, quora.com ….
2 Awraq, Birzeit University Digital Palestinian Archive
3 Palestinian and Lebanese dialect corpora

Downloads and Demo

Birzeit University, in cooperation with the Edward Said Foundation, is organizing a folklore festival. The festival will start at 4:00 pm, on May 16, 2016, with the sponsorship of Bank of Palestine for an amount of five thousand dollars.

جامعة بيرزيت وبالتعاون مع مؤسسة ادوارد سعيد تنظم مهرجان للفن الشعبي، سيبدأ المهرجان الساعة الرابعة عصرا، بتاريخ 16/5/2016، وذلك برعاية من بنك فلسطين بمبلغ خمسة الاف ولار.

output format : highlighted

جامعة بيرزيت (ORG GPE) وبالتعاون مع مؤسسة ادوارد سعيد (ORG PERS) تنظم مهرجان للفن الشعبي، (EVENT)

سيبدأ المهرجان الساعة الرابعة عصرا، (TIME) بتاريخ 16/5/2022, (DATE) وذلك برعاية من بنك فلسطين (ORG GPE)

بمبلغ (ORG) خمسة الاف دولار (MONEY CURRENCY)

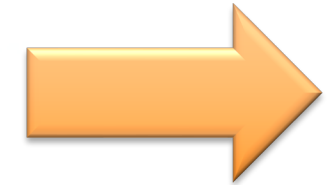
https://ontology.birzeit.edu/wojood

ArabicNER

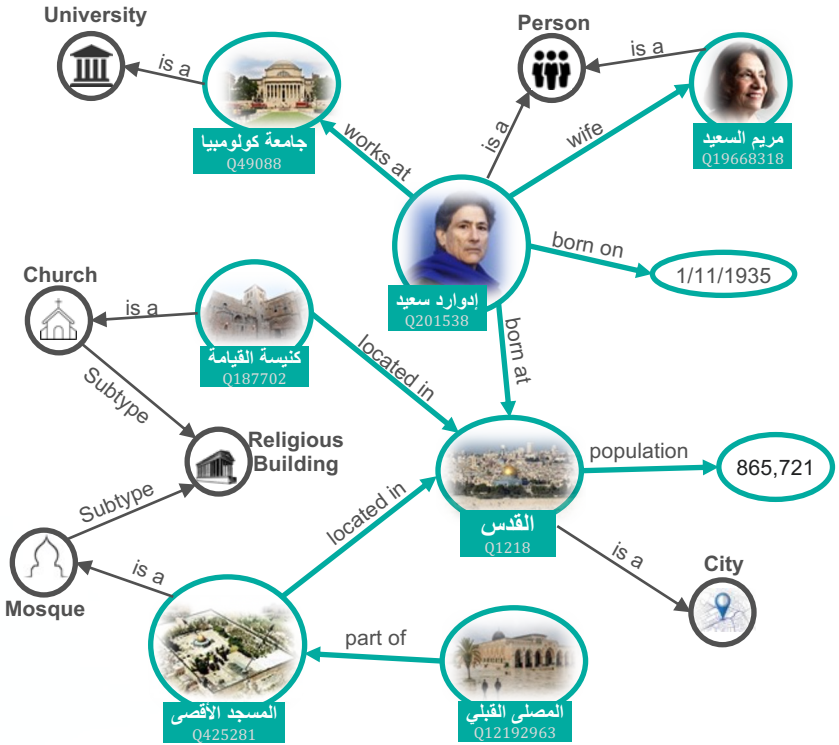
ArabicNER-Wojood

Text to Intelligence

Text



Knowledge Graph



What Is Named Entity Recognition (NER) Used For?

Customer Support
Recognize and route per products, models, places, phones,

Content Recommendation
understand use profiles and history

Resumes
Recognize name, address, phone, etc.

Chatbots and Question Answering
Recognize names, amounts, places, etc.

Machine Translation
Transliterate entities….

Event Detection
Identity event as they occur (Finance/business events, social media, ….)

Knowledge Management
knowledge graphs, Digital libraries, content filtering

Classifying content for news providers

Efficient Search Algorithms

Finance
extract figures, loans, credit risk.

News providers

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) — حوسبة المعاجم
- + **Arabic Ontology** — الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** — كراسات مدونة العاميات
- + **Arabic Synonyms** — استخراج مترادفات
- + **Named Entity Recognition (Wojood)** — وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** — سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** — تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** — خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

Arabic Word-Sense Disambiguation

ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD

Moustafa Al-Hajj
Lebanese University
Beirut, Lebanon
moustafa.alhajj@ul.edu.lb

Mustafa Jarrar*
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

Context-Gloss Augmentation for Improving Arabic Target Sense Verification

Sanad Malaysha, Mustafa Jarrar, Mohammed Khalilia
Birzeit University, Palestine

SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks

Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammed Khalilia
Birzeit University, Palestine
{mjarrar, smalaysha, thammouda, mkhalilia}@birzeit.edu

SALMA
ArabGlossBERT for Arabic Word-Sense Disambiguation

The Word Sense Disambiguation (WSD) Task

Given a word in a context, which sense (i.e. meaning) this word denotes?

You have beautiful eyes

Set of senses/meanings

1. the organ of sight
2. good discernment (either visually or as if visually)
3. attention to what is seen
4. an area that is approximately central within some larger region)
5. a small hole or loop

Al-Hajj, M., Jarrar, M., (2021). ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In Proceedings – the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), PP 40--48.

SALMA
ArabGlossBERT for Arabic Word-Sense Disambiguation

The Word Sense Disambiguation (WSD) Task

Given a word in a context, which sense (i.e. meaning) this word denotes?

قصيدة من عيون الشعر

Set of senses

1. **عُضو الإبصار** في الإنسان والحيوان: له عينان كعَيْنَيْ الصقر - ألا إنّما العينان للقلب رائدٌ ...
2. **جاسوس**، "كان عينًا لدولةٍ أجنبيَّة . بثَّ العيون : تجسَّس، راقب - فلانٌ عَيْن على فلان : ناظر عليه
3. **أجود كلّ شيء** وأحسنه ونفيسُه: عيونُ الفنِّ.
4. **حارس:** فلان عين على المكان.
5. **الحاضر من كل شيء** أصبح أثرًا بعد عين ...
6. **عَيْنُ الماء**:- ينبوعه، تُحلِّق الطّيورُ فوق عيون الماء
7. **عَيْن الشَّيء**:- نفسه، ذاته (تستعمل للتوكيد): جاء القوم أعينهم - كنّا في المكان عينه.
8. **عَيْن العقل**:- قدرة ذهنيَّة مَوْروثة على التخيُّل وتذكّر الأحداث.
9. .....

Al-Hajj, M., Jarrar, M., (2021). ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In Proceedings – the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), PP 40--48.

SALMA
ArabGlossBERT for Arabic Word-Sense Disambiguation

❖ Arabic context-gloss pairs Dataset (167k)

- Extracted from Birzeit University's Lexicographic database
- Annotated target words in context;

Gloss	Context	Label
[SEP] أجود كل شيء وأحسنه ونفيسه	[CLS] قصيدة من عيون الشعر	True
[SEP] عين الشيء : نفسه ، ذاته (تستعمل للتوكيد)	[CLS] قصيدة من عيون الشعر	False
[SEP] عين الشيء : نفسه ، ذاته (تستعمل للتوكيد)	[CLS] جاء القوم أعينهم	True
[SEP] أجود كل شيء وأحسنه ونفيسه	[CLS] جاء القوم أعينهم	False

❖ Three Fine-tuned BERT Models

- WSD into **binary sequence-pair classification task**
- Accuracy 84%
- 4 types of signals to emphasize target words in context

Fine-Tuning

- Fine-tuned three Arabic pre-trained BERT models

The WSD task is converted into **binary sequence-pair classification task**

True [CLS] قصيدة من عيون الشعر [SEP] أجود كل شيء وأحسنه ونفيسه [SEP]

False [CLS] قصيدة من عيون الشعر [SEP] عين الشيء : نفسه ، ذاته (تستعمل للتوكيد) [SEP]

True [CLS] جاء القوم أعينهم [SEP] عين الشيء : نفسه ، ذاته (تستعمل للتوكيد) [SEP]

False [CLS] جاء القوم أعينهم [SEP] أجود كل شيء وأحسنه ونفيسه [SEP]

- Results

Model		True	False	Accuracy
AraBERTv02	Precision	81	85	84
	Recall	66	93	
	F1-score	72	89	
CAMeLBERT	Precision	77	83	82
	Recall	60	92	
	F1-score	67	87	
QARiB	Precision	73	82	80
	Recall	58	90	
	F1-score	65	86	

Constructing a dataset of context-gloss pairs

Statistics:

	count
Unique Lemmas (undiacritized)	26169
Avg glosses per Lemmas	1.25
Unique Glosses	32839
Unique Contexts	60272
Avg context per gloss	1.83
True context-gloss pairs	60323
False context-gloss pairs	106884
Total True and False pairs	167207

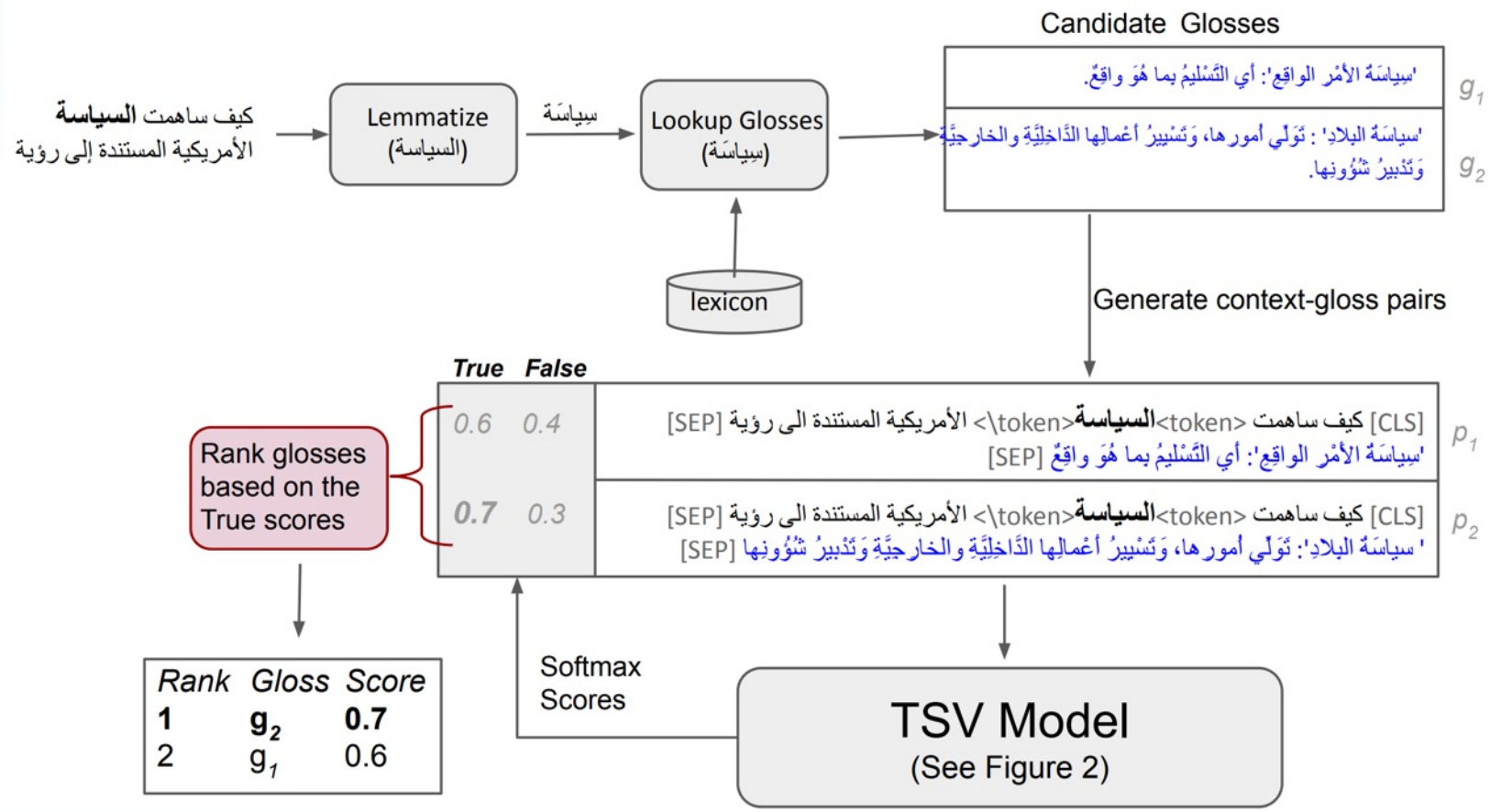
Training and Test Datasets

- every context selected in the test set should not be selected in the training set;
- every gloss should be selected in both the training and the test sets.

Datasets	Pairs	Count	Total
Training	True pairs	55,585	
	False pairs	96,450	152,035
Test	True pairs	4,738	
	False pairs	10,434	15,172
		Total	167,207

Download: https://ontology.birzeit.edu/downloads

SALMA
End-to-end WSD system (using TSV)



Try

https://sina.birzeit.edu/salma/

SinaLab

News Team Resources

SALMA سلمى

A corpus and model for Arabic Word Sense Disambiguation (WSD).
Version: 1.0 (updated on 22/10/2023)

قصيدة من عيون الشعر

WSD

◂ **قصيدة** (قَصِيدَة 1_303044571): مجموعة من الأبيات الشِّعريّة متّحدة في الوزن والقافية والرَّويّ وهي تتكوّن من سبعة أبيات فأكْثر "قصيدة غزليّة ".

بَيْتُ القصيدة : البيت المتضمّن غاية الشّاعر، أو أنفس أبياتها، أو مثل يُضرب في تفضيل بعض الشَّيء على كُلّه - مطلع القصيدة : أوّل بيت منها

◂ **من** (مِن 1)

◂ **عيون** (عَيْن 2_303038477): أجود كلّ شيء وأحسنه ونفيسُه "قصيدة من عيون الشّعر - عيونُ الفنِّ ".

◂ **الشعر** (شِعْر 1_303029103): كلام موزون مقفّى قصدًا يعتمد على التخييل والتأثير؛ ليوحي بإحساسات مؤثِّرة وصور خياليّة "شعر صافي الديباجة

- نظم الشِّعْرَ - ما الشِّعْر إلاّ شعورُ المرء يُرسلُه ... عفو البديهة عن صدق وإيمانِ - إنَّ مِنَ الشِّعْرِ لَحِكْمَةً [حديث] - حوَمَا عَلَّمْنَاهُ الشِّعْرَ وَمَا يَنْبَغِي لَهُ> يس/

69 ". أنشده الشِّعرَ : قرأه عليه - أوابِدُ الشِّعْر : ما لا تُماثَل جودتُه أو قوافيه الشاردة - ربَّة الشِّعْر : إلهة الشِّعْر عند الوثنيِّين - شطرا بيت الشِّعْر : الصدر

+ Description

+ Downloads

ArabGlossBERT
Arabic Word-Sense Disambiguation

https://sina.birzeit.edu/salma/

ArabGlossBERT
Arabic Word-Sense Disambiguation

https://sina.birzeit.edu/salma/

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) حوسبة المعاجم
- + **Arabic Ontology** الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** كراسات مدونة العاميات
- + **Arabic Synonyms** استخراج مترادفات
- + **Named Entity Recognition (Wojood)** وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

Intent Detection

ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic

Mustafa Jarrar
Birzeit University
Birzeit, Palestine
mjarrar@birzeit.edu

Ahmet Birim
Sestek
Istanbul, Türkiye
ahmet.birim@sestek.com

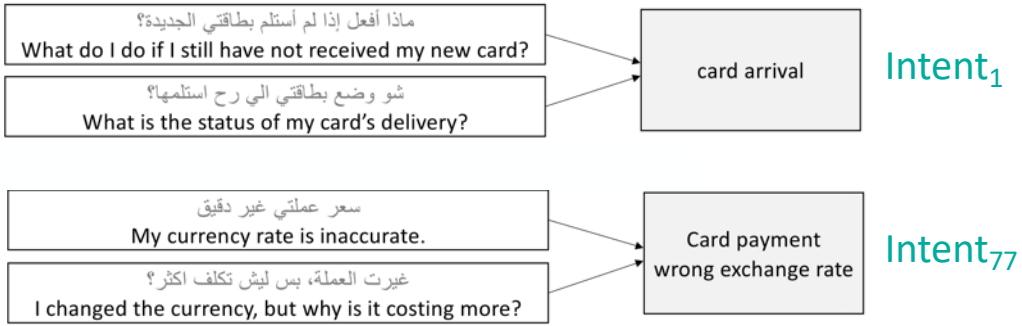
Mohammed Khalilia
Birzeit University
Birzeit, Palestine
mkhalilia@birzeit.edu

Mustafa Erden
Sestek
Istanbul, Türkiye
mustafa.erden@sestek.com

Sana Ghanem
Birzeit University
Birzeit, Palestine
swghanem@birzeit.edu

Chatbots need to detect intents before answering

➔ build an Arabic intent dataset, and train BERT.



Our Contributions

❖ArBanking77 dataset

- ArBanking77 dataset consists of **31,404 queries**.
- 2.4x larger than the Banking77 dataset.
- On average, there are 408 queries per intent

C	D	E	F	G	H	I
Qid	Question_en	Question_MSA1	n_MSA2	Question_PAL1	Question_PAL2	Intent_en
1	I am still waiting on my card?	ما زلت أنتظر بطاقتي؟		بعدني بستنى في البطاقة؟		card arrival
2	What can I do if my card still hasn't arrived after 2 w	ماذا أفعل إذا لم تصل بطاقتي بعد أسبوعين؟	ماذا علي أن أفعل	ايش بلزم اعمل اذا بطاقتي ما وصلت بعد اسبوعين	شو اساوي اذا ما وصلت بطاقتي	card arrival
3	I have been waiting over a week. Is the card still co	انا أنتظر منذ أكثر من أسبوع. هل ما زالت البطاقة قادمة؟		صارلي أكثر من أسبوعين بستنى، لساتها البطاقة م	الي بستنى أكثر من أسبوعين، كم	card arrival
4	Can I track my card while it is in the process of deliv	هل يمكنني تتبع بطاقتي أثناء عملية التسليم؟	أبإمكاني تتبع بطا	بقدر أتتبع بطاقتي وهي بعملية التسليم؟	فيي أتتبع بطاقتي لتوصلني؟	card arrival
5	How do I know if I will get my card, or if it is lost?	كيف أعرف إذا ما كنت سأحصل على بطاقتي ، أو إذا ضاعت	كيف يمكنني أن أ	كيف بقدر أعرف اذا راح أخد بطاقتي ولا ضاعت		card arrival
6	When did you send me my new card?	متى أرسلت لي بطاقتي الجديدة؟	متى بعثت لي بط	متى بعتتلي بطاقتي الجديدة؟	ايمتى بعتت بطاقتي الجديدة؟	card arrival
7	Do you have info about the card on delivery?	هل لديك معلومات عن البطاقة عند التسليم؟	هل هناك معلومات	عندكم معلومات عن بطاقة الدفع عند الاستلام؟		card arrival
8	What do I do if I still have not received my new card	ماذا أفعل إذا لم أستلم بطاقتي الجديدة؟	ما الذي علي فعله	شو أعمل اذا ما اخدت بطاقتي الجديدة؟	ايش بقدر اعمل اذا ما استلمت بط	card arrival
9	Does the package with my card have tracking?	هل يتم تتبع الحزمة مع بطاقتي؟	هل هناك تتبع للح	بقدر أتبع الحزمة يلي بتيجي مع بطاقتي؟		card arrival
10	I ordered my card but it still isn't here	طلبت بطاقتي لكنها ما زالت غير موجودة	طلبت بطاقتي لكن	طلبت بطاقتي بس لساتها مش موجودة.	بطاقتي الي وصيت عليها لساتها	card arrival
11	Why has my new card still not come?	لماذا لم تصل بطاقتي الجديدة؟	لماذا بطاقتي الجد	ليش بطاقتي لساتها ما اجت؟	ليه بطاقتي مش موجودة بعدها؟	card arrival
12	I still haven't received my card after two weeks, is it	ما زلت لم أستلم بطاقتي بعد أسبوعين ، هل فقدت؟	للأن لم أستلم بطا	الي اسبوعين ما استلمت بطاقتي، معقول ضاعت؟	صارلي اسبوعين ما استلمت بطا	card arrival
13	Can you track my card for me?	هل يمكنك تتبع بطاقتي من أجلي؟	أبإمكانك تتبع بطا	ممكن تتبعلي بطاقتي؟		card arrival
14	Is there a way to track the delivery of my card?	هل هناك طريقة لتتبع تسليم بطاقتي؟		في طريقة أتتبع تسليم بطاقتي؟		card arrival
15	It's been a week since I ordered my card and it's no	لقد مر أسبوع منذ أن طلبت بطاقتي وهي ليست هنا. ارجو الم	مر أسبوع على ط	صارلي أسبوع بستنى بالبطاقة، ممكن تساعدوني؟	فيكم تساعدوني بخصوص البطاق	card arrival
16	Will I be able to track the card that was sent to me?	هل سأتمكن من تتبع البطاقة التي تم إرسالها إلي؟	أيمكنني تتبع البط	بقدر أتتبع البطاقة الي انبعتتلي؟	في إمكانية أتتبع البطاقة الي بعتوا	card arrival
17	I don't have my card in 1 week. Should I be worried	منذ أسبوع ولم أستلم بطاقتي خلال أسبوع واحد. يجب أن أكون قلقا؟	لا أملك بطاقتي	الها أسبوع بطاقتي مش معي، لازم أقلق؟	في اشي داعي للقلق عشان بطاقتي	card arrival

❖Intent detection model

- F1-scores: MSA (%92) and PAL (%90)

Download

SinaLab

News Team Resources

ArBanking77

A dataset and source-code for ArBanking77
Version: 1.0 (updated on 1/9/2023)

ArBanking77 consists of 31,404 (MSA and Palestinian dialect) that are manually Arabized and localized from the original English Banking77 dataset; which consists of 13,083 queries. Each query is classified into one of the 77 classes (intents) including card arrival, card linking, exchange rate, and automatic top-up. A neural model based on AraBERT was fine-tuned on the ArBanking77 dataset (F1-score 92% for MSA, 90% for PAL). Try the service (type sentences seperated by newLine or ؟ or ? or ! or .):

يا ريت يتم توقيف البطاقة

Detect

- Downloads

ArBanking77 is available to download upon request for academic and commercial use.
Request to download ArBanking77 (whole dataset 31,404 queries, MSA 15,537 queries, Palestinian Dialect 15,867 queries)
GitHub (download BERT training source code + sample data (~1K queries))
Hugging Face (download fine-tuned BERT model, ready to use)

+ API

- Publications

Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, Sana Ghanem: ArBanking77: Intent Detection Neural Model

BIRZEIT UNIVERSITY

https://sina.birzeit.edu/arbanking77/

Offensive Language Detection in Hebrew

Offinsive content {Hate | Violence | Racist | Abusive} in Hebrew is growing.

Type of Offence:

Class	Sub-Class	Definition
Offensive	Abusive	If the tweet contains direct or implicit insults using vulgar or street words.
	Hate	If the tweet contains criticism, attack, or degrade, directly or implicitly, because of race, color, religion, nationality, or gender.
	Pornographic	If the tweet promotes or invites any pornographic or sexual arousal.
	Violence	If the tweet endorses an act that involves physical harm towards any party, regardless of the reason.

Offensive Language Detection in Hebrew

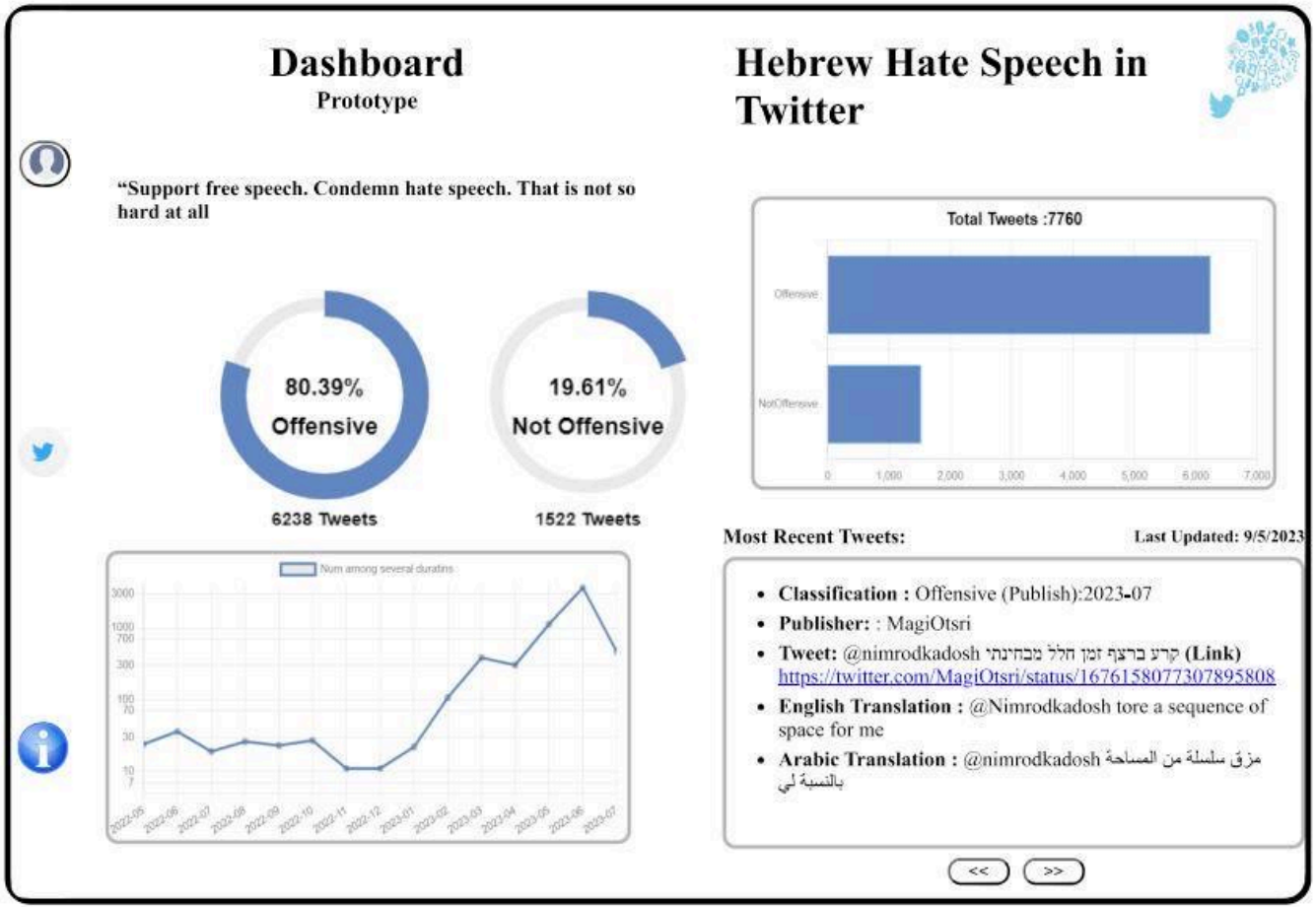
- Collect 16K tweets,
- classify each tweet with different tags,
- then, train a deep learing model – to automatically detect offensive content.

Tweet	English Translation	Annotations
גירוש, הריסת בתים, מאסרי עולם ללא אפשרות חנינה, אחרת יהיה עוד יותר גרוע. להתייחס אליהם כמו אל מחבלים.	Deportation, demolition of houses, life sentences without the possibility of amnesty, otherwise it will be even worse. Treat them like terrorists.	**Class**: Violence, Hate **Target**: Palestinians **Topic**: punish Palestinians **Phrase**: Deportation, Demolition of houses, life sentences, terrorists
אין כבר הרתעה.לא מפחדים מהמשטרה.אני חושב שהגיע הזמן על פי מראות ההפגנות וההתפרעויות בימים האחרונים,כמו שאמר פעם רבין בתחילת האינתיפאדה: „לשבור להם את העצמות".פה יש כבר אינתיפאדה של התפרעויות.	There is no more deterrence. We are not afraid of the police. I think the time has come to face the demonstrations and riots, as Rabin once said at the beginning of the intifada: "to break their bones." There is already an intifada of riots here.	**Class**: Violence, Hate **Target**: Palestinians **Topic**: Demonstrations **Phrase**: Break their bones
@Onetruth011 ימח שמה וזכרה של אילנה דיין. העיתונאית הכי מנוולת ושקרנית שאני מכיר. ממש מרשעת.	May the name and memory of Ilana Dayan be remembered. The most depraved and lying journalist I know. Really sinister.	**Class**: Abusive **Target**: Ilana Dayan **Topic**: Journalism **Phrase**: Sinister, Depraved, Lying
@judash0 פרצופו האמיתי של אבי ביטון נחשף לעיני כל. מדובר בשמאלני, אנטי ציוני, עוכר ישראל, בוגד שממומן ע"י הקרן החדשה להפיל את שלטון הימין ולהעלות את המפלגות הערביות לשלטון כדי להוביל למדינת כל אזרחיה	Avi Beaton's true face clear now. This is a leftist, anti-Zionist, oppressor of Israel, a traitor who is financed to overthrow the right-wing government and bring the Arab parties to power in order to lead to a state for all its citizens.	**Class**: Hate, Abusive **Target**: Avi Bitton, Arab Parties **Topic**: politics **Phrase**: Traitor, Anti-Zionist
@rabea_bader ה. וא לא רלוונטי אם אתה דרוזי, סורי, אנטי ציוני ומגעיל שכמוך.	@rabea_bader is irrelevant if you are Druze, Syrian, anti-Zionist and disgusting like you.	**Class**: Hate, Abusive **Target**: Rabea Bader, Druze, Syrian **Topic**: Racism **Phrase**: disgusting, anti-Zionist
@Ahmad_tibi אתה לפחות לא משקר - היית ונשארת לאומן ערבי שרוצה בחורבן ישראל כמדינה יהודית.	@Ahmad_tibi At least you're not lying - you were and remain an Arab nationalist who wants the destruction of Israel as a Jewish state.	**Class**: Hate **Target**: Ahmad Tibi **Topic**: Political views **Phrase**:

Offensive Language Detection in Hebrew

Data set	# of training examples		Validation data	Test data	Accuracy	
	Our data	D_{OLaH}			He BERT	Aleph BERT
D_1	1,750	0	250 (ours)	500 (ours)	63%	68%
D_2	1,750	1,013	250 (ours)	500 (ours)	58%	63%
D_3	1,750	2,026	250 (ours)	500 (ours)	61%	63%
D_4	0	1,418	203 (D_{OLaH})	405 (D_{OLaH})	79%	**86%**
D_5	1,250	1,418	203 (D_{OLaH})	405 (D_{OLaH})	**81%**	79%
D_6	2,500	1,418	203 (D_{OLaH})	405 (D_{OLaH})	**81%**	82%
D_7	0	2,026	203 (D_{OLaH})	500 (ours)	60%	57%
D_8	2,500	0	250 (ours)	405 (D_{OLaH})	64%	69%

Offensive Language Detection in Hebrew



Offensive Language Detection in Hebrew

Download Corpus and models

SinaLab / OffensiveHebrew

Type / to search

Code | Issues | Pull requests | Actions | Projects | Security

OffensiveHebrew Public

Edit Pins | Watch 0

main | 1 branch | 0 tags | Go to file | Add file | Code

naghamghanim Add files via upload c0fcc0d 2 hours ago 19 commits

Training	Add files via upload	3 months ago
data	Add files via upload	2 months ago
LICENSE	Update LICENSE	3 months ago
README.md	Update README.md	2 months ago
tag_vocab.pkl	Add files via upload	2 hours ago

README.md

Hebrew Corpus

This corpus contains offensive language in Hebrew manually annotated. The data includes 15,881 tweets, labeled with one or more of five classes (abusive, hate, violence, pornographic, or non-offensive). The corpus is annotated manually by Arabic-Hebrew bilingual speakers.

The Corpus

Hugging Face

Search models, datasets, users...

SinaLab / OffensiveHebrew

like 0

License: cc-by-nc-sa-4.0

Use with library

Model card | Files | Community 1 | Settings

Edit model card

Downloads last month
0

Hebrew Corpus

This corpus contains offensive language in Hebrew manually annotated. The data includes 15,881 tweets, labeled with one or more of five classes (abusive, hate, violence, pornographic, or non-offensive). The corpus is annonated manually by Arabic-Hebrew bilingual speakers.

Model Download

Huggingface:

https://huggingface.co/SinaLab/OffensiveHebrew

Natural Language Understanding Tools and Datasets

Open Source

https://sina.birzeit.edu/resources

SinaLab

News Team Resources

Resources

Download and try NLP/NLU datasests, corpora, tools and services

- + **Lexicographic Database** (150 lexicons) حوسبة المعاجم
- + **Arabic Ontology** الأنطولوجيا العربية
- + **Dialect Corpora (Currasat)** كراسات مدونة العاميات
- + **Arabic Synonyms** استخراج مترادفات
- + **Named Entity Recognition (Wojood)** وجود –لاستخراج أسماء الاعلام
- + **Word Sense Disambiguation (Salma)** سلمى – محلل دلالي
- + **ArBanking77 Intent Detection** تحديد المقصود في المساعدات الآلية
- + **Offensive Language Detection** خطاب الكراهية بالعبرية
- + **Lemmatizer**
- + **NLP Tools**

BIRZEIT UNIVERSITY

References

1. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammad Khalilia: SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
2. Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: Nâbra: Syrian Arabic Dialects with Morphological Annotations. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
3. Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem: ArBanking77: Intent Detection Neural Model and a New Dataset in Modern and Dialectical Arabic. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
4. Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, Muhammad AbdulMageed: Arabic Fine-Grained Entity Eecognition. In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023. ACL.
5. Mustafa Jarrar, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim El-madany, Nagham Hamad, Alaa' Omar: WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In Proceedings of the 1st Arabic Natural Language Processing Conference (Arabic- NLP), Part of the EMNLP 2023. ACL.
6. Nouran Khallaf, Elin Arfon, Mo El-Haj, Jon Morris, Dawn Knight, Paul Rayson, Tymaa Hammouda, Mustafa Jarrar: Open-source thesaurus development for under-resourced languages: a Welsh case study. The 4th LDK Conference on Language, Data and Knowledge, Vienna, Austria, 12-15 September 2023
7. Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, Nadim Nashif: Offensive Hebrew Corpus and Detection using BERT. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). IEEE. Egypt. 2023
8. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023
9. Sanad Malaysha, Mustafa Jarrar, Mohammad Khalilia: Context-Gloss Augmentation for Improving Arabic Target Sense Verification. The 12th International Global Wordnet Conference (GWC2023), Global Wordnet Association. (pp.). San Sebastian, Spain, 2023
10. Mustafa Jarrar, Mohammed Khalilia, Sana Ghanem: Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
11. Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations. The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(-). IEEE. Egypt. 2023 arXiv, DOI 10.48550/ARXIV.2212.06468. 2023
12. Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, Fadi Zaraket: Curras + Baladi: Towards a Levantine Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
13. Mustafa Jarrar: The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content. Applied Ontology Journal, 16:1, 1-26. IOS Press. 2021
14. Moustafa Al-Hajj, Mustafa Jarrar: ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021
15. Moustafa Al-Hajj, Mustafa Jarrar: LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation. In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). PP 748--755, Association for Computational Linguistics. 2021
16. Eman Naser-Karajah, Nabil Arman, Mustafa Jarrar: Current Trends and Approaches in Synonyms Extraction: Potential Adaptation to Arabic. In Proceedings of the 2021 International Conference on Information Technology (ICIT). PP 748--755, Association for Computational Linguistics. pp. 428-434, IEEE. 2021
17. Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, Khaled Shaalan: Extracting Synonyms from Bilingual Dictionaries. The 11th International Global Wordnet Conference (GWC2021), Global Wordnet Association. (pp. 215-222). Pretoria, South Africa, 2021
18. Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, Hamdy Mubarak: A Panoramic Survey of Natural Language Processing in the Arab World. Communications of the ACM, April 2021, Vol. 64 No. 4, Pages 72-81
19. Mustafa Jarrar: Digitization of Arabic Lexicons. Arabic Language Status Report. UAE Ministry of Culture and Youth. Pages 214-2017. Dec 2020
20. Mustafa Jarrar, Hamzeh Amayreh: An Arabic-Multilingual Database with a Lexicographic Search Engine. The 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Pages(234-246). LNCS 11608, Springer. 2019
21. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: Representing Arabic Lexicons in Lemon - a Preliminary Study. The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
22. Diana Alhafi, Anton Deik, Mustafa Jarrar: Usability Evaluation of Lexicographic e-Services. The 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Abu Dhabi, UAE. 2019
23. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: Diacritic-Based Matching of Arabic Words. ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
24. Mustafa Jarrar: Search Engine for Arabic Lexicons. The 5th Conference on Translation and the Problematics of Cross-cultural Understanding. The Forum for Arab and International Relations. Doha, Qatar. December, 2018
25. Diab Abuaiadah, Dileep Rajendran, Mustafa Jarrar: Clustering Arabic Tweets for Sentiment Analysis. The 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications. Pages(499-506). IEEE Computer Society. ISBN:9781538635810. (doi.10.1109/AICCSA.2017.162). Hammamet, Tunisia. 2017
26. Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, Nasser Zalmout: Curras: An Annotated Corpus for the Palestinian Arabic Dialect. Journal Language Resources and Evaluation. Pages(745-775). Volume(51).