مدونة وجود لاستخراج الكينونات

https://sina.birzeit.edu/wojood

بُنِيَت مدونة ضخمة تُسمَّى "وجود" وجرى تدريب نماذج لغوية لاستخراج الكينونات المسماة (Named-Entity Recognition NER) كأسماء الأشخاص، والمؤسسات، والأماكن، والأحداث، والمنتجات، وغيرها (انظر الأمثلة المرفقة). تُعَدُّ مهمة استخراج الكينونات من أهم مهمات حوسبة اللغة والذكاء الاصطناعي، ولا يوجد مدونات عربية مشابهة سوى بعض مدونات صغيرة محدودة التوسيمات. المدونة التي طُوِّرَت تحتوى على 550 ألف كلمة تشمل الفصحى والعامية وتغطي حقول متنوعة مثل السياسة والاقتصاد والقانون والتاريخ والصحة والهندسة وغيرها. والأهم أن المدونة تغطي 21 نوعاً من الكينونات وتدعم توسيم الكينونات الضمنية (Nested Named Entities). يقارن الجدول التالي بين مدونة "وجود" والمدونات الأخرى، وتجدر الإشارة إلى أن معدل دقة النموذج اللغوي الذي تم تدريبه على مدونة وجود 88.4%.

يمكن الوصول إلى كل من المدونة والنموذج اللغوي وتحميلهما من الموقع، ويجري حاليا تنظيم مسابقة (SharedTask) اعتمادا على مدونة "وجود".

Corpus	Nested	Tokens	Entities	Classes	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**

أوراق علمية

للمزيد يمكن الضغط على الرابط بلون (تركواز)

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

ورقة تصف مدونة وجود ومنهجية بناؤها والنماذج اللغوية التي تم تطويرها

*

فيديو

فيديو عرض مدونة وجود في مؤتمر (LERC)

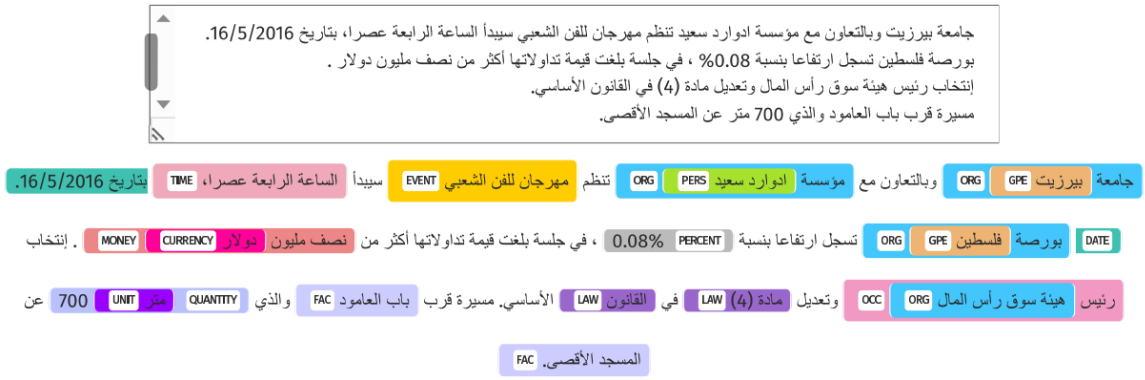
مثال (screenshot) لاستخراج الكينونات من النص

https://sina.birzeit.edu/wojood/

Wojood

A corpus and model for nested Arabic Named Entity Recognition

Wojood consists of about 550K tokens (MSA and dialect) that are manually annotated with 21 entity types (e.g., person, organization, location, event, date, etc). It covers multiple domains and was annotated with nested entities. The corpus contains about 75K entities and 22.5% of which are nested. A nested named entity recognition (NER) model based on BERT was trained (F1-score 88.4%). Try the service:



Description

Corpus size: 550K tokens (MSA and dialects)
Richness: 21 entity classes, contains ~75K entities and 22.5% of them are nested entities
Domains: Media, History, Culture, Health, Finance, ICT, Law, Elections, Politics, Migration, Terrorism, social media
IAA: 97.9% (Cohen's Kappa)
NER Model: AraBERTV2 (88.4% F1-score)
Entity Classes (21):

PERS (person)	EVENT	CARDINAL
NORP (group of people)	DATE	ORDINAL
OCC (occupation)	TIME	PERCENT
ORG (organization)	LANGUAGE	QUANTITY
GPE (geopolitical entity)	WEBSITE	UNIT
LOC (geographical location)	LAW	MONEY
FAC (facility: landmarks places)	PRODUCT	CURR (currency)

Downloads

Wojood is available to download upon request for academic and commercial use.
Request to download Wojood (Nested NER corpus, 550K tokens)
GitHub (download BERT training source code + sample data (~35K tokens))
Hugging Face (download fine-tuned BERT model, ready to use)