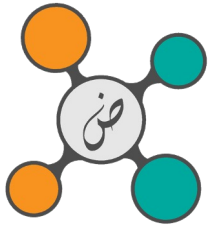
The 12th International Global Wordnet Conference (GWC2023)
Donostia / San Sebastian (Spain)
23-27/01/2023



Context-Gloss Augmentation for Improving Arabic Target Sense Verification

Sanad Malaysha **Mustafa Jarrar** **Mohammed Khalilia**

Birzeit University
Palestine

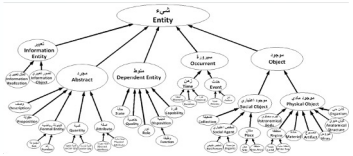

جامعة بيرزيت
BIRZEIT UNIVERSITY
Building linguistic resources for NLP

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, …

WSD 84%

Synonyms

NER 88.4%

Big Linguistic Data Graph

https://ontology.birzeit.edu

Semantic/meaning Understanding Tasks

1. Word Sense-Disambiguation (WSD)

2. Target Sense Verification (TSV)

3. Word-in-Context (WiC)

Word-Sense Disambiguation (WSD)

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. .....

Target Sense Verification (TSV)

- Given a context, target word and gloss, TSV aims to decide whether it is true that this gloss is the intended meaning of the target in this context.
- Whether a (Context-Gloss pair) is true or false

Example:

Context	Gloss	Label
تمشي بين الجداول والازهار Walking among streams and flowers	مجرى صغير متفرع من نهر A small stream branching from a river	True
تمشي بين الجداول والازهار Walking among streams and flowers	تنظيم للبيانات والمعلومات في صورة صفوف وأعمدة Organization of data in the form of rows and columns	False

Word-in-Context (WiC)

Determines whether a target word in two contexts (sentences) is used in the same sense or not

Example:

Context 1	Context 2	Label
تمشي بين الجداول والازهار Walking among streams and flowers	كنا نمرح ونستمتع بجداول الربيع We were playing and enjoying the spring streams	True
تمشي بين الجداول والازهار Walking among streams and flowers	انظر الجداول في الصفحة الثالثة See the table in third page	False

ArabGlossBERT (TSV Dataset)

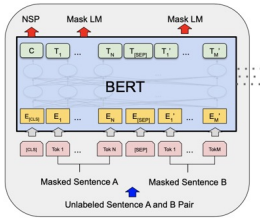
❖ Arabic context-gloss pairs dataset (167k) (Al-hajj & Jarrar, 2021)

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

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

❖ Fine-tuned BERT Model

- Trained as **binary sequence-pair classification task**
- Accuracy 84%



Pre-training

ArabGlossBERT TSV Dataset

❖ **Arabic context-gloss pairs Dataset (167k)** (Al-hajj & Jarrar, 2021)

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

	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

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

Goal: Context-Gloss Augmentation

Augment the ArabGlossBERT dataset using back-translated (Arabic→English→Arabic)

	Context	Gloss
(Arabic)	فكرة أو مسألة تقدم للبحث	جلس المسؤولون يناقشون أطروحات المشروع
(Arabic->English)	An idea or question is progressing to research	Officials sat discussing project proposals
Back-Translated (English->Arabic)	فكرة أو سؤال مقدم للبحث	جلس المسؤولون لمناقشة مقترحات المشاريع

Contributions

- Augmented ArabGlossBERT using back- translation (352K pairs).
- 13 experiments with different dataset configurations - to measure whether the back- translation enrichment can improve TSV performance.
- In-depth analysis of the TSV accuracy for each part-of-speech.

Dataset Augmentation

Term	Original ArabGlossBERT	Back-Translations Pairs	Augmented ArabGlossBERT
Unique un-diacritized lemmas	26,169	--	26,169
Unique Glosses	32,839	32,839	65,678
Unique Contexts	60,272	60,272	120,544
Training pairs	152,035	32,839 + 152,035	**336,909**
Positive pairs	55,585	32,839 + 55,585	144,009
Negative pairs	96,450	96,450	192,900
Testing pairs	15,172	--	**15,172**
Positive pairs	4,738	--	4,738
Negative pairs	10,434	--	10,434
Total: Train + Test	167,207	32,839 + 152,035	**352,081**

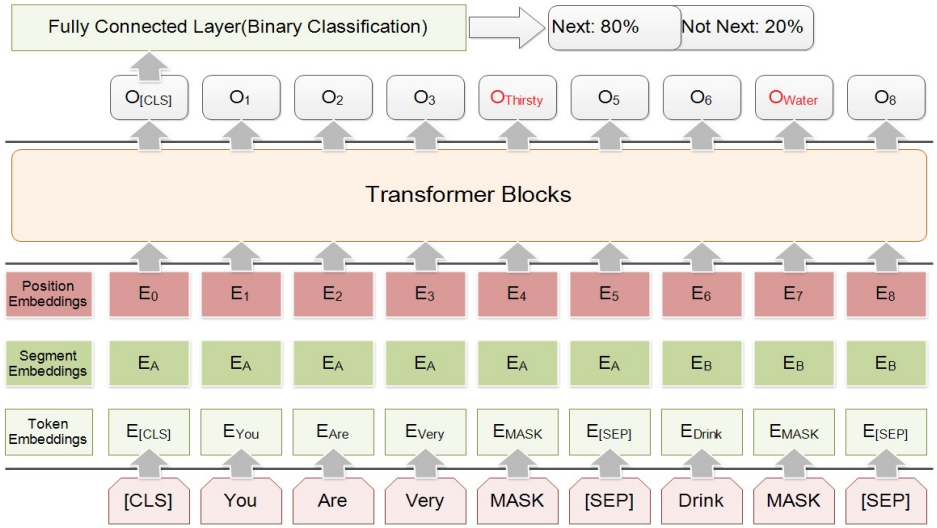
Google's Back-Translations

- Generally acceptable, but there are wrong translations
  - No manual improvement to these translations - as we to measure whether automated back-translations can improve the task accuracy

- Sometimes translates half of the sentence:
  - we added a dot (.) at the end of each sentence, and check whether it was translated back
  - We compared the length of the original and back-translated sentences. If the difference is significant, then we reviewed and fixed them manually

Experiments

- We fine-tuned AraBERTv2 with these hyperparameters:
  - η = 2e−5
  - batch size B = 16
  - max sequence length of 512
  - warm-up steps 1,412
  - number of epochs 4



Fine-Tuning Datasets used for AraBERT

Dataset	Description	Positive Pairs	Negative Pairs	Total
D_1	The original ArabGlossBERT dataset	55,585	96,450	152,035
D_2	D1 with target signal	55,585	96,450	152,035
D_3	D1 with context replaced by back-translated context	55,585	96,450	152,035
D_4	D_1 + Positive pairs of D_3	111,170	96,450	207,620
D_5	D_1 + D_3	111,170	192,900	304,070
D_6	D_1 + Positive pairs (original gloss - back-translated gloss)	88,424	96,450	184,874
D_7	D_4 + Positive pairs (original gloss - back-translated gloss)	144,009	96,450	240,459
D_8	D_5 + Positive pairs (original gloss - back-translated gloss)	144,009	192,900	336,909
D_9	D_1 + Positive pairs (original context-back-translated gloss)	111,170	96,450	207,620
D_{10}	D_1 + Pairs of cross relating the glosses against each other	88,424	373,955	462,379
D_{11}	D_1 (exclude pairs of functional words)	54,878	92,730	147,608
D_{12}	D_1 (only the pairs of the noun POS)	36,487	37,998	74,485
D_{13}	D_1 (only the pairs of the verb POS)	18,178	54,945	73,123

Augmented ArabGlossBERT

Results

D_1: original dataset
152,035 (55,585 + 96,450)

D_3 = D_1 but bk-trans. contexts
152,035 (55,585 + 96,450)

D_1 but bk-trans. glosses
184,874 (88,424 + 96,450)

signal

Dataset	Metric	All POS		Accuracy	Noun		Accuracy	Verb		Accuracy	Functional Words		Accuracy
		Positive	Negative		Positive	Negative		Positive	Negative		Positive	Negative	
D1 Baseline 152,035 pairs	Precision	76	85		75	85		78	85		63	84	
	Recall	66	90	83	70	88	82	65	91	83	46	92	81
	F1-Score	71	88		72	82		71	88		53	88	
D2 152,035 pairs	Precision	81	85		79	85		82	85		71	82	
	Recall	65	93	84	68	91	83	64	94	84	36	95	81
	F1-Score	72	89		73	88		72	89		48	88	
D3 152,035 pairs	Precision	68	80		65	79		70	80		55	79	
	Recall	52	88	77	54	85	75	52	90	78	19	95	77
	F1-Score	59	84		59	82		60	85		29	86	
D4 207,620 pairs	Precision	80	81		79	80		81	81		69	80	
	Recall	53	94	81	55	92	80	53	94	81	23	97	79
	F1-Score	64	87		65	86		64	87		34	88	
D5 304,070 pairs	Precision	76	82		77	79		76	84		70	80	
	Recall	57	92	81	53	92	80	62	91	82	24	97	79
	F1-Score	65	87		63	85		68	87		36	88	
D6 184,874 pairs	Precision	76	85		76	84		76	87		71	82	
	Recall	67	90	**83**	66	89	81	70	90	84	32	96	81
	F1-Score	71	88		71	**86**		**73**	88		44	88	
D7 240,459 pairs	Precision	79	82		77	81		80	83		71	79	
	Recall	56	93	81	57	91	80	58	93	82	17	98	79
	F1-Score	66	87		66	86		67	88		17	98	
D8 336,909 pairs	Precision	80	81		79	80		81	81		69	80	
	Recall	54	94	81	55	92	80	53	94	81	23	97	79
	F1-Score	65	87		65	86		64	87		34	88	
D9 207,620 pairs	Precision	78	84		77	83		78	86		73	81	
	Recall	63	92	**83**	62	91	81	66	92	84	31	96	81
	F1-Score	70	88		69	**86**		**72**	88		43	88	
D10 462,379 pairs	Precision	71	80		70	78		71	81		66	79	
	Recall	51	90	78	50	89	76	54	90	79	19	97	78
	F1-Score	59	85		58	83		61	85		30	87	
D11 147,750 pairs	Precision	80	81		79	80		81	81				
	Recall	54	94	81	55	92	80	53	94	81			
	F1-Score	65	87		65	86		64	87				
D12 74,485 pairs	Precision				80	82							
	Recall				60	92	81						
	F1-Score				69	87							
D13 73,123 pairs	Precision							74	84				
	Recall							62	90	81			
	F1-Score							68	87				

Back-translation pairs achieved 77%.

Automatic translations of glosses and contexts is not high but is generally acceptable.

Augmented ArabGlossBERT

Results

D_1 + Positive pairs of D3
207,620(111,170 + 96,450)

D1 + D3
304,070 (111,170 + 192,900)

D_4+ Positive Gloss-GlossBT
184,874 (144,009 + 240,459)

D_5+Positive Gloss-GlossBT
336,909 (144,009 + 192,900)

D_4 + Positive Context-glossBT
207,620 (111,170 + 96,450)

D_1 +GlossBT+many False pairs
462,379 (88,424 + 373,955)

Dataset	Metric	All POS		Accuracy	Noun		Accuracy	Verb		Accuracy	Functional Words		Accuracy
		Positive	Negative		Positive	Negative		Positive	Negative		Positive	Negative	
D1 Baseline 152,035 pairs	Precision	76	85		75	85		78	85		63	84	
	Recall	66	90	83	70	88	82	65	91	83	46	92	81
	F1-Score	71	88		72	82		71	88		53	88	
D2 152,035 pairs	Precision	81	85		79	85		82	85		71	82	
	Recall	65	93	84	68	91	83	64	94	84	36	95	81
	F1-Score	72	89		73	88		72	89		48	88	
D3 152,035 pairs	Precision	68	80		65	79		70	80		55	79	
	Recall	52	88	77	54	85	75	52	90	78	19	95	77
	F1-Score	59	84		59	82		60	85		29	86	
D4 207,620 pairs	Precision	80	81		79	80		81	81		69	80	
	Recall	53	94	81	55	92	80	53	94	81	23	97	79
	F1-Score	64	87		65	86		64	87		34	88	
D5 304,070 pairs	Precision	76	82		77	79		76	84		70	80	
	Recall	57	92	81	53	92	80	62	91	82	24	97	79
	F1-Score	65	87		63	85		68	87		36	88	
D6 184,874 pairs	Precision	76	85		76	84		76	87		71	82	
	Recall	67	90	**83**	66	89	81	70	90	84	32	96	81
	F1-Score	71	88		71	**86**		**73**	88		44	88	
D7 240,459 pairs	Precision	79	82		77	81		80	83		71	79	
	Recall	56	93	81	57	91	80	58	93	82	17	98	79
	F1-Score	66	87		66	86		67	88		17	98	
D8 336,909 pairs	Precision	80	81		79	80		81	81		69	80	
	Recall	54	94	81	55	92	80	53	94	81	23	97	79
	F1-Score	65	87		65	86		64	87		34	88	
D9 207,620 pairs	Precision	78	84		77	83		78	86		73	81	
	Recall	63	92	**83**	62	91	81	66	92	84	31	96	81
	F1-Score	70	88		69	**86**		**72**	88		43	88	
D10 462,379 pairs	Precision	71	80		70	78		71	81		66	79	
	Recall	51	90	78	50	89	76	54	90	79	19	97	78
	F1-Score	59	85		58	83		61	85		30	87	
D11 147,750 pairs	Precision	80	81		79	80		81	81				
	Recall	54	94	81	55	92	80	53	94	81			
	F1-Score	65	87		65	86		64	87				
D12 74,485 pairs	Precision				80	82							
	Recall				60	92	81						
	F1-Score				69	87							
D13 73,123 pairs	Precision							74	84				
	Recall							62	90	81			
	F1-Score							68	87				

Augmented ArabGlossBERT

Results

Excluding the pairs of functional words from the dataset (experiment 11) did not improve the performance

all POS categories yields better performance than fine-tuning separate models for nouns and verbs (experiments 12- 13)

D_1 (exclude pairs func lemmas)
147,608 (54,878 + 92,730)

D_1 (only the pairs of nouns)
74,485 (36,487 + 37,998)

D_1 (only the pairs of verbs)
73,123 (18,178 + 54,945)

Dataset	Metric	All POS		Accuracy	Noun		Accuracy	Verb		Accuracy	Functional Words		Accuracy
		Positive	Negative		Positive	Negative		Positive	Negative		Positive	Negative	
D1 Baseline 152,035 pairs	Precision	76	85		75	85		78	85		63	84	
	Recall	66	90	83	70	88	82	65	91	83	46	92	81
	F1-Score	71	88		72	82		71	88		53	88	
D2 152,035 pairs	Precision	81	85		79	85		82	85		71	82	
	Recall	65	93	84	68	91	83	64	94	84	36	95	81
	F1-Score	72	89		73	88		72	89		48	88	
D3 152,035 pairs	Precision	68	80		65	79		70	80		55	79	
	Recall	52	88	77	54	85	75	52	90	78	19	95	77
	F1-Score	59	84		59	82		60	85		29	86	
D4 207,620 pairs	Precision	80	81		79	80		81	81		69	80	
	Recall	53	94	81	55	92	80	53	94	81	23	97	79
	F1-Score	64	87		65	86		64	87		34	88	
D5 304,070 pairs	Precision	76	82		77	79		76	84		70	80	
	Recall	57	92	81	53	92	80	62	91	82	24	97	79
	F1-Score	65	87		63	85		68	87		36	88	
D6 184,874 pairs	Precision	76	85		76	84		76	87		71	82	
	Recall	67	90	**83**	66	89	81	70	90	84	32	96	81
	F1-Score	71	88		71	**86**		**73**	88		44	88	
D7 240,459 pairs	Precision	79	82		77	81		80	83		71	79	
	Recall	56	93	81	57	91	80	58	93	82	17	98	79
	F1-Score	66	87		66	86		67	88		17	98	
D8 336,909 pairs	Precision	80	81		79	80		81	81		69	80	
	Recall	54	94	81	55	92	80	53	94	81	23	97	79
	F1-Score	65	87		65	86		64	87		34	88	
D9 207,620 pairs	Precision	78	84		77	83		78	86		73	81	
	Recall	63	92	**83**	62	91	81	66	92	84	31	96	81
	F1-Score	70	88		69	**86**		**72**	88		43	88	
D10 462,379 pairs	Precision	71	80		70	78		71	81		66	79	
	Recall	51	90	78	50	89	76	54	90	79	19	97	78
	F1-Score	59	85		58	83		61	85		30	87	
D11 147,750 pairs	Precision	80	81		79	80		81	81				
	Recall	54	94	81	55	92	80	53	94	81			
	F1-Score	65	87		65	86		64	87				
D12 74,485 pairs	Precision				80	82							
	Recall				60	92	81						
	F1-Score				69	87							
D13 73,123 pairs	Precision							74	84				
	Recall							62	90	81			
	F1-Score							68	87				

Conclusion

- Augmented ArabGlossBERT dataset to 352K context-gloss pairs using Back-Translation
- Accuracy: 81% - 84% accuracy on all POS
- Functional words have the lowest performance
- Positive pairs have lower accuracy than Negative pairs
- Back-Translation might generalize the model - future work to test

References

- Moustafa Al-Hajj and Mustafa Jarrar. 2021. Lu-bzu at semeval-2021 task 2: Word2vec and lemma2vec per- formance in arabic word-in-context disambiguation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 748– 755, Online. Association for Computational Linguis- tics.
- Moustafa Al-Hajj and Mustafa Jarrar. 2022. Arabgloss- bert: Fine-tuning bert on context-gloss pairs for wsd. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3September, 2021*, pages 35–43.
- 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 Jar- rar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.
- Mustafa Jarrar. 2006. Towards the notion of gloss, and the adoption of linguistic resources in formal ontol- ogy engineering. In *Proceedings of the 15th interna- tional conference on World Wide Web (WWW2006)*, pages 497–503. ACM Press, New York, NY.
- Mustafa Jarrar. 2011. Building a formal arabic ontology (invited paper). In *Proceedings of the Experts Meet- ing on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.
- Mustafa Jarrar. 2018. Search engine for arabic lexicons. Mustafa Jarrar. 2020. *Digitization of Arabic Lexicons*,
- pages 214–217. UAE Ministry of Culture and Youth.
- Mustafa Jarrar. 2021. The arabic ontology - an arabic wordnet with ontologically clean content. *Applied Ontology Journal*, 16(1):1–26.
- Mustafa Jarrar and Hamzeh Amayreh. 2019. An arabic- multilingual database with a lexicographic search engine. In *The 24th International Conference on Ap- plications of Natural Language to Information Sys- tems (NLDB 2019)*, volume 11608 of *LNCS*, pages 234–246. Springer.
- Mustafa Jarrar, Hamzeh Amayreh, and John P. McCrae. 2019. Representing arabic lexicons in lemon - a preliminary study. In *The 2nd Conference on Lan- guage, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings.
- Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 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.
- Mustafa Jarrar, Fadi Zaraket, Rami Asia, and Hamzeh Amayreh. 2018. Diacritic-based matching of arabic words. *ACM Asian and Low-Resource Language Information Processing*, 18(2):10:1–10:21.