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

Abstract

SALMA, the first Arabic sense-annotated corpus, consists of ~34K tokens, which are all sense-annotated. The corpus is annotated using two different sense inventories simultaneously (Modern and Ghani). SALMA novelty lies in how tokens and senses are associated. Instead of linking a token to only one intended sense, SALMA links a token to multiple senses and provides a score to each sense. A smart web-based annotation tool was developed to support scoring multiple senses against a given word. In addition to sense annotations, we also annotated the corpus using six types of named entities. The quality of our annotations was assessed using various metrics (Kappa, Linear Weighted Kappa, Quadratic Weighted Kappa, Mean Average Error, and Root Mean Square Error), which show very high inter-annotator agreement. To establish a Word Sense Disambiguation baseline using our SALMA corpus, we developed an end-to-end Word Sense Disambiguation system using Target Sense Verification. We used this system to evaluate three Target Sense Verification models available in the literature. Our best model achieved an accuracy with 84.2% using Modern and 78.7% using Ghani. The full corpus and the annotation tool are open-source and publicly available at https://sina.birzeit.edu/salma/.

1 Introduction

WSD aims to determine a word's intended meaning (sense) in a given context. WSD is underdeveloped in Arabic due to the lack of sense-annotated datasets. This is in addition to the challenging nature of the WSD task due to the semantic polysemy of the words (Al-Hajj and Jarrar, 2021). For instance, the Arabic word (عَين *ʿayn*) has sixteen meanings in the Contemporary Arabic Dictionary (Omar, 2008). In the context (رأيتُه رأي العَين *rʾaytuh rʾay ālʿayn*), word (عَين *ʿayn*) refers to *eye*, while in (شربت مِن عَين الماء *šribt min ʿayn ālmāʾ*), it refers to *water spring*. Similarly, the English word *book* as a noun has ten different senses in Princeton WordNet (Miller et al., 1990), such as (a written work or composition that has been published), or (number of pages bound together). WSD has been considered a challenging task for many years (Weaver, 1949/1955), but it has recently gained more attention due to the advances in learning contextualized word representations from language models, such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018).

As glosses are short descriptions of senses (Jarrar, 2006, 2005), recent research has demonstrated promising results in WSD task by framing the problem as a sentence-pair (context-gloss) binary classification task, referred to as Target Sense Verification (TSV), where the context is a sentence containing the ambiguous word (Huang et al., 2019; Yap et al., 2020; Blevins and Zettlemoyer, 2020). Al-Hajj and Jarrar (2021) proposed an approach for Arabic WSD (using TSV) based on context-gloss pairs extracted from the Arabic Ontology and lexicons and they achieved 84% accuracy, but this evaluation was done on a TSV dataset rather than a WSD evaluation using a sense-annotated corpus. Additionally, Al-Hajj and Jarrar (2021) presented an attempt for Arabic Word-in-Context (WiC) disambiguation using the dataset provided by the SemEval shared task (Martelli et al., 2021).

This article presents SALMA, the first sense-annotated Arabic corpus consisting of about 34K tokens, which are manually annotated with senses. Since there are no available sense inventories for Arabic, We used two Arabic lexicons as sense inventories: Contemporary Arabic Dictionary (اللغة العربية المعاصرة *āllġh ālʿrbyh ālmʿāṣrh*), hereafter we refer to as **Modern** (Omar, 2008), and Al-Ghani Al-Zaher (الغني الزاهر *ālġny ālzāhr*), hereafter we refer to as **Ghani** (Abul-Azm, 2014). These two lexicons are part of the lexicon digitization project and lexicographic database at SinaLab[1] (Jarrar and Amayreh, 2019; Alhafi et al., 2019; Amayreh et al., 2019; Ghanem et al., 2023; Jarrar et al., 2021). We introduce a novel sense-annotation framework (Section 3), in which all candidate senses, from both lexicons, are scored to indicate their semantic

[1] https://sina.birzeit.edu/

relatedness to a token appearing within a context. The higher the score, the more semantically related the sense is. For better coverage, we annotated each token in our corpus using both lexicons independently and in parallel. The scores assigned to senses of the Modern do not influence the scoring of the Ghani senses. In addition, we also annotated our corpus using six types of named entities: person (PERS), organization (ORG), geopolitical entity (GPE), location (LOC), facility (FAC), and currency (CURR). The corpus was annotated by three linguists and we assessed the inter-annotator agreement (IAA) using 2.6% of the annotated words in the corpus. To establish a baseline for WSD in Arabic, we developed an end-to-end WSD system, in which we benchmarked three available TSV models, with different settings. The best model resulted in 84.2% accuracy using Modern and 78.7% using Ghani. The main contributions of this paper are:

- *Sense-annotated corpus*, annotated with two sense inventories independently, and six named entities; and most importantly, each word is linked with all of its senses, and each sense is given a score.
- *Web-based sense-annotation framework* to score all senses of a given word.
- *End-to-end WSD system*, implemented and evaluated using three different TSV models.
- *WSD baseline for Arabic*, with different settings.

The remainder of the article is organized as follows: Section 2 highlights the related work, Section 3 presents the corpus, Section 4 describes the inter-annotator agreement, Sections 5 and 6 present how the baselines are produced, we conclude in Section 7 and outline the limitations and future work in Section 8.

2 Related Work

We will first review related sense-annotated corpora, then we will review related sense inventories.

One of the known English sense-annotated corpora is SemCor (Miller et al., 1993), which is annotated using the Princeton WordNet (Miller et al., 1990). It contains about 200K sense annotations for around 700K words, but not all words are sense-annotated in the SemCor corpus, especially multi-word expressions, articles, and prepositions. The AnCora corpus for Spanish and Catalan languages (Taulé et al., 2008) was collected from newspapers and consists of 500K words, but only 200K noun words are semantically annotated using the Spanish WordNet. AnCora also includes morphological, semantic, and syntactic annotations. TuBa-D/Z is a German annotated corpus, manually collected from newspapers and annotated using the GermanNet senses (Telljohann et al., 2004). TuBa-D/Z was later used as a gold standard for the WSD task by (Petrolito and Bond, 2014). The Italian Syntactic-Semantic Treebank (ISST) is a corpus built for the Italian language with 89,941 sense-annotated words (Montemagni and Venturi, 2003). The ISST annotations cover five levels that are related to lexico-semantics such as orthographic, morpho-syntactic, semantic, and syntactic aspects.

The NTU-MC corpus (Tan and Bond, 2012) covers eight languages including Thai, Vietnamese, Arabic, Korean, Indonesian, Japanese, Mandarin Chinese, and English. However, the Arabic version is not publicly available. This corpus was collected from short stories, essays, and tourism articles resulting in a total of 116K words, but only 63K words are annotated. KPWr, a Polish corpus, contains text from multiple domains including science, law, religion, and press (Broda et al., 2012) with a total of 438,327 words, but only 9,157 words are annotated using the Polish WordNet (Maziarz et al., 2012).

For Arabic, the focus of research has been primarily on developing corpora for morphological and syntactic tagging (Darwish et al., 2021) rather than semantic and sense annotation, as noted by Elayeb (2019) and Naser-Karajah et al. (2021). For instance, part of the OntoNotes corpus (Weischedel et al., 2013) covers limited semantic annotations for Arabic using a small sense inventory of size 261 senses (150 verbs and 111 nouns). Additionally, AQMAR corpus (Schneider et al., 2012) is annotated with 25 super-sense labels representing broad semantic fields such as ARTIFACT and PERSON, which can be considered as general types of named entities, rather than word-sense annotations. They annotated ~22K nouns out of 65K tokens corpus. Table 1 compares our proposed corpus and related Arabic resources.

In addition to the lack of sense-annotated corpora, Arabic lacks reliable sense inventories. Although there are some available semantic resources, they are not mature enough to be used as sense

Corpus	Unique Senses	Annotation Type	Corpus Size (tokens)	Annotations				
				Nouns	Verbs	Func. Words	Punc.+ Digits	Total
AQMAR	25 semantic fields (closer to named entities)	selected words each one sense	65K	~22K	–	–	–	~22K
OntoNotes5	261 semantic fields (high-level grouped senses)	selected words each one sense	300K	8,700	4,300	–	–	13K
SALMA (ours)	4,151 word senses (from each sense inventory) 6 types of named entities	all senses of all words	34K	**19,030**	**2,763**	**7,116**	**5,344**	**34,253**

Table 1: Overview of related Arabic sense-annotated corpora.

inventories. For example, the Arabic WordNet (Black et al., 2006) contains about 10K senses, and the Arabic Ontology (Jarrar, 2021, 2011) contains about 18K synsets. However, both resources cannot be used as sense inventories as they do not provide a complete set of senses for a given lemma (i.e., lexicon entry). The lexicographic database developed at Birzeit University contains about 150 Arabic lexicons (Jarrar and Amayreh, 2019; Jarrar et al., 2019), but these lexicons are not well-structured or suitable to be used as sense inventories (Jarrar and Amayreh, 2019). Due to the lack of dependable Arabic sense inventory, we decided to obtain a license to digitize and use two Arabic lexicons as sense inventories, namely, **Modern** (Omar, 2008) and **Ghani** (Abul-Azm, 2014).

3 Corpus Construction and Annotation

3.1 Corpus Collection

Our SALMA corpus is part of the Wojood corpus (Jarrar et al., 2022), and was collected from 33 online media sources written in Modern Standard Arabic (MSA) and covering general topics. Some of those sources include mipa.institute, sanaacenter.org, hrw.org, diplomatie.ma, sa.usembassy.gov, eeas.europa.eu, crisisgroup.org, and mofaic.gov.ae. The corpus was then segmented into sentences and tokenized, resulting in 1439 sentences and ~34K tokens, with an average of 23.8 tokens per sentence.

3.2 Annotation Framework

This section presents a novel sense annotation framework, where instead of linking a word to one sense, we propose to score all semantically related senses to the word. The score ranges between 1-100% and a sense with a score $\geq 60\%$ is considered a correct sense of the word. The ranking scale is divided into six categories:

- *Explicate* /مباشرة (100%): direct and explicate semantics (دلالة صحيحة وصريحة).
- *General* /معنى عام (80%): correct but implicate semantics (دلالة صحيحة غير مباشرة).
- *Referral* /دلالة لغوية (60%): generally correct semantics, but is referred to another lemma (صحيحة ولكن عامة جداً مثل مصدر، اسم فاعل). For example, the word *drinker* and its gloss (*active participle of drink*).
- *Related* /ذات علاقة (40%): weak semantics (مشتركة في الدلالة العامة فقط، أختها دلالياً). For example, the term (سياسة الشركة *syāsh ālšrkh*) / *company's policy*, is related to the sense (*the policy used to collect taxes*) which is not a sense of the lemma (سياسة *syāsh*), but semantically related.
- *Root semantics* /دلالة جذر (20%): share root semantics (دلالة مختلفة ولكن تشترك في الدلالة المجردة التي يحملها الجذر، مثل الدلالة المجازية). In Arabic lexical semantics, all words with the same root share part of the semantics of this root (Ryding, 2014; Boudelaa and Marslen-Wilson, 2004; Boudelaa et al., 2010). For example, all senses of the lemma (سياسة *syāsh*), such as *politics* and *policies* share an abstract meaning (e.g., issues related to governing and acting).
- *Different* /مختلفة (1%): unrelated semantics (دلالة مختلفة تماماً).

This framework serves several purposes. First, in case of underdeveloped sense inventories (such as the Modern and Ghani lexicons), in which glosses might be vague, redundant, or overlapping, our framework allows the annotators to score each sense. In this paper, we linked every word in the corpus with all semantically related senses in Modern and Ghani, thus we were able to compare and evaluate the lexical coverage in both lexicons (see Section 3.5). Another advantage of using this framework (i.e., scoring all senses) is that our corpus can be used to benchmark ranking-based WSD methods (Conia and Navigli, 2021; Yap et al.,

Figure 1: Screenshot of our web-based annotation tool.

2020), which is not possible in the case of one-sense annotated corpora.

3.3 Annotation Tool

We developed a web-based tool optimized for our sense annotation framework and methodology. On the right side of Figure 1, the linguist selects a word to be annotated (such as "السياسة *ālsyāsh*"). The tool will then retrieve all sentences (i.e. contexts) in the corpus containing the selected word. The tool will also automatically fetch the lemma of the selected word, and the linguist has the ability to search for the lemma manually. After selecting a lemma, the tool retrieves senses associated with the lemma from both lexicons, Modern and Ghani. The linguist can then select the score category for each sense according to our guideline and apply these scores to all selected words (in contexts) as shown in Figure 1. The scores are selected from a `ComboBox` of the six categories (See Section 3.2), however, the tool internally stores their corresponding numeric values.

3.4 Annotation Process

The annotation was carried out in three phases:

Phase 1 (training): we recruited three undergraduate students majoring in linguistics. The students were trained in three steps in order to produce consistent annotations. We first assigned 50 words to each linguist and trained them to conduct the annotation jointly. Second, we assigned the same 150 words to each student separately, then asked them to compare and consolidate their annotations, which helps in calibrating their scoring. Third, we repeated the second phase, but using 300 words and again we asked them to compare their annotations.

Phase 2 (annotation): out of ~34K tokens, excluding digits and punctuations, we assigned about 9.6K words to each of the three linguists. Each linguist was asked to annotate all occurrences of each word in the corpus - resulting in about ~29K annotations for the whole words.

Phase 3 (validation): after finishing the annotations, we used the tool to automatically validate the annotations and flag those that violated the following cases: (i) a word is annotated with more than one *Explicit* or *General* sense in the same lexicon, which is an indication of either a mistake or redundant or overlapping senses in the lexicon. (ii) a word is missing either an *Explicit* or a *General*

sense; this is an indication of a mistake or the lexicon is missing this sense. (iii) if the selected sense is a proper noun, then all other senses should be ranked as *Different*. The linguists were asked to review these flagged annotations and revise them if necessary.

The linguists were encouraged to discuss among themselves and take joint decisions when facing difficulties, especially in the case of vague glosses or contexts. In addition, as will be discussed in Section 3.5, missing lemmas and senses are manually added to the lexicons. Table 2 provides general statistics about the annotations. It is worth noting that sense annotations are typically costly and time-consuming. The linguists spent about 600 working days (i.e., 4800 working hours) to carry out the three phases described above.

Term	Noun	Verb	Func. Words	Punc+ Digits	Total
Tokens	19,030	2,763	7,116	5,344	**34,253**
Unique Tokens	6,670	1,593	322	175	**8,760**
Unique Lemmas	2,904	677	119	175	**3,875**
Unique Senses	3,151	792	206	2	**4,151**

Table 2: Statistics of the SALMA corpus.

Term	Modern	Ghani
Lemmas	80% (2,788/3,522)	78% (2,724/3,522)
Senses (Without Proper nouns)	83% (3,430/4,151)	78% (3,226/4,151)
Proper Nouns Senses	4% (9/213)	14% (30/213)

Table 3: Coverage of Modern and Ghani lexicons.

3.5 Discussion and Lexical Coverage

We evaluated the coverage of both lexicons based on the sense-annotated tokens. As Table 3 shows, Modern has higher coverage of lemmas (80%) compared to Ghani's coverage (78%), and has higher sense coverage (83%) compared to Ghani (78%). Moreover, glosses in Modern are more precise, less ambiguous and well-formulated as discussed in Section 4.1. The proper nouns are the main reason for the missing lemmas and senses, as the Modern and Ghani cover 4% and 14% of proper nouns in SALMA corpus, respectively. Lemmas and senses that are not covered by any of the two lexicons were added manually by the linguists. All numerical values are annotated with the same "digit" sense that covers ordinal and nominal numbers, and similarly, punctuation marks are all annotated with "Punc".

3.6 Named Entity Annotations

Named-entity annotations are important in sense-annotated corpora because sense inventories do not typically cover names of organizations, towns, people, landmarks, and others.

Tag	Description
PERS	Person names: first, middle, last, nickname ...
ORG	Organizations: company, team, government ...
GPE	Geopolitical entities: country, city, state ...
LOC	Geographical locations: river, sea, mountain...
FAC	facilities: landmark, road, building, airport ...
CURR	Currency names or symbols.

Table 4: Types of named entities.

In addition to word-sense annotations, we annotated our corpus using six types of named entities listed in Table 4. As our corpus is a part of the Wojood, which is annotated with 21 types of nested named-entities (Jarrar et al., 2022), in this article we annotated SALMA with six flat entities only. We used the IOB2 tagging scheme (Sang and Veenstra, 1999), where B indicates the beginning of the entity mention, I the inside token, and O outside token.

Tag	Named Entity Mentions	Tokens in the Entity Mentions
PERS	294	568
ORG	1,123	2,108
GPE	1,086	1,295
LOC	166	318
FAC	22	59
CURR	37	41
Total	**2,728**	**4,389**

Table 5: Statistics of named entities in SALMA corpus.

We applied the NER guidelines that were used to annotate the OntoNotes5 corpus (Weischedel et al., 2011). Table 5 presents statistics about all named entities in the SALMA corpus, which shows that 4389 (about 15%) of the tokens are part of an entity mention.

4 Inter-Annotation Agreement (IAA)

To evaluate our annotations, we selected 250 annotated words from each annotator $A \in \{A_1, A_2, A_3\}$, and assigned them to a different annotator to perform double annotations. This yielded a total of 750 words (2.6% of the annotated words) divided among three pairs of annotators, $\{(A_1, A_2), (A_1, A_3), (A_2, A_3)\}$. Because

our sense annotations contain scores (i.e., not discrete values), computing IAA is not straightforward. We chose to use various evaluation metrics especially those that take ranking into consideration. The IAA metrics used are: (i) Kappa, (ii) Linear Weighted Kappa (LWK), (ii) Quadratic Weighted Kappa (QWK), (iv) Mean Average Error (MAE), and (v) Root Mean Square Error (RMSE).

Kappa is usually used when the data is nominal (Eugenio and Glass, 2004), so we set a threshold on the score ($\geq$60%) in the six categories to be able to calculate Cohen's Kappa. The senses with scores above or equal this threshold carry the intended meanings that map with the context of the targeted word (See section 3.2). Nonetheless, a more suitable metric for ranked labels is either the LWK or QWK, as specified in the following equations, which we adopt from (Vanbelle, 2016):

$$QWK = 1 - \frac{\sum_{i,j=1}^{K} \frac{(y_i - y_j)^2}{(K-1)^2} . fo_{ij}}{\sum_{i,j=1}^{K} \frac{(y_i - y_j)^2}{(K-1)^2} . fe_{ij}} \quad (1)$$

$$LWK = 1 - \frac{\sum_{i,j=1}^{K} \frac{|y_i - y_j)|}{(K-1)} . fo_{ij}}{\sum_{i,j=1}^{K} \frac{|y_i - y_j)|}{(K-1)} . fe_{ij}} \quad (2)$$

where fo_{ij} is the observed frequency of the categories (i and j) per the annotators selection, fe_{ij} is the expected frequency for both annotators' selected categories, $(y_i - y_{jx})$ denotes the distance between the categories, and K is number of categories.

Both LWK and QWK take the distance between categories into consideration, where the distance is defined as the number of categories separating the two annotators' selection. The difference is that LWK calculates the distance linearly while QWK calculates it quadratically. For measuring the ranking error deviation among annotators we used MAE and RMSE.

4.1 IAA Results

Table 6 summarizes the result of the inter-annotator-agreement, the value in parenthesis is the standard deviation among pairs of annotators. Overall, we see higher agreement among the annotators for the Modern. The higher agreement is clear from all IAA metrics and the standard deviation. We see less confidence in the Ghani annotations as the IAA dropped across all metrics with higher variability among annotators, presented in higher standard deviation. Kappa was affected the most with a drop of 11.8% when measured on the Ghani, followed by LWK with a drop of 8.73%. QWK has the smallest drop of 5.91% and also has the least variability among annotators. We believe the reason for the higher IAA on Modern is because Modern has better quality glosses compared to the Ghani, which has shorter glosses and in many cases are ambiguous. However, regardless of the lexicon used, we observed higher agreement among annotators as measured by LWK and QWK since they take advantage of the scores assigned to each gloss, while Kappa ignores the scoring information.

Metric	Lexicons	Average (STD)
Kappa	Modern	90.48 ($\pm$2.97)
	Ghani	78.68 ($\pm$8.49)
LWK	Modern	88.29 ($\pm$5.37)
	Ghani	79.56 ($\pm$9.35)
QWK	Modern	91.94 ($\pm$3.42)
	Ghani	86.03 ($\pm$5.41)
RMSE	Modern	13.44 ($\pm$3.08)
	Ghani	19.12 ($\pm$3.06)
MAE	Modern	4.46 ($\pm$2.04)
	Ghani	8.27 ($\pm$3.52)

Table 6: Inter-Annotator Agreement (IAA) average among the three linguists using different metrics.

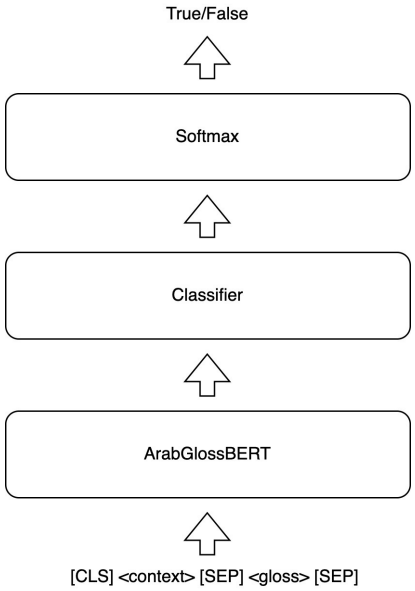

Figure 2: BERT-based TSV Architecture.

We reach similar conclusions for RMSE and MAE. Both metrics are lower for Modern compared to Ghani. The Average RMSE among all annotator pairs on the Modern is 13.44 compared

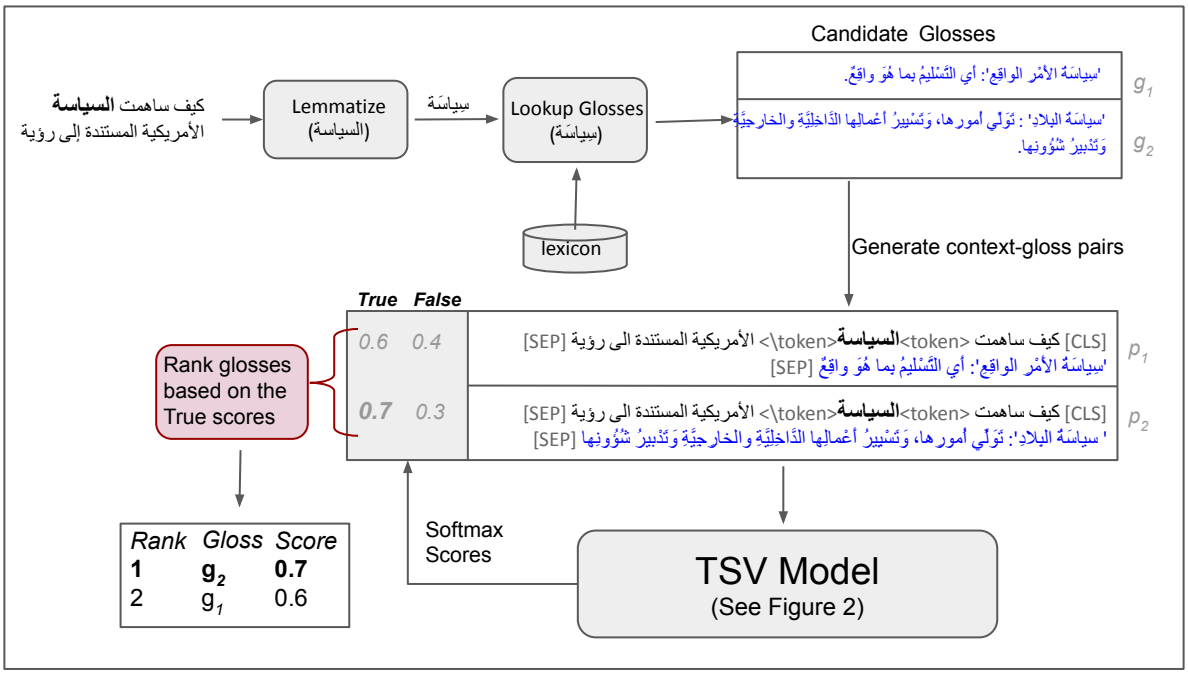

Figure 3: An end-to-end WSD using the TSV model (SALMA system).

to 19.12 for Ghani, while the average MAE for the Modern is 4.46 compared to 8.27 on the Ghani.

5 Computing WSD Baselines using SALMA

In this section, we present the baseline for Arabic WSD using our SALMA corpus. To the best of our knowledge, there are no available Arabic WSD systems to evaluate. The only available Arabic models are TSV, which are related, but not the same as WSD. In what follows, we explain the difference between WSD and TSV tasks, and propose an end-to-end WSD system using TSV.

5.1 The TSV Task

The TSV task is a binary classification task used to determine whether a pair of sentences (context and gloss) are True or False (see Figure 2). In other words, given a context c containig the target word w, and a gloss g_i, TSV aims to classify the context-gloss pair (c, g_i) as True or False. It is True if the gloss g_i is the intended sense of w in c, otherwise, it is False (Breit et al., 2020). It is important to note that TSV is different from WSD, which determines which gloss, among a set of glosses, is the intended meaning for the target word.

There are three available Arabic TSV models with the same architecture: (1) the Razzaz model, trained using 31K context-gloss pairs extracted from Modern (El-Razzaz et al., 2021); (2) the ArabGlossBERT model, trained on a larger dataset (167K context-gloss pairs) extracted from several Arabic lexicons (Al-Hajj and Jarrar, 2021); and (3) the Aug-ArabGlossBERT (D9) model, trained on an augmented data, generated using back-translation of the ArabGlossBERT dataset (Malaysha et al., 2023).

In what follows, we propose to develop an end-to-end WSD system using TSV (called SALMA system) and in Section 6, we benchmark our proposed system using the SALMA corpus.

5.2 Building WSD System Using TSV

In this section, we propose an end-to-end solution for WSD using TSV. The solution consists of the following phases (Figure 3): 1) candidate glosses lookup, 2) target sense verification, and 3) gloss ranking.

1. Candidate Glosses Lookup: given a target word w in a context c, we first lemmatize w (i.e., determine its lemma l), where we use our own in-house lemmatizer, then retrieve the set of n candidate glosses, $G = \{g_1, g_2, ..., g_n\}$, of l from the lexicon (i.e., sense inventory).

Example: the word w (السياسة *ālsyāsh*) in c (كيف ساهمت السياسة الأمريكة المستندة الى رؤية) has the lemma (سِيَاسَةٌ *siyaāsatun*) with two corresponding glosses ($\{g_1, g_2\}$) in the Ghani, as shown in Figure 3.

2. TSV: once we have the set of n candidate glosses, we input to the TSV model a set of n context-gloss pairs, $P = \{(c, g_i)|\forall g_i \in G\}$, as illustrated with (p_1, p_2) in Figure 3. The target word w in c is wrapped with special tokens "<token>w</token>", to emphasize the target word during training and testing of the TSV models. For each context-gloss pair, the TSV model returns confidence scores for the True and False labels, but the TSV model does not compare or rank glosses in this phase.

3. Gloss Ranking: we determine the intended meaning by ranking the glosses based on their True confidence scores calculated in the previous step. The gloss with the highest score is selected as the intended gloss for w.

6 Experiments and Results

6.1 Experimental Setup

To evaluate the three available Arabic TSV models using our SALMA corpus, we implemented three instances of the WSD system depicted in Figure 3, each with a different TSV model. For each word in each context in the SALMA corpus, we generated context-gloss pairs similar to the example shown in Figure 3. Because our corpus was sense-annotated using two lexicons (i.e., two sense inventories), we generated two sets of context-gloss pairs. In this way, we compute a separate baseline for each of the Modern and Ghani. We neither included annotations of digits and punctuations, nor the named-entity annotations presented in Section 3.6.

The length of the contexts may impact the WSD accuracy, so in addition to using the full context around w, we also experimented with different context sizes, $s \in \{3, 5, 7, 9, 11\}$. For example, the context size $s = 5$ means that there are two tokens before and two tokens after w.

As will be discussed in the next subsection, we evaluated three TSV models: Razzaz[2], ArabGlossBERT[3], and Aug-ArabGlossBERT(D9)[4]. We used context size $s = 11$, which gave the best results. Following the authors of these models, we did not use any signal to mark up target words in the case of the Razzaz and Aug-ArabGlossBERT(D9); however, we used UNUSED0 for ArabGlossBERT.

The experiments have been implemented in Python, specifically using the Transformers library provided by HuggigFace[5], which is used to load and test the models. To speed-up the models evaluation, we have run the codes using a GPU (SVGA II) instance, where each run took around 20 hours.

TSV Model	Lexicons	Accuracy
Razzaz	Modern	66.0%
	Ghani	68.4%
ArabGlossBERT	Modern	**84.2%**
	Ghani	77.6%
Aug-ArabGlossBERT(D9)	Modern	82.6%
	Ghani	78.7%

Table 7: WSD baselines for three TSV models, with context length = 11.

6.2 Baselines and Discussion

Table 7 presents our evaluation of the three TSV models using both Modern and Ghani with context size $s = 11$. As shown in this table, the ArabGlossBERT is the best-performing model(84.2%), which most probably because it was trained on a larger and higher quality dataset of lexicon definitions. The accuracy was calculated for nouns and verbs. We excluded the functional words as they mostly do not carry semantics.

Window	Lexicon	Accuracy Target Sense Rank			Accuracy (Top1) per POS		
		Top1	Top2	Top3	Noun	Verb	Func.
All	Modern	82.8	94.2	97.4	83.5	77.9	41.2
	Ghani	77.0	89.3	94.1	78.5	66.0	36.0
11	**Modern**	**84.2**	95.1	98.1	85.4	76.1	37.9
	Ghani	**77.6**	90.1	94.9	79.4	61.7	31.8
9	Modern	83.5	95.0	97.9	84.4	78.3	37.7
	GHani	77.3	90.1	94.8	79	63.7	32.2
7	Modern	83.8	95.1	97.9	84.8	77.4	38.9
	Ghani	77.3	90.0	94.9	79.1	62.9	31.8
5	Modern	84.0	95.1	98.1	85.3	75.6	40.0
	Ghani	77.6	90.1	94.9	79.5	61.6	31.7
3	Modern	82.8	94.4	97.6	84.4	71.8	42.1
	Ghani	77.4	90.0	94.8	79.4	59.7	32.1

Table 8: Baselines - evaluation of ArabGlossBERT on two sense inventories, with different context windows and sense orderings.

Table 8 presents further evaluation of ArabGlossBERT, which illustrates the following: (i) using Modern is better than using Ghani in all experiments. This might be because of the better quality

[2]We reproduced the TSV model using the code and data available at https://github.com/MElrazzaz/Arabic-word-sense-disambiguation-bench-mark

[3]ArabGlossBERT fine-tuned model Version 1 (CC-BY-4.0) at https://huggingface.co/SinaLab/ArabGlossBERT/tree/main

[4]Fine-tuned model D9 (CC-BY-4.0) at https://huggingface.co/SinaLab/ArabGlossBERT/tree/Augment

[5]https://huggingface.co/docs/transformers/index

of Modern glosses (refer to IAA in Section 4); (ii) While window 11 and 5 have the highest WSD accuracy, the use of context windows does not make major difference (only 1.4% for Modern and 0.6% for Ghani); (iii) the ranking of the intended sense among the top 1, 2, and 3 senses illustrates a consistent and reasonable increase in the WSD accuracy; and (iv) when evaluating the model accuracy for noun and verb, the accuracy of nouns is about 8.5% better than verbs for Modern, which might be because verbs are typically more ambiguous (Malaysha et al., 2023). The WSD accuracy for functional words is very low with both lexicons. This is because functional words are highly polysemous and their glosses describe their functions rather than semantics.

7 Conclusion

We presented SALMA, the first sense-annotated Arabic corpus. The novelty of SALMA lies in utilizing two sense inventories and named entity annotations. In addition, instead of linking a word to one intended sense, we scored all semantically related senses of each token in the corpus. The quality of the annotations was assessed using various inter-annotator agreement metrics (Kappa, LWK, QWK, MAE, and RSME). To compute a WSD baseline using our corpus, we proposed to build an end-to-end WSD system using TSV, and evaluated this system using three different TSV models. The full corpus, annotations, and the tool, are open source and publicly available on GitHub.

8 Limitations and Future Work

Although Modern provides a better quality of glosses compared with the Ghani, some of Modern's glosses are referrals, i.e., referred to another related lemma. At this stage, we annotated these referrals as senses. Nevertheless, in order to use the Modern as a general sense inventory, these referrals need to be treated differently. We plan to replace all referral glosses with the senses they refer to, which can be done semi-automatically. For missing lemmas in Modern, we plan to map between the lemmas in both lexicons and then import missing lemmas and their senses from Ghani to Modern. In this way, we expect to have a richer Arabic sense inventory. Additionally, our sense annotations are limited to the senses of a single-word lemma. We plan to annotate the corpus with multiword expressions (Jarrar et al., 2018). Furthermore, the corpus we presented in this article is limited to MSA. To extend this corpus with dialectal text, plan to sense-annotate portions of the available corpora Curras (Haff et al., 2022; Jarrar et al., 2017), Baladi (Haff et al., 2022), Nabra (Nayouf et al., 2023) and Lisan (Jarrar et al., 2023).

Acknowledgment

We would like to thank Shimaa Hamayel, Tamara Qaimari, Raghad Aburahma, Hiba Zayed, and Rwaa Assi for helping us in the corpus annotation.

References

Abdul-Ghani Abul-Azm. 2014. Al-ghani al-zaher dictionary. *Rabat: Al-Ghani Publishing Institution*.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. 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)*, pages 40–48, Online. INCOMA Ltd.

Diana Alhafi, Anton Deik, and Mustafa Jarrar. 2019. Usability evaluation of lexicographic e-services. In *The 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEE.

Hamzeh Amayreh, Mohammad Dwaikat, and Mustafa Jarrar. 2019. Lexicon digitization-a framework for structuring, normalizing and cleaning lexical entries. *Technical Report, Birzeit University*.

William Black, Sabri Elkateb, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease, Christiane Fellbaum, et al. 2006. Introducing the arabic wordnet project. In *Proceedings of the third international WordNet conference*, pages 295–300. Jeju Korea.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

Sami Boudelaa and William D Marslen-Wilson. 2004. Abstract morphemes and lexical representation: The cv-skeleton in arabic. *Cognition*, 92(3):271–303.

Sami Boudelaa, Friedemann Pulvermüller, Olaf Hauk, Yury Shtyrov, and William Marslen-Wilson. 2010. Arabic morphology in the neural language system. *Journal of cognitive neuroscience*, 22(5):998–1010.

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2020. Wic-tsv: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1635–1645.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a free corpus of Polish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3218–3222, Istanbul, Turkey. European Language Resources Association (ELRA).

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3269–3275. Association for Computational Linguistics.

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, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab worlds. *Commun. ACM*, 64(4):72–81.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mohammed El-Razzaz, Mohamed Waleed Fakhr, and Fahima A Maghraby. 2021. Arabic gloss wsd using bert. *Applied Sciences*, 11(6):2567.

Bilel Elayeb. 2019. Arabic word sense disambiguation: a review. *Artif. Intell. Rev.*, 52(4):2475–2532.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Sana Ghanem, Mustafa Jarrar, Radi Jarrar, and Ibrahim Bounhas. 2023. A benchmark and scoring algorithm for enriching arabic synonyms. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*, pages 215–222. Global Wordnet Association.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022)*, Marseille, France.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3507–3512. Association for Computational Linguistics.

Mustafa Jarrar. 2005. *Towards Methodological Principles for Ontology Engineering*. Ph.D. thesis, Vrije Universiteit Brussel.

Mustafa Jarrar. 2006. Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In *Proceedings of the 15th international 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 Meeting on Arabic Ontologies and Semantic Networks*. ALECSO, Arab League.

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 Applications of Natural Language to Information Systems (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 Language, Data and Knowledge (LDK 2019)*, volume 2402, pages 29–33. CEUR Workshop Proceedings.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: An annotated corpus for the palestinian arabic dialect. *Journal Language Resources and Evaluation*, 51(3):745–775.

Mustafa Jarrar, Eman Karajah, Muhammad Khalifa, and Khaled Shaalan. 2021. Extracting synonyms from bilingual dictionaries. In *Proceedings of the 11th International Global Wordnet Conference (GWC2021)*, pages 215–222. Global Wordnet Association.

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.

Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Waahlisch. 2023. Lisan: Yemeni, irqi, libyan, and sudanese arabic dialect copora with morphological annotations. In *The 20th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*. IEEE.

Sanad Malaysha, Mustafa Jarrar, and Mohammad Khalilia. 2023. Context-gloss augmentation for improving arabic target sense verification. In *Proceedings of the 12th International Global Wordnet Conference (GWC2023)*. Global Wordnet Association.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic). In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 24–36. Association for Computational Linguistics.

Marek Maziarz, Maciej Piasecki, and Stan Szpakowicz. 2012. Approaching plwordnet 2.0. In *Proceedings of 6th International Global Wordnet Conference, The Global WordNet Association*, pages 189–196.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Simonetta Montemagni and Guglielmo Venturi. 2003. Building sense-tagged corpora for all: The itec project. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Eman Naser-Karajah, Nabil Arman, and Mustafa Jarrar. 2021. Current trends and approaches in synonyms extraction: Potential adaptation to arabic. In *Proceedings of the 2021 International Conference on Information Technology (ICIT)*, pages 428–434, Amman, Jordan. IEEE.

Amal Nayouf, Mustafa Jarrar, Fadi zaraket, Tymaa Hammouda, and Mohamad-Bassam Kurdy. 2023. Nâbra: Syrian arabic dialects with morphological annotations. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*. ACL.

Ahmed Mukhtar Omar. 2008. Contemporary arabic dictionary.(i1). *World of Books, Cairo, Egypt. Retrieval Date*, 14(8):2020.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global Wordnet Conference, GWC 2014, Tartu, Estonia, January 25-29, 2014*, pages 236–245. University of Tartu Press.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 1(8).

Karin C Ryding. 2014. *Arabic: A linguistic introduction*. Cambridge University Press.

Erik F. Tjong Kim Sang and Jorn Veenstra. 1999. Representing text chunks. In *EACL 1999, 9th Conference of the European Chapter of the Association for Computational Linguistics, June 8-12, 1999, University of Bergen, Bergen, Norway*, pages 173–179. The Association for Computer Linguistics.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A. Smith. 2012. Coarse lexical semantic annotation with supersenses: An arabic case study. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers*, pages 253–258. The Association for Computer Linguistics.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU - multilingual corpus). *Int. J. Asian Lang. Process.*, 22(4):161–174.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Heike Telljohann, Erhard Hinrichs, and Ra Ubler. 2004. The tüba-d/z treebank: Annotating german with a context-free backbone. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*.

Sophie Vanbelle. 2016. A new interpretation of the weighted kappa coefficients. *Psychometrika*, 81(2):399–410.

Warren Weaver. 1949/1955. Translation. In William N. Locke and A. Donald Boothe, editors, *Machine Translation of Languages*, pages 15–23. MIT Press, Cambridge, MA. Reprinted from a memorandum written by Weaver in 1949.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 ldc2013t19. *Philadelphia: Linguistic Data Consortium*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Boon Peng Yap, Andrew Koh, and Eng Siong Chng. 2020. Adapting bert for word sense disambiguation with gloss selection objective and example sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 41–46.