Active Learning for Multidialectal Arabic POS Tagging

Diyam Akra[λ] **Mohammed Khalilia**[λ] **Mustafa Jarrar**[σ,λ]
[λ] Birzeit University, Palestine
[σ] Hamad Bin Khalifa University, Qatar
{dakra,mkhalilia,mjarrar}@birzeit.edu

Abstract

Multidialectal Arabic POS tagging is challenging due to the morphological richness and high variability among dialects. While POS tagging for MSA has advanced thanks to the availability of annotated datasets, creating similar resources for dialects remains costly and labor-intensive. Increasing the size of annotated datasets does not necessarily result in better performance. Active learning offers a more efficient alternative by prioritizing annotating the most informative samples. This paper proposes an active learning approach for multidialectal Arabic POS tagging. Our experiments revealed that annotating approximately $15,000$ tokens is sufficient for high performance. We further demonstrate that using a fine-tuned model from one dialect to guide the selection of initial samples from another dialect accelerates convergence—reducing the annotation requirement by about $2,000$ tokens. In conclusion, we propose an active learning pipeline and demonstrate that, upon reaching its defined stopping point of $16,000$ annotated tokens, it achieves an accuracy of 97.6% on the Emirati Corpus.

1 Introduction

POS tagging assigns a part of speech to each word in a sequence. It is a crucial step in various NLP tasks such as Named Entity Recognition (NER) (Güngör et al., 2018; Hamad et al., 2025; Jarrar et al., 2024b, 2023a; Liqreina et al., 2023; Jarrar et al., 2022), machine translation (Yazar and Kiliç, 2025), text summarization (Nambiar et al., 2023), word sense disambiguation (Khalilia et al., 2024; Al-Hajj and Jarrar, 2021), and synonymous extraction (Naser-Karajah et al., 2021; Jarrar et al., 2021) . In Arabic, POS tagging is particularly challenging due to its highly inflectional morphology across MSA and dialects(Darwish et al., 2021). First, Arabic's inflectional and derivational nature leads to numerous affixes that modify stems (Ryding, 2014). Second, dialectal variation adds complexity,

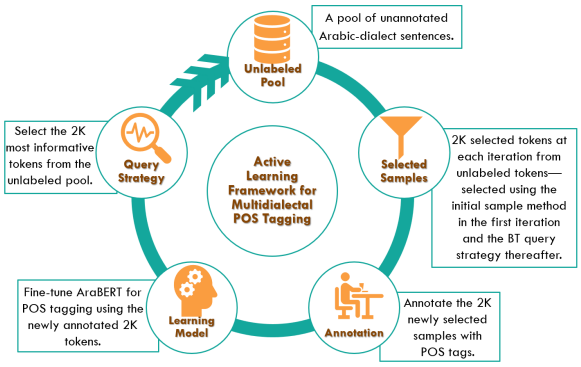

Figure 1: Active learning framework for multidialectal POS tagging

as an example, different affixes for the progressive tense—e.g., (ب / *b*) in Levantine and Egyptian, (د / *d*) in Iraqi, and (ك / *k*) in Moroccan. Third, each dialect includes unique vocabulary absent in MSA and other dialects (Hamed et al., 2025; Jarrar et al., 2023c), complicating POS tagging further.

The importance of POS tagging, combined with the complexity of Arabic morphology, has motivated extensive research aimed at improving POS tagging accuracy, coverage, and speed. Approaches differ in datasets, tag sets, and NLP techniques. Recent methods have achieved strong results on MSA, supported by a large morphologically annotated corpus (Inoue et al., 2021). Research on POS tagging for Arabic dialects is still in its early stages and has yet to produce a tagger that covers most dialects (Inoue et al., 2021; Darwish et al., 2020). This is due to morphological challenges specific to dialects, as discussed earlier, and the lack of morphologically annotated corpora. Creating such corpora for each dialect is costly.

Active learning approaches, which prioritize annotating the most informative and representative data samples (Kirsch, 2024), effectively reduce annotation cost and effort while maintaining or improving model performance. Studies show that it can cut labeling effort by up to 50% (Settles,

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24960–24973
November 4-9, 2025

2009; Sener and Savarese, 2018), and recent work demonstrates strong results in many domains: 40% labeled data sufficed in document image classification (Krishnan and Satish, 2023), 20.83%–24.34% in retail recognition achieved 95% performance (Bhatia and Kumar, 2024), and over 90% time reduction was reported in biomedical analysis (Chen et al., 2024). Thus, active learning provides a cost-effective method for developing multidialectal POS taggers by producing high-quality morphologically annotated corpora with minimal annotation.

This paper evaluates active learning for POS tagging of four Arabic dialects (Palestinian, Syrian, Egyptian, and Emirati) along with MSA. The evaluation examines four key factors affecting active learning cycle (see Figure 1): learning models, initial sample selection methods, annotation levels, and query strategies. Results show that annotating approximately 15,000 informative tokens from MSA, Palestinian, Egyptian, and Syrian corpora is sufficient to achieve high performance, reaching 97.3%, 95.5%, 94.8%, and 92.9%, respectively. It also explores using fine-tuned models from other dialects for initial sample selection, showing a 2.5–5% accuracy gain in the first iteration and faster convergence. Finally, it introduces an active learning pipeline applicable to any unannotated dialect. In summary, our contributions are:

- Evaluation of active learning for multidialectal Arabic POS tagging (29 experiments).
- Insightful analysis and 10 assessments for initial sample selection across dialects.
- Active learning pipeline, e.g., achieved 97.6% with only 16,000 tokens in Emirati.

The paper is structured as: Section 2 reviews related work; Sections 3 and 4 describe the corpora and the evaluation methodology; Sections 5 and 6 present results and findings; Section 7 proposes an active learning pipeline; and we conclude in Section 8.

2 Related Work

Various methods have been used for Arabic and dialectal POS tagging, evolving from limited rule- and database-based approaches (Boudchiche et al., 2017; Buckwalter, 2004; Graff et al., 2009; Jarrar et al., 2024a), through classical machine learning models (Pasha et al., 2014; Darwish et al., 2014, 2017, 2018), to neural networks (Darwish et al., 2017, 2018; Kondratyuk et al., 2018), with recent advancements focusing on transformer-based models that achieved state-of-the-art results (Inoue et al., 2021; Kondratyuk, 2019).

All of the aforementioned machine learning and deep learning approaches rely on passive learning, where a large dataset is required for training.

As a result, most MSA approaches relied on one or more versions of the Arabic Treebank (Maamouri et al., 2004), which is sourced from different Arabic news sources. The POS annotation procedure used in the Arabic Treebank (ATB) involves segmenting raw input, applying the Buckwalter Morphological Analyzer (Buckwalter, 2004) to generate candidate tags, and having human annotators select the correct POS. Review passes follow to correct errors and ensure quality. While automatic analysis helps improve efficiency, substantial manual correction is needed, making the process time-consuming and reliant on trained annotators. Similarly, the Egyptian Arabic Treebank (ARZTB) (Maamouri et al., 2018), which is sourced from various Egyptian informal data sources (discussion forums, text messaging, and chat), was built using the same approach.

For the Levantine dialects, the first morphologically annotated corpus was the Palestinian corpus **Curras** (Jarrar et al., 2017; Habash et al., 2015; Jarrar et al., 2014), annotated using the DIWAN tool (Al-Shargi and Rambow, 2015), which integrated outputs from MADAMIRA (Pasha et al., 2014) and was completed by two annotators over one year. The Lebanese corpus **Baladi** (Haff et al., 2022) was manually annotated by four annotators over ten months using AnnoSheet, a Google Sheets-based tool with JavaScript aids and Curras-based suggestions. The Syrian corpus **Nabra** (Nayouf et al., 2023) adopted this methodology, using the Tawseem portal for annotation with smart features and was completed by nine annotators in one year.

On the other hand, Lisan (Jarrar et al., 2023c), which cover multiple dialects, followed the same annotation methodology as Curras, Baladi, and Nabra and was manually annotated using the ADAT tool by 35 annotators over two years, with separate teams per dialect. Similarly, the Gumar corpus (Khalifa et al., 2018) used the MADARi tool for manual annotation by a dedicated team. All these corpora were sourced from diverse platforms such as Facebook, Twitter, blogs, forums, YouTube, and TV shows.

None of these corpora have utilized active learn-

ing approaches in their annotation process, resulting in extended annotation timelines to produce corpora suitable for POS tagger training. Active learning for POS tagging remains underexplored, with a notable recent study by (Chaudhary et al., 2021) applying it to six languages—excluding Arabic. To our knowledge, this work is the first to apply active learning strategies for POS tagging to Arabic and its dialects, demonstrating that a large annotated corpus is not required; annotating only the most informative tokens yields comparable results to training on the full dataset.

3 Datasets Preparation and Composition

This paper investigates four Arabic dialects—Palestinian, Syrian, Egyptian, and Emirati—alongside MSA, using seven datasets annotated primarily with POS and other morphological features: (1) Arabic Treebank Part 3 v3.2 (ATB) (Maamouri et al., 2010), (2) SALMA (Jarrar et al., 2023b), (3) BOLT Egyptian Treebank (ARZTB) (Maamouri et al., 2018), (4) Curras (Haff et al., 2022), (5) Nabra (Nayouf et al., 2023), (6) Baladi (Haff et al., 2022), (7) Gumar (Khalifa et al., 2018)[1].Table 1 summarizes these datasets, and Section 2 provides more details.

Dataset	Tokens	Unique Tokens
ATB (MSA)	339,710	51,820
SALMA (MSA)	34,253	8,718
ARZTB (Egyptian)	400,448	66,899
Curras (Palestinian)	56,700	16,573
Nabra (Syrian) + Baladi (Lebanese)	69,582	24,664
Gumar (Emirati)	201,596	22,924

Table 1: Datasets Statistics

As the seven datasets use different POS tagsets, we employed a two-step unification process to align all corpora with the tagset used in ALMA (Jarrar et al., 2024a, 2018), a lemmatizer and POS tagger for Arabic that is part of the morphology module in SinaTools (Hammouda et al., 2024). ALMA relies on the Qabas lexicographic database (Jarrar and Hammouda, 2024), which links 110 lexicons (Jarrar and Amayreh, 2019; Jarrar et al., 2019) and the Arabic Ontology (Jarrar, 2021, 2011)—and 12 morphologically annotated corpora, including those used in addition to QuranMorph (Akra et al., 2025). Our two unification steps are:

- Label normalization: Tags that express the same category but are named differently were unified. **Example:** DEM_PRON (Curras, Nabra, Baladi, ARZTB, ATB) and PRON_DEM (Gumar) were unified as DEM_PRON in every corpus.
- Value-level mapping: Tokens that carry equivalent meanings but receive different tags across corpora were manually harmonized. **Example:** the numeral "7" is labeled NOUN_NUM in ARZTB and ATB but DIGIT in the other corpora; we mapped all occurrences to DIGIT.

See full mapping list in Table 3 in Appendix A.

After unifying the tags across all datasets, the datasets —except SALMA, which was used for testing only—were split following the methodology in (van der Goot, 2021): TRAIN for training, TUNE (5% of TRAIN) for model selection, DEV for initial evaluation, and TEST for final evaluation. Table 2 shows the token counts for TRAIN, DEV, and TEST. Some tokens were excluded due to missing POS annotations in the original datasets.

Dataset	TRAIN	DEV	TEST
ATB	221,262	39,790	68,242
Egyptian	267,555	81,650	38,040
Curras	44,600	5,698	5,808
Nabra + Baladi	56,035	6,447	6,927
Gumar	161,441	20,138	20,017

Table 2: Splits statistics for each dataset

4 Active Learning Methodology

This section presents the methodology for identifying the most effective query strategy to achieve high performance with minimal annotated data. To ensure fair comparison across query strategies, we first evaluated various pretrained learning models on Arabic dialects from the literature to select the most suitable one (subsection 4.1). We then ran preliminary experiments to determine the best initial sample selection method (subsection 4.2). Finally, with both components selected, we conducted experiments to identify the optimal query strategy (subsection 4.4).

[1]ATB and ARZTB are licensed from the Linguistic Data Consortium (LDC), while Curras, Baladi, Nabra and SALMA are CC-BY-4.0, and Gumar is under a custom license

In our methodology, the active learning process continues until all tokens in each corpus's training set are annotated. This enables us to track the impact of each newly selected batch of tokens in each iteration and evaluate how increased annotation affects performance.

4.1 Model Selection

To identify the best query strategy for multidialectal Arabic POS tagging, we first evaluate the best performing model. To ensure an unbiased comparison, both the initial sample selection and query strategy are fixed to random, isolating the impact of the learning model.

Accordingly, we assess three commonly pretrained BERT models for Arabic dialects: AraBERT ("arabertv02-twitter") (Antoun et al., 2020), MARBERTV2 (Abdul-Mageed et al., 2021), and CAMeLBERT-Mix (Inoue et al., 2021). The evaluation was conducted on Curras, as Palestinian Arabic was covered across all three models.

4.2 Initial Sample Selection Method

The second step in identifying the best query strategy for multidialectal Arabic POS tagging is selecting a consistent initial sample selection method. To determine the most effective approach, we conducted preliminary experiments using a random query strategy to isolate the effect of the initial sample selection method.

The initial sample comprises sentences used in the first active learning iteration to train the model described in 4.1. We evaluated three selection methods: (i) Most Dissimilar Sentences, (ii) Probabilistic Selection, (iii) Longest Sentences. The first two rely on a TF-IDF matrix. Most Dissimilar selects sentences with the lowest cosine similarity, while Probabilistic Selection applies softmax to the normalized top-k mean of matrix rows and samples indices accordingly. Longest Sentences are selected based on word count. Additionally, we explored prediction-based methods using a model fine-tuned on one dialect to guide initial sample selection in another.

To ensure consistency across corpora, we fixed the initial sample size at 2000 tokens. Although sentences are selected, we apply a token cap—once 2000 tokens are reached, remaining tokens and sentences are excluded.

4.3 Annotation Level Selection

Another key consideration is the annotation level, which can be either sentence-level (selecting full sentences) or token-level (selecting individual tokens). We adopt token-level annotation, as sentences often include redundant functional words that add little value. Instead, we aim to target key tokens the model finds uncertain.

The token-level annotation raises the issue of handling unselected tokens within selected sentences during training. We evaluate two strategies from the literature: (1) Mask-All-Unknowns: the model is fed full sentences, but loss is computed only on the selected tokens; (2) Drop-All-Unknowns: unselected tokens are removed entirely, resulting in incomplete sentence inputs. For example, given the sentence (بديش العب هسا برا الدار / *bdyš ālʿb hsā brā āldār*), if only (برا / *brā* ,بديش / *bdyš*) are annotated, Mask-All-Unknowns computes loss only on these words while the full sentence is used as input, whereas Drop-All-Unknowns feeds only (برا / *brā* ,بديش / *bdyš*) to the model. Both methods are discussed in (Vacareanu et al., 2024).

We assess these approaches using the Curras corpus for Palestinian Arabic, applying the best learning model, best initial sample selection method, and a random query strategy.

4.4 Query Strategy Selection

we outline the query strategies that used to identify the most effective method for achieving high performance with minimal annotated data.

- Random: Selects 2000 tokens at random from the unannotated data pool in each iteration.
- Prediction Entropy: We selected 2000 tokens with the highest entropy in their tag probability distributions, formally a token x is selected from unlabeled pool U based on (Roy and McCallum, 2001):

$$\underset{x \epsilon U}{\operatorname{argmax}} - \sum_{i=1}^{c} P(y_i|x) \log P(y_i|x)$$

- Breaking Ties: We selected 2000 tokens with the smallest difference between the top two predicted labels, formally a token x is selected from unlabeled pool U based on (Scheffer et al., 2001):

$$\underset{x \epsilon U}{\operatorname{argmin}} P(y_{l1}|x) - P(y_{l2}|x)$$

where l_1 and l_2 are the most and second most likely labels.

- Least Confidence: Selects 2000 tokens with the lowest probability for the top predicted label, formally a token x is selected from unlabeled pool U based on (Culotta and McCallum, 2005)

$$\underset{x \epsilon U}{\operatorname{argmax}}\, 1 - P(y_{l1}|x)$$

where l_1 is the most likely label.

- Most Common Ambiguous Words: We propose a new strategy to select 2,000 tokens that receive multiple predicted tags across various contexts, indicating ambiguity. In each iteration, the number of distinct tags predicted for each token is computed, tokens are ranked in descending order by this count, and the top 2,000 are chosen.

5 Evaluation Results

This section presents the evaluation results based on the methodology outlined in Section 4, the selection of the best model (5.2), initial sample selection method (5.3), unselected token handling strategy (5.4), best query strategies (5.5), and use of a fine-tuned model from one dialect to select the initial sample in another (5.6).

5.1 Experimental Setup

All models were implemented with the Hugging Face Transformers library (Wolf et al., 2020) and trained in PyTorch (Paszke et al., 1912). Unless stated otherwise, we used the same hyper-parameter configuration in every experiment:10 epochs per active-learning iteration, a learning rate of $5e{-}5$, a batch size of 6, and a fixed random seed (12345).

The number of active-learning iterations was proportional to the size of the training corpus: it ranged from 22 iterations for the smallest dataset (Curras) to 133 iterations for the largest (ARZTB). Experiments were run on a server with 62GiB RAM, 1.2TB disk storage, and a single NVIDIA T4 GPU. Total wall-clock time consequently varied with corpus size, from about 4.5 hours for Curras to roughly 65 hours for ARZTB.

Additional parameters—learning model, initial sample selection, and unselected token handling strategy—were chosen based on preliminary results.

5.2 Model Evaluation

Figure 2 shows that AraBERTV2 and MARBERTV2 perform similarly, but AraBERTV2 slightly outperforms MARBERTV2 on Curras (Palestinian). Given that the state-of-the-art result is 94%, AraBERTV2 exceeds it at 18,000 tokens with 94.1%, while MARBERTV2 requires 20,000 tokens and CAMelBERT-Mix needs more. Therefore, AraBERTV2 is selected as the learning model for subsequent experiments.

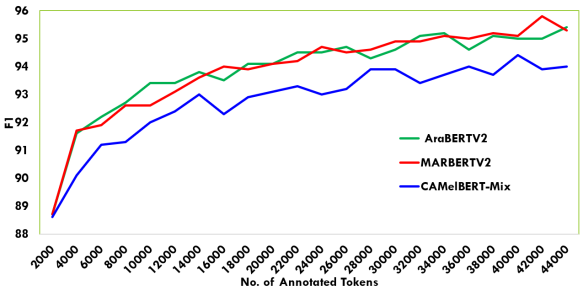

Figure 2: Learning Models Experiments-Curras

5.3 Initial Sample Selection Evaluation

Figure 3 compares initial sample selection methods, excluding the approach using a fine-tuned model from another dialect, which is evaluated separately later. Among Most Dissimilar Sentences, Probabilistic Sampling, and Longest Sentences, the Most Dissimilar Sentences method performs best. This is likely because it captures a representative subset of the dataset. However, the overall performance shows minimal sensitivity to the initial sample method, reinforcing that the query strategy plays the most critical role. Therefore, Most Dissimilar Sentences is adopted for all subsequent experiments.

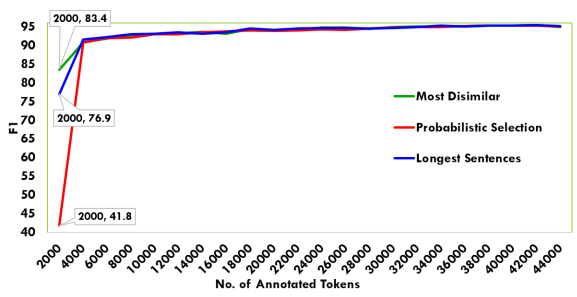

Figure 3: Initial Sample Selection Method Experiments-Curras

5.4 Annotation Level Evaluation

Figure 4 compares methods for handling unselected tokens, showing that Mask-All-Unknowns slightly outperforms Drop-All-Unknowns during the first seven iterations due to retained context. As more tokens become annotated, the context improves in Drop-All-Unknowns; both methods converge in performance. Therefore, Mask-All-Unknowns is adopted for the remaining experiments to ensure better early-stage performance with minimal annotation as aimed.

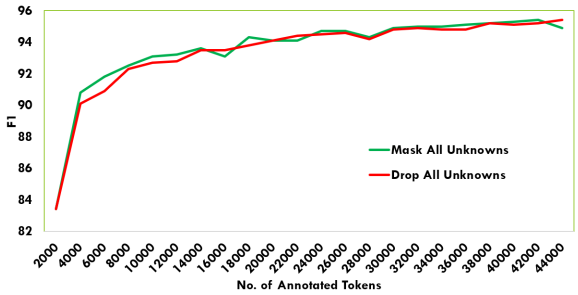

Figure 4: Handling Unselected tokens Experiments-Curras

5.5 Query Strategy Evaluation

Notably, all prior preliminary experiments maintained a fixed query strategy which is Random, focusing instead on identifying parameters to be held constant. We now present the results of the query strategy experiments. Figures 5, 6, 7, and 8 show the performance of each query strategy across dialects.In Figure 5, covering Curras, Breaking Ties, Least Confidence, and Entropy outperform the state-of-the-art, reaching F1 scores of 94.7%, 94.6%, and 94.5% with only 8,000 tokens (about 18% of the training set). These results confirm that high performance is achievable with partial annotation, significantly reducing time and cost. Notably, the Most Ambiguous Words strategy performs the worst, behaving similarly to random selection. This is likely because it selects tokens with the most diverse tag predictions across contexts without considering model confidence. Future improvements to this method should factor in confidence levels.

The same pattern appears in Nabra (Syrian) (Figure 6), where Breaking Ties, Least Confidence, and Entropy reach F1 scores of 93%, 93%, and 92.9% using only 18,000 tokens (32% of the training data), with no performance gains from further annotation. Similarly,in ARZTB (Egyptian) (Figure 7), the state-of-the-art is surpassed at only 5.2%

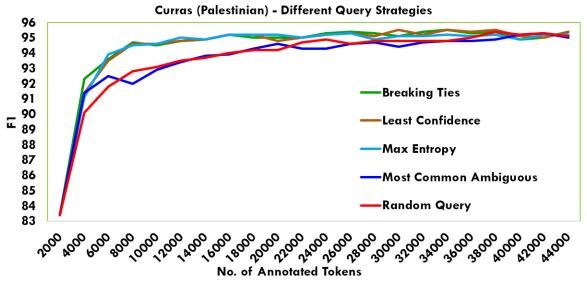

Figure 5: Curras (Palestinian) with Different Query Strategies

of the training data, reaching 94.7%. Finally, in ATB (MSA) (Figure 8), a 98% F1 score is achieved with just 46,000 tokens (23% of the training data).

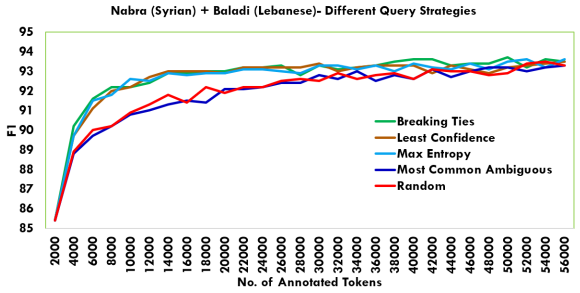

Figure 6: Nabra (Syrian) with Different Query Strategies

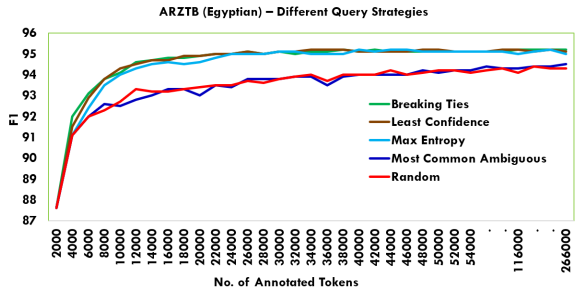

Figure 7: ARZTB (Egyptian) with Different Query Strategies

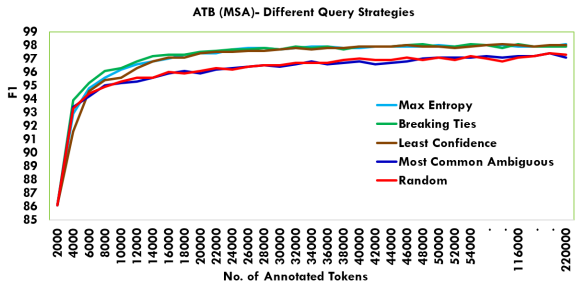

Figure 8: ATB (MSA) with Different Query Strategies

5.6 Another-Dialect Model for Initial Sample Selection Evaluation

Experiments in Figures 9, 10, 11, 12, and 13 aim to examine the impact of using a model fine-tuned on one dialect as an inference model for another dialect, then selecting the tokens the model is uncertain about as the initial sample for that dialect. (1) Figure 9 shows that using a model finetuned on Curras (Palestinian-Levantine) to guide selection for Nabra (Syrian-Levantine) yields a 5% performance boost in the first iteration and reaches 93.2% with only 12,000 tokens (vs. 22,000 without a fine-tuned model on Curras). The reverse—using Nabra for Curras—shows similar benefits (Figure 10), reaching 94.5% with just 6,000 tokens (vs. 8,000). (2) Figure 11 indicates that using an ATB (MSA) model for Curras boosts first-iteration performance by 4.7%. (3) Figure 12 shows a 2.3% gain in first-iteration when using ARZTB (Egyptian-a different dialect family) for Nabra (Syrian), and reaching 93.2% at 14,000 tokens (vs. 22,000). (4) Lastly, Figure 13 uses a model fine-tuned incrementally on ATB (MSA) → Curras (Palestinian) → Nabra (Syrian) for ARZTB (Egyptian), further validating the effectiveness of this approach.

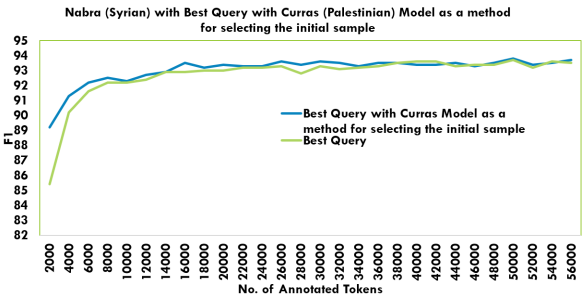

Figure 9: Nabra (Syrian) with Curras (Palestinian)

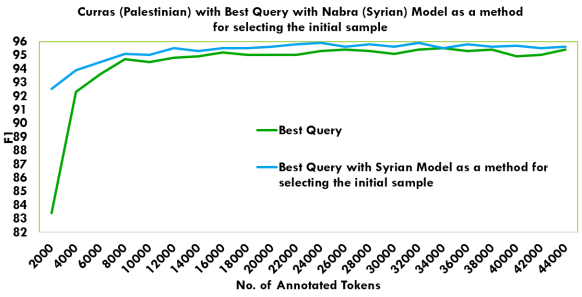

Figure 10: Curras (Palestinian) with Nabra (Syrian)

6 Discussion

This section discusses and analyzes the results, beginning with an extra validation step to check for potential overfitting (Section 6.1). Then, Sections 6.2, 6.3, and 6.4 present our findings with corresponding analysis.

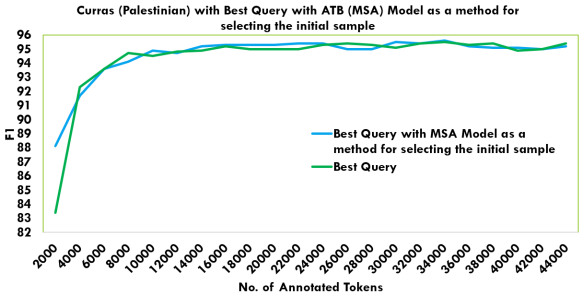

Figure 11: Curras (Palestinian) with ATB (MSA)

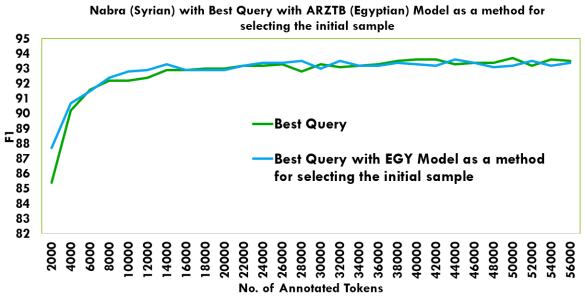

Figure 12: Nabra (Syrian) with ARZTB (Egyptian)

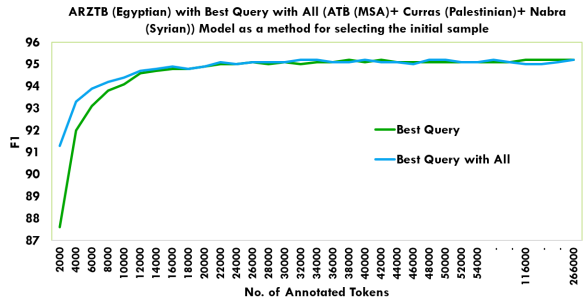

Figure 13: ARZTB (Egyptian) with All (Curras, Nabra, MSA)

6.1 Extra Validation

To evaluate our methodology and check for potential overfitting, we tested the model at each iteration—while training on Curras (Palestinian) with the Breaking Ties query strategy—not only on the Curras test set, but also on Nabra (Syrian) and SALMA (MSA). As shown in Figure 14, the model exhibited similar behavior across all three datasets. Through iterative selection and annotation of the most informative tokens from Curras, the model improved performance not only on Curras but also on the unseen Nabra and SALMA datasets, despite having no prior exposure to them during training.

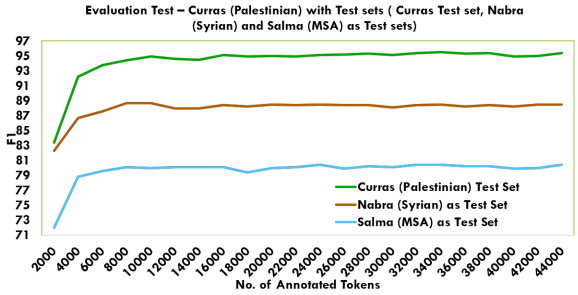

Figure 14: Evaluation Test – Curras (Palestinian) with Test it on Nabra (Syrian), Salma (MSA)

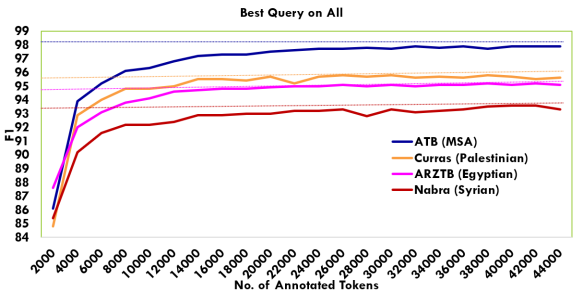

Figure 15: Performance on All corpora using Best Active Learning Method

6.2 Best Query Strategy

The first key finding is that all three uncertainty-based strategies—Breaking Ties, Least Confidence, and Entropy—achieved similarly high performance, as they all rely on prediction scores for token selection. However, Breaking Ties consistently converged earlier, probably due to its focus on tokens with the smallest margin between top two predictions, capturing uncertainty sooner.

To understand the slight performance differences, we analyzed selected tokens at iterations 3, 5, and 7 on the Curras dataset. All strategies prioritized purely dialectal words (e.g., بس / *bs* , فش / *fš*) and tokens with dialectal prefixes (e.g., ب / *b* /PROG_PART) or suffixes (e.g., ش / *š* /NEG_PART), which are typically ambiguous. Early selection of such tokens helped reduce confusion, and the best-performing strategy appeared to target more of these informative tokens earlier.

6.3 Analysis Stabilization

The second notable finding across all experiments is that model performance stabilizes around $15,000$ informative tokens, as illustrated in Figure 15. This trend is further supported by the selected tokens distributions to the overall training set in the first eight iterations from Curras (Palestinian) in Figure 17. Initially, active learning targets dense and diverse regions—typically ambiguous or uncertain cases—allowing the model to quickly learn essential pattern distinctions. Over time, newly selected tokens increasingly cover more of the training set's distribution; finally, by the end of eight iterations (totaling $16,000$ tokens), the selections are broadly distributed (see last plot in Figure 17). This wide coverage explains the degrading contribution of later tokens, resulting in performance stabilization.

6.4 Comparing Between Dialects

Figure 15 also shows that Nabra (Syrian) underperformed, prompting a quantitative analysis that revealed rare and ambiguous Syrian words like جودلي / *ğwdly* , نبريج / *nbryğ* , شن / *šn* , خاشوكة / *hāšwkh* , سقرق / *sqrq* . Conversely, ATB (MSA) performed best due to its more systematic morphology, where verbs and nouns follow consistent forms, unlike dialects where, for example, a noun like جودلي / *ğwdly* can resemble a verb.

7 Proposed Active Learning Pipeline

Based on the evaluation in 5 and findings in 6, we propose an active learning pipeline for multi-dialectal Arabic POS tagging on new unannotated datasets (see Figure 1). The pipeline consists of three phases: (1) selecting the initial sample using a model fine-tuned on another dialect. (2) applying the Breaking Ties strategy to select 2,000 new tokens per iteration. (3) stopping once performance stabilizes. As discussed earlier, stabilization typically occurs at approximately $15,000$ tokens. Therefore, the stopping criterion is set at $16,000$ tokens to ensure convergence. This pipeline was validated on the Gumar (Emirati) Corpus, reaching a performance of 97.6% at stopping point. (see Figure 16).

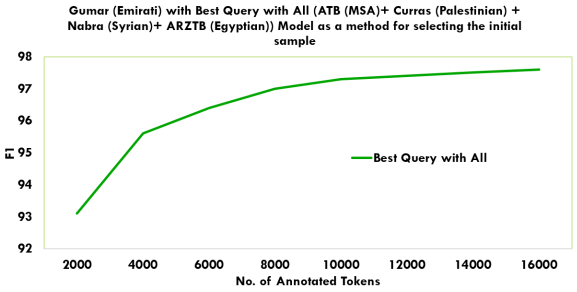

Figure 16: Gumar (Emirati) with Best Query with All

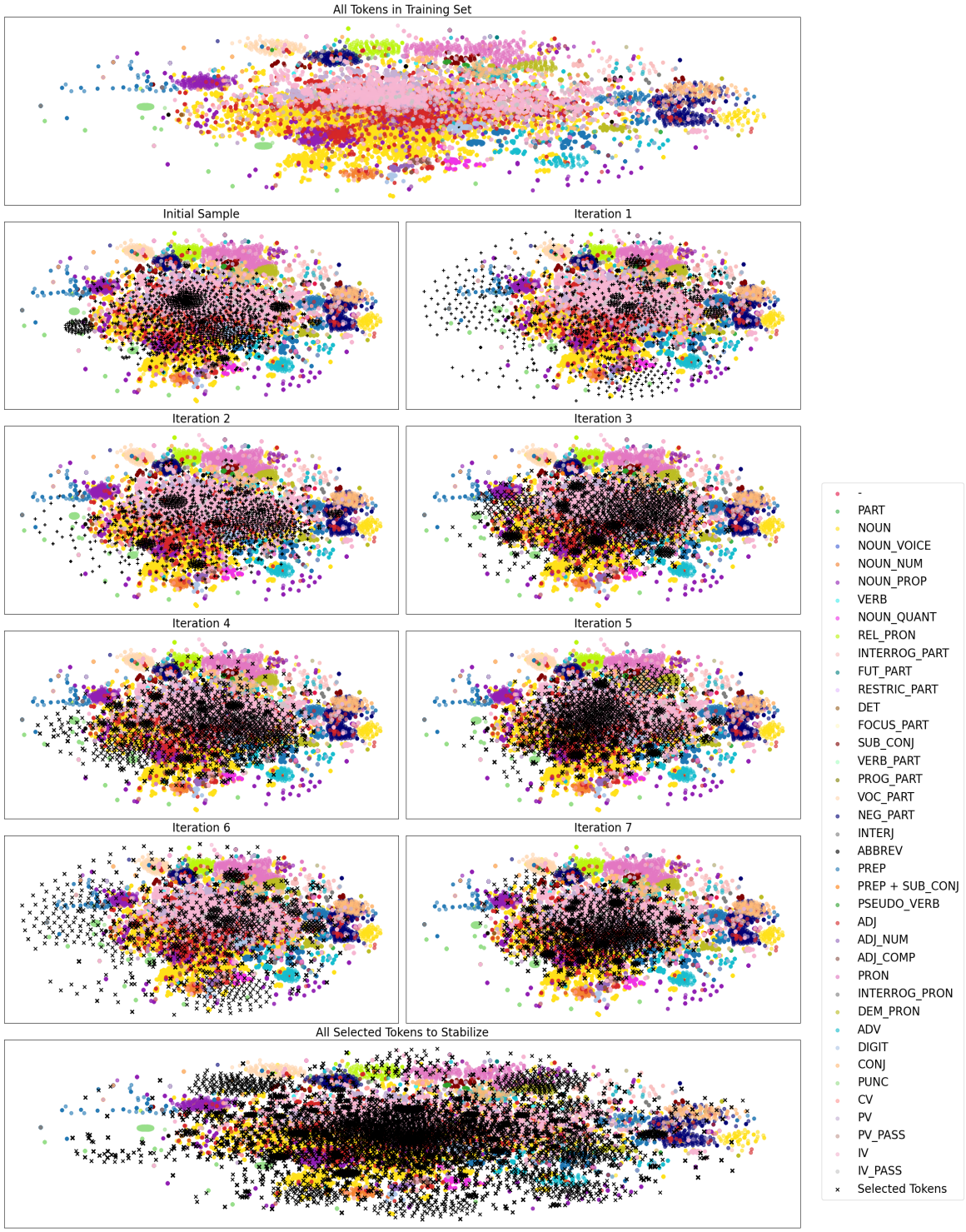

Figure 17: t-SNE visualization of selected tokens distribution on the first eight iterations of the Curras (Palestinian) training set. Notably, the Last plot is for all selected tokens in eight iterations together.

8 Conclusion

This paper evaluated using active learning for POS tagging in MSA and Arabic dialects—Palestinian, Syrian, and Egyptian. Selecting $15,000$ informative tokens per corpus proved sufficient for high performance. Using a fine-tuned model from another dialect with the Breaking Ties strategy yielded the best results. Thus, new dialectal POS taggers should adopt active learning to reduce annotation effort and cost.

9 Limitations

Like many deep learning approaches, our work relies heavily on GPU resources. However, unlike typical training setups where the model is trained once on a fixed dataset, active learning requires repeated training cycles as new data points are iteratively selected and added to the training set. This iterative nature significantly increases the overall computational cost and GPU demand. An additional limitation arises when working with large-scale datasets, as both the training time and resource requirements grow substantially with each active learning iteration, potentially impacting scalability and experimentation efficiency.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Diyam Akra, Tymaa Hammouda, and Mustafa Jarrar. 2025. QuranMorph: Morphologically Annotated Quranic Corpus. Technical report, Birzeit University.

Moustafa Al-Hajj and Mustafa Jarrar. 2021. LUBZU at SemEval-2021 Task 2: Word2Vec and Lemma2vec Performance 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 Linguistics.

Faisal Al-Shargi and Owen Rambow. 2015. DIWAN: A dialectal word annotation tool for Arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 49–58, Beijing, China. Association for Computational Linguistics.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Arpit Bhatia and Rohit Kumar. 2024. Active learning for retail product recognition with minimal annotations. *SN Computer Science*, 5(1):1–12.

Mohamed Boudchiche, Azzeddine Mazroui, Mohamed Ould Abdallahi Ould Bebah, Abdelhak Lakhouaja, and Abderrahim Boudlal. 2017. Alkhalil morpho sys 2: A robust arabic morpho-syntactic analyzer. *Journal of King Saud University - Computer and Information Sciences*, 29:141–146.

Tim Buckwalter. 2004. Buckwalter arabic morphological analyzer version 2.0 ldc2004l02. philadelphia: Linguistic data consortium. Linguistic Data Consortium.

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9:1–16.

Ling Chen, Ming Zhao, and 1 others. 2024. Efficient active learning for biomedical image segmentation with minimal annotations. *arXiv preprint arXiv:2405.01701*.

Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2*, AAAI'05, page 746–751. AAAI Press.

Kareem Darwish, Ahmed Abdelali, and Hamdy Mubarak. 2014. Using stem-templates to improve arabic pos and gender/number tagging. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 2926–2931.

Kareem Darwish, Mohammed Attia, Hamdy Mubarak, Younes Samih, Ahmed Abdelali, Lluís Màrquez, Mohamed Eldesouki, and Laura Kallmeyer. 2020. Effective multi-dialectal arabic pos tagging. *Natural Language Engineering*, 26:677–690.

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.

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, and Mohamed Eldesouki. 2017. Arabic pos tagging: Don't abandon feature engineering just yet. *Proceedings of The Third Arabic Natural Language Processing Workshop (WANLP)*, pages 130–137.

Kareem Darwish, Hamdy Mubarak, Mohamed Eldesouki, Ahmed Abdelali, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy, and Laura Kallmeyer. 2018. Multi-dialect arabic pos tagging: A crf approach. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.

David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tem Buckwalter. 2009. Standard arabic morpholgical analyzer(sama) version 3.1. Linguistic Data Consortium.

Onur Güngör, Suzan Uskudarli, and Tunga Güngör. 2018. Improving named entity recognition by jointly learning to disambiguate morphological tags. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2082–2092, Santa

Fe, New Mexico, USA. Association for Computational Linguistics.

Nizar Habash, Mustafa Jarrar, Faeq Alrimawi, Diyam Akra, Nasser Zalmout, Eric Bartolotti, and Mahdi Arar. 2015. Palestinian arabic conventional orthography guidelines. Technical report, Birzeit University.

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.

Nagham Hamad, Mohammed Khalilia, and Mustafa Jarrar. 2025. Konooz: Multi-domain Multi-dialect Corpus for Named Entity Recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 0–0, Vienna, Austria. Association for Computational Linguistics.

Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara, Zeina Aldallal, Sara Chrouf, and Safwan AlModhayan. 2025. Lahjawi: Arabic cross-dialect translator. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 12–24, Abu Dhabi, UAE. Association for Computational Linguistics.

Tymaa Hammouda, Mustafa Jarrar, and Mohammed Khalilia. 2024. SinaTools: Open Source Toolkit for Arabic Natural Language Understanding. In *Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024)*, Procedia Computer Science, Dubai. ELSEVIER.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

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, Muhammad Abdul-Mageed, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Nagham Hamad, and Alaa' Omar. 2023a. WojoodNER 2023: The First Arabic Named Entity Recognition Shared Task. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 748–758. ACL.

Mustafa Jarrar, Diyam Akra, and Tymaa Hammouda. 2024a. ALMA: Fast Lemmatizer and POS Tagger for Arabic. In *Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024)*, Procedia Computer Science, Dubai. ELSEVIER.

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, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian arabic: a preliminary study. In *Proceedings of the EMNLP 2014, Workshop on Arabic Natural Language*, pages 18–27. Association For Computational Linguistics.

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):2-s2.0-85001544989.

Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024b. WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Mustafa Jarrar and Tymaa Hasanain Hammouda. 2024. Qabas: An Open-Source Arabic Lexicographic Database. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13363–13370, Torino, Italy. ELRA and ICCL.

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, Sanad Malaysha, Tymaa Hammouda, and Mohammed Khalilia. 2023b. SALMA: Arabic Sense-annotated Corpus and WSD Benchmarks. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 359–369. ACL.

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. 2023c. 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)*, pages 1–7. IEEE.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of emirati arabic. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 3839–3846.

Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, and Imed Zitouni. 2024. ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task. In *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*, Bangkok, Thailand. Association for Computational Linguistics.

Andreas Kirsch. 2024. Advancing deep active learning & data subset selection: Unifying principles with information-theory intuitions. *arXiv preprint arXiv:2401.04305*.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy. Association for Computational Linguistics.

Daniel Kondratyuk, Tomáš Gavenčiak, Milan Straka, and Jan Hajič. 2018. Lemmatag: Jointly tagging and lemmatizing for morphologically rich languages with brnns. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

V Krishnan and BS Satish. 2023. Document image classification using active learning with ensemble and transfer learning. *International Journal on Document Analysis and Recognition (IJDAR)*, 26:239–252.

Haneen Liqreina, Mustafa Jarrar, Mohammed Khalilia, Ahmed Oumar El-Shangiti, and Muhammad Abdul-Mageed. 2023. Arabic Fine-Grained Entity Recognition. In *Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the EMNLP 2023*, pages 310–323. ACL.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR conference on Arabic language resources and tools*, volume 27, pages 466–467.

Mohamed Maamouri, Ann Bies, Seth Kulick, Sondos Krouna, Fatma Gaddeche, and Wajdi Zaghouani. 2010. Arabic treebank: Part 3 v 3.2 ldc2010t08. Linguistic Data Consortium.

Mohamed Maamouri, Ann Bies, Seth Kulick, Sondos Krouna, Dalila Tabassi, and Michael Ciul. 2018. Bolt egyptian arabic treebank - discussion forum ldc2018t23. Linguistic Data Consortium.

Sindhya K. Nambiar, S. Peter David, and Sumam Mary Idicula. 2023. Abstractive summarization of text document in malayalam language: Enhancing attention model using pos tagging feature. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22:1–14.

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*, pages 12–23. ACL.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1094–1101. European Language Resources Association (ELRA).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 1912. Pytorch: An imperative style, high-performance deep learning library. arxiv 2019. *arXiv preprint arXiv:1912.01703*, 10.

Nicholas Roy and Andrew McCallum. 2001. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 441–448, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Karin C. Ryding. 2014. *Arabic: A Linguistic Introduction*. Cambridge University Press.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *Advances in Intelligent Data Analysis*, pages 309–318, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*.

Burr Settles. 2009. Active learning literature survey. *University of Wisconsin-Madison Department of Computer Sciences*.

Robert Vacareanu, Enrique Noriega-Atala, Gus Hahn-Powell, Marco A. Valenzuela-Escarcega, and Mihai Surdeanu. 2024. Active learning design choices for NER with transformers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 321–334, Torino, Italia. ELRA and ICCL.

Rob van der Goot. 2021. We need to talk about train-dev-test splits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4485–4494, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Bilge Kağan Yazar and Erdal Kiliç. 2025. Improving low-resource kazakh-english and turkish-english neural machine translation using transfer learning and part of speech tags. *IEEE Access*, 13:32341–32356.

A Appendix

ATB, ARZTB	Curras, Nabra, Baladi, SALMA	Gumar	Unified Tag (Sina)
NOUN	NOUN	NOUN	**NOUN**
NOUN_PROP	NOUN_PROP	NOUN_PROP	**NOUN_PROP**
NOUN_NUM	NOUN_NUM	NOUN_NUM	**NOUN_NUM**
NOUN_QUANT	NOUN_QUANT	NOUN_QUANT	**NOUN_QUANT**
-	NOUN_VOICE	-	**NOUN_VOICE**
ADJ	ADJ	ADJ	**ADJ**
ADJ_COMP	ADJ_COMP	ADJ_COMP	**ADJ_COMP**
ADJ_NUM	ADJ_NUM	ADJ_NUM	**ADJ_NUM**
PV	PV	VERB:P	**PV**
IV	IV	VERB:I	**IV**
CV	CV	VERB:C	**CV**
PV_PASS	PV_PASS	-	**PV_PASS**
IV_PASS	IV_PASS	VERB:PI	**IV_PASS**
PRON	PRON	PRON	**PRON**
DEM_PRON	DEM_PRON	PRON_DEM	**DEM_PRON**
INTERROG_PRON	INTERROG_PRON	PRON_INTERROG	**INTERROG_PRON**
EXCLAM_PRON	EXCLAM_PRON	PRON_EXCLAM	**EXCLAM_PRON**
REL_PRON	REL_PRON	PRON_REL	**REL_PRON**
ADV	ADV	ADV	**ADV**
REL_ADV	REL_ADV	ADV_REL	**REL_ADV**
INTERROG_ADV	INTERROG_ADV	ADV_INTERROG	**INTERROG_ADV**
PART	PART	PART	**PART**
EMPHATIC_PART	EMPHATIC_PART	PART_EMPHATIC	**EMPHATIC_PART**
INTERROG_PART	INTERROG_PART	PART_INTERROG	**INTERROG_PART**
RESTRIC_PART	RESTRIC_PART	PART_RESTRIC	**RESTRIC_PART**
FOCUS_PART	FOCUS_PART	PART_FOCUS	**FOCUS_PART**
VOC_PART	VOC_PART	PART_VOC	**VOC_PART**
DET	DET	PART_DET	**DET**
FUT_PART	FUT_PART	PART_FUT	**FUT_PART**
SUB_CONJ	SUB_CONJ	CONJ_SUB	**SUB_CONJ**
PROG_PART	PROG_PART	PART_PROG	**PROG_PART**
NEG_PART	NEG_PART	PART_NEG	**NEG_PART**
VERB_PART	VERB_PART	PART_VERB	**VERB_PART**
PSEUDO_VERB	PSEUDO_VERB	VERB_PSEUDO	**PSEUDO_VERB**
VERB	VERB	VERB_NOM	**VERB**
PREP	PREP	PREP	**PREP**
CONJ	CONJ	CONJ	**CONJ**
INTERJ	INTERJ	INTERJ	**INTERJ**
NOUN_NUM	DIGIT	DIGIT	**DIGIT**
FOREIGN	FOREIGN	FOREIGN	**FOREIGN**

Table 3: POS Tagsets Mapping