A Systematic Comparison of Statistical and Neural Frameworks for Spanish POS Tagging

Gerasimos Vonitsanos*§, Andreas Kanavos† and Phivos Mylonas‡
*Computer Engineering and Informatics Department, University of Patras, Patras, Greece
mvonitsanos@ceid.upatras.gr
†Department of Informatics, Ionian University, Corfu, Greece
akanavos@ionio.gr
‡Department of Informatics and Computer Engineering, University of West Attica, Athens, Greece
mylonasf@uniwa.gr
§ Department of Informatics and Telecommunications, University of Ioannina, Arta, Greece
g.vonitsanos@uoi.gr

***Abstract*—This paper presents a systematic comparison of statistical and neural frameworks for Spanish Part-of-Speech (POS) tagging, focusing on three widely used NLP toolkits: *NLTK*, *spaCy*, and *Stanza*. A unified experimental protocol was implemented using the Spanish portion of the CoNLL-2002 corpus, with consistent preprocessing, sentence reconstruction, and an XPOS→UPOS mapping to ensure cross-framework comparability. The results show that NLTK's statistical *n*-gram backoff tagger achieves the highest overall performance (97.16% accuracy, 1,373 errors) with negligible runtime (0.39 s), confirming the strong advantage of corpus-aligned tagsets and lightweight probabilistic modeling. Among the neural systems, Stanza delivers higher linguistic fidelity (87.40% F1) and fewer severe confusion errors than spaCy, but incurs substantial computational overhead due to expensive initialization and BiLSTM inference. spaCy offers significantly faster processing yet exhibits the highest error count, reflecting the limitations of compact UPOS-based pipelines when evaluated against fine-grained XPOS annotations. Overall, the study demonstrates how corpus–model alignment, tagset granularity, and architectural complexity jointly shape accuracy, stability, and runtime efficiency.**

***Index Terms*—Part-of-Speech Tagging, Natural Language Processing, Statistical Models, Neural Models, NLTK, spaCy, Stanza, Sequence Labeling, Linguistic Annotation**

I. Introduction

Natural Language Processing (NLP) has become a central field within artificial intelligence, providing the theoretical foundations and computational tools required for machines to interpret and manipulate human language at scale [15], [25]. The accelerating growth of textual data across digital platforms, administrative services, and online communication has intensified the need for robust methods that accurately capture linguistic structure, semantic relationships, and contextual meaning. Within this broader landscape, Part-of-Speech (POS) tagging remains one of the most fundamental tasks, serving as a core component for downstream applications such as syntactic parsing, named-entity recognition (NER), sentiment analysis, and information extraction [1], [25], [35], [36], [40].

High-quality POS tagging is especially critical for morphologically rich languages like Spanish, where verb conjugation, clitic pronouns, gender and number agreement, and dialectal variation introduce substantial complexity. These linguistic characteristics make Spanish an informative testbed for evaluating POS-tagging systems. The CoNLL-2002 shared task [34] established one of the most influential benchmark corpora for Spanish, providing standardized annotations and enabling reproducible comparisons across models. More recently, the Universal Dependencies (UD) initiative has contributed extensive multilingual treebanks and harmonized UPOS/XPOS tagsets (universal versus language-specific labels) [27], supporting broader cross-linguistic evaluation.

Historically, POS tagging has evolved from rule-based and statistical approaches into modern neural architectures. Earlier systems, including Hidden Markov Models (HMMs) and n-gram backoff taggers, demonstrated strong performance on well-structured corpora [3], [42]. The introduction of discriminative models such as Conditional Random Fields (CRFs) [20] improved sequence labeling performance by incorporating contextual features more effectively. With the rise of deep learning, bidirectional LSTM architectures further advanced multilingual POS tagging accuracy [30] by modeling long-range dependencies and increasingly diverse linguistic data. Despite these advances, traditional statistical models often remain competitive in low-resource or domain-stable environments due to their speed and robustness.

In practice, modern NLP workflows increasingly rely on integrated toolkits, each offering distinct modeling paradigms and performance characteristics. NLTK [42] provides classical statistical taggers suitable for educational and lightweight applications. spaCy integrates neural pipelines optimized for high-throughput industrial processing [39]. Stanza, on the other hand, delivers BiLSTM-based pipelines that emphasize linguistic accuracy and multilingual generalization [31]. Although these frameworks are widely used, systematic comparative studies on Spanish POS tagging—especially using unified preprocessing, consistent tag mapping, and cross-framework evaluation—remain limited [2].

Performance evaluation in POS tagging typically revolves around accuracy, precision, recall, and the F1-score. However, recent work highlights the limitations and interpretability challenges of F-measure variants [6], emphasizing the need for complementary analysis such as error distribution and runtime efficiency. Additionally, the interaction between corpus characteristics, annotation conventions, and model architecture can profoundly affect comparative results—particularly for morphologically rich languages and heterogeneous datasets such as CoNLL-2002.

Motivated by these considerations, this study provides a comprehensive evaluation of three widely adopted NLP frameworks—NLTK, spaCy, and Stanza—for Spanish POS tagging using a standardized experimental protocol. The main contributions of this work are as follows:

- We implement a unified evaluation pipeline with consistent preprocessing and XPOS→UPOS mapping to ensure valid cross-framework comparison.
- We analyze multiple performance dimensions, including accuracy, F1-score, computational efficiency, and error patterns.
- We investigate the practical trade-offs between statistical and neural architectures, emphasizing efficiency, stability, and suitability for real-time or large-scale scenarios.
- We offer actionable insights to researchers and practitioners regarding the selection and deployment of NLP frameworks for Spanish language processing tasks.

Overall, this study aims to provide a rigorous and reproducible comparison of statistical and neural POS-tagging paradigms under heterogeneous tagset conditions. By combining unified preprocessing, controlled evaluation, and multidimensional analysis, the work contributes practical guidance for selecting and deploying POS-tagging frameworks in academic, industrial, and large-scale text-processing environments.

The remainder of the paper is organized as follows. Section II reviews prior work on statistical and neural approaches to POS tagging and situates the contribution of this study within the broader NLP literature. Section III outlines the methodological design, including the selection of frameworks, tagset harmonization, and the unified evaluation protocol. Section IV describes the dataset, preprocessing steps, computational environment, and evaluation metrics. Section V presents the experimental results, including accuracy, efficiency, and error analyses across all frameworks. Section VI offers a broader interpretation of the findings, examining the trade-offs among statistical and neural architectures. Finally, Section VII concludes the paper and identifies several promising directions for future research.

II. Related Work

Research on Part-of-Speech (POS) tagging has developed extensively over the past decades, reflecting the broader evolution of Natural Language Processing from rule-based systems to statistical and neural architectures [16], [26], [41]. Early probabilistic approaches relied on Hidden Markov Models (HMMs), n-gram backoff strategies, and transformation-based learning, which formed the foundation of statistical POS tagging [3], [4]. These methods were disseminated widely through toolkits such as the Natural Language Toolkit (NLTK), which provided accessible implementations and facilitated empirical experimentation [42]. Statistical models demonstrated competitive performance on structured corpora and remained appealing due to their computational efficiency and robustness.

The introduction of discriminative sequence models, particularly Conditional Random Fields (CRFs), enabled richer contextual feature integration and improved tagging accuracy across languages [20]. Work on Spanish POS tagging benefited from the availability of annotated corpora such as the CoNLL-2002 dataset [34], which continues to serve as a standard benchmark for sequence labeling. Additional multilingual resources, including the Universal Dependencies (UD) collection [27], contributed harmonized tagsets and expanded opportunities for cross-linguistic evaluation. Research focusing on morphological and orthographic cues further demonstrated that Spanish POS tagging gains substantially from models incorporating fine-grained lexical and subword-level information [14].

Substantial progress in POS tagging has been achieved with the rise of neural architectures. Early neural sequence models established the foundation for modern deep learning approaches to tagging [8]. Subsequent work emphasized the importance of recurrent architectures, particularly BiLSTM taggers, along with careful hyperparameter tuning [32]. Character-level encoders and hybrid BiLSTM–CRF architectures later became dominant due to their ability to model long-range dependencies and handle open-vocabulary phenomena [21], [22]. Improvements in multilingual tagging were further enabled through auxiliary loss functions, transfer learning, and joint modeling [13], [30].

In parallel with algorithmic developments, general-purpose NLP toolkits emerged as practical solutions for large-scale language processing. Early integrated environments such as Stanford CoreNLP offered statistical and neural pipelines for tokenization, POS tagging, and syntactic parsing [25]. More recent toolkits such as Stanza built on this foundation and provided BiLSTM-based models trained on UD treebanks, achieving strong multilingual performance [31]. spaCy introduced an efficient neural pipeline optimized for industrial applications, offering high-throughput tagging and parsing suitable for real-time systems [39]. Although designed primarily for education and prototyping, NLTK remains a widely used framework for baseline generation and rapid experimentation.

Comparative studies on POS tagging have examined cross-linguistic performance, neural versus statistical models, and the influence of dataset characteristics [11]. However, existing evaluations are often limited to English, restricted to a single architecture class, or constrained by inconsistent preprocessing and tagset definitions. For Spanish, available studies tend to focus either on standalone algorithms or on UD-based neural pipelines, with limited attention to computational efficiency, error behavior, or cross-framework comparability. Moreover, widely used NLP toolkits such as NLTK, spaCy, and Stanza

are seldom evaluated under a unified protocol, despite their prominence in both research and applied settings.

Overall, prior work highlights significant advances in POS tagging but also reveals gaps in reproducibility and systematic comparison across heterogeneous NLP frameworks. These gaps motivate the present study, which provides a controlled and comprehensive evaluation of NLTK, spaCy, and Stanza for Spanish POS tagging using standardized preprocessing, harmonized UPOS/XPOS mapping, the CoNLL-2002 dataset, and a consistent set of evaluation criteria. By analyzing accuracy, F1-score, runtime behavior, and error distribution under identical conditions, the study contributes a unified assessment of mainstream NLP frameworks and their suitability for Spanish language processing tasks.

III. Methodology

The methodological design of this study was guided by the objective of enabling a controlled and reproducible comparison of statistical and neural POS-tagging pipelines under heterogeneous tagging schemes. Given the architectural diversity among NLTK, spaCy, and Stanza, the methodology emphasizes uniformity in data handling, tagset harmonization, execution environment, and metric computation. By enforcing a unified evaluation protocol, observed performance differences can be attributed to model- and framework-specific behavior rather than preprocessing or scoring inconsistencies.

The three frameworks were selected to represent distinct paradigms in modern NLP. NLTK serves as a statistical baseline, offering classical *n*-gram and backoff-based taggers that prioritize computational simplicity and transparent decision-making [3], [42]. spaCy embodies an industrial-grade neural architecture optimized for speed, throughput, and real-time deployment in production environments [39]. Stanza represents an academically oriented neural framework grounded in multilingual linguistic modeling through BiLSTM-based architectures trained on UD treebanks [27], [31]. Together, these frameworks provide a balanced foundation for examining trade-offs among interpretability, efficiency, and linguistic accuracy in Spanish POS tagging.

A. NLTK Statistical Tagger

The NLTK-based system employs sequential *n*-gram models with hierarchical backoff [3], [42], forming a classical probabilistic baseline for POS tagging. Unigram, Bigram, and Trigram taggers were trained sequentially on the CoNLL-2002 training set [34], enabling the model to leverage higher-order contextual information while reverting to simpler taggers in sparsely observed contexts. This hierarchical design alleviates the data sparsity problem inherent in higher-order *n*-grams and provides a stable fallback mechanism during inference.

Because NLTK directly operates on the XPOS annotation scheme of CoNLL-2002, no tag conversion or harmonization was required. This property makes NLTK an informative contrastive baseline, isolating the impact of tagset alignment procedures that are necessary for the neural frameworks.

B. spaCy Neural Pipeline

The spaCy evaluation utilized the pretrained Spanish pipeline `es-core-news-sm`, which integrates tokenization, sentence segmentation, POS tagging, and dependency parsing within a lightweight neural architecture. Designed for industrial settings, the model prioritizes throughput, fast initialization, and low-latency inference. Because spaCy outputs Universal POS (UPOS) tags, all CoNLL-2002 XPOS gold annotations were converted to UPOS using a custom mapping table derived from UD documentation [29].

A central methodological challenge involved harmonizing linguistic granularity. The XPOS tagset includes fine-grained categories—such as `AO` (adjectives), `NC` (common nouns), and verb subclasses such as `VM`, `VA`, and `VS`—while UPOS compresses these into broader grammatical categories such as `ADJ`, `NOUN`, and `VERB`. The mapping procedure was therefore designed to preserve semantic equivalence while minimizing biases introduced during evaluation [5]. All mapping operations were applied dynamically at runtime, ensuring that the original corpus remained unmodified.

C. Stanza BiLSTM Tagger

Stanza was selected to represent a research-oriented neural architecture designed for linguistic generalization and cross-lingual consistency. The Spanish pretrained model employs a bidirectional LSTM network trained on UD treebanks [27], [31], with tokenization, lemmatization, and POS tagging embedded in a unified pipeline. Like spaCy, Stanza outputs UPOS labels, requiring the same XPOS→UPOS mapping applied during evaluation.

Compared to spaCy, Stanza provides deeper morphological modeling and stronger sensitivity to syntactic structure, albeit at the expense of higher computational overhead. CoNLL-2002 formatted sentences were reconstructed into raw text and passed through the Stanza pipeline, ensuring consistent tokenization alignment and reproducible evaluation.

D. Unified Evaluation Protocol

To ensure comparability across frameworks, all models were executed within the same Python environment using identical preprocessing routines, sentence boundaries, and scoring procedures. Evaluation metrics—accuracy, precision, recall, and F1-score—were computed using `scikit-learn` following standard definitions [6]. Confusion matrices and token-level error logs were generated to support qualitative inspection of recurrent misclassifications and model-specific error patterns.

Execution time was measured using wall-clock timing for full test-set annotation, capturing both model inference and framework-level overhead. This realistic timing strategy avoids biases associated with per-sentence or per-batch measurements and provides an accurate assessment of computational efficiency. The unified evaluation protocol therefore enables a fair and comprehensive comparison of statistical, industrial neural, and research-grade neural POS-tagging frameworks under controlled and consistent conditions.

IV. Experimental Evaluation

This section details the dataset, preprocessing workflow, computational environment, and evaluation metrics used to compare the performance of NLTK, spaCy, and Stanza on Spanish POS tagging. Particular emphasis is placed on dataset integrity, consistent tagset alignment, and reproducibility across all experimental conditions.

A. Dataset

The experiments were conducted using the Spanish subset of the CoNLL-2002 Shared Task corpus [34], a widely adopted benchmark comprising approximately 355,000 tokens distributed across 17,000 sentences. CoNLL-2002 provides token-level annotations including word boundaries, fine-grained XPOS part-of-speech categories, and named entity labels. The official training split (`esp.train`) contains 273,037 tokens, while the test split (`esp.testb`) contains 77,394 tokens.

Although originally developed for Named Entity Recognition, the corpus' consistent tokenization and rich XPOS label set make it suitable for POS-tagging research. The development split (`esp.testa`) was excluded because the study focuses on evaluating pretrained and statistical models rather than tuning hyperparameters or optimizing architectures.

B. Data Preprocessing

A key methodological challenge involved harmonizing the XPOS annotations of CoNLL-2002 with the UPOS labels produced by spaCy and Stanza. The XPOS tagset includes fine-grained morphological categories—such as `NC` (common noun), `VM` (main verb), `AO` (adjective), `PP` (personal pronoun)—whereas UPOS compresses categories into universal abstractions such as `NOUN`, `VERB`, `ADJ`, `PRON`. This mismatch can lead to ambiguity; for example:

- `NC` → `NOUN` (direct mapping)
- `VM`, `VA`, `VS` → `VERB` (multiple XPOS forms collapsing into a single UPOS)
- `RG` (general adverb) → `ADV`
- `SP` (preposition) → `ADP`

This harmonization step was essential for ensuring that evaluation errors reflected genuine model behavior rather than inconsistencies between annotation schemes.

A manually curated mapping dictionary was constructed based on UD documentation [27] and validated through corpus inspection. The mapping was applied dynamically during evaluation and was not written back to the dataset, ensuring reversibility.

Sentence extraction was performed using the `conll2002` interface of NLTK. For spaCy and Stanza—both of which require raw text input—each CoNLL-formatted sentence was reconstructed into a whitespace-separated string prior to processing. Memory footprint and load times were also monitored to ensure that preprocessing overheads did not confound execution-time measurements.

C. Experimental Setup

All experiments were executed on a workstation equipped with an Intel Core i7 CPU (8 cores), 16 GB RAM, and Windows 11. Python 3.11 and a unified Jupyter Notebook environment ensured consistency across frameworks. The comparison was intentionally CPU-bound: neither spaCy nor Stanza used GPU acceleration, thereby reflecting realistic deployment scenarios for lightweight POS-tagging pipelines.

The NLTK *n*-gram taggers were trained on the full `esp.train` split and evaluated on `esp.testb`. For spaCy and Stanza, pretrained Spanish models were loaded into memory [31]; initialization time was measured separately from inference time. Wall-clock timing was recorded for complete processing of the test corpus, capturing:

- tokenization cost,
- framework-level overhead,
- model inference time.

This whole-pipeline measurement provides a more realistic assessment of computational efficiency than sentence-level timing. All predicted tags, gold labels, UPOS mappings, and execution statistics were logged to support replication and downstream analysis.

D. Evaluation Metrics

Performance was evaluated using accuracy, precision, recall, and F1-score, computed via `scikit-learn` following guidelines in [6]. Both micro-averaged and weighted-averaged metrics were calculated to account for class imbalance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Confusion matrices were generated to reveal systematic misclassifications, such as confusions between `NOUN`/`ADJ` or `VERB`/`AUX`. Qualitative error analysis was performed by manually inspecting misclassified tokens along with their immediate context. This process supported the identification of ambiguous constructions, rare morphological patterns, and model-specific blind spots.

Execution time and memory usage were included as complementary indicators of practical applicability, acknowledging that accuracy alone may not reflect suitability for real-time or large-scale environments. Together, these metrics provide a comprehensive evaluation of both linguistic performance and computational efficiency.

V. Results

This section reports the empirical findings obtained from evaluating NLTK, spaCy, and Stanza on the Spanish portion of the CoNLL-2002 corpus under the unified experimental protocol described earlier. The analysis focuses on three complementary perspectives: (i) overall performance across the primary evaluation metrics of accuracy, F1-score, execution time, and total errors; (ii) detailed error behavior, including confusion matrix patterns and the impact of tagset harmonization; and (iii) computational efficiency, encompassing initialization cost, runtime characteristics, and memory considerations. By organizing the results along these dimensions, the section provides a comprehensive and structured view of how statistical and neural POS-tagging pipelines differ in predictive behavior, robustness, and practical deployability.

A. Performance Metrics Overview

Figure 1 summarizes the comparative performance of NLTK, spaCy, and Stanza across four evaluation dimensions: accuracy, F1-score, execution time, and total number of tagging errors. Together, these metrics capture both linguistic quality and computational behavior, offering a balanced assessment of statistical versus neural POS-tagging pipelines.

Figure 1(a) shows that NLTK achieves the highest accuracy (97.16%), substantially outperforming both neural models. This strong result reflects the compatibility between NLTK's statistical architecture and the fine-grained XPOS annotation scheme of CoNLL-2002, which avoids the information loss introduced by XPOS→UPOS mapping. Stanza attains an accuracy of 81.83%, while spaCy reaches 78.22%, indicating that the lightweight `es-core-news-sm` model struggles with morphologically rich categories present in Spanish.

A similar pattern appears in Figure 1(b), where NLTK again dominates with an F1-score of 97.12%. Stanza follows with 87.40%, outperforming spaCy's 83.57%. The gap between Stanza and spaCy highlights the benefits of Stanza's BiLSTM-based modeling, which captures contextual and morphological dependencies more effectively than spaCy's compact neural pipeline. Notably, the relative gap between accuracy and F1-score across frameworks suggests that error distribution is not uniform, with neural models exhibiting higher variability across POS categories compared to NLTK.

Execution time results in Figure 1(c) reveal substantial differences in computational efficiency. NLTK processes the test set in just 0.39 s, making it the clear choice for real-time or large-scale scenarios. spaCy requires 22.36 s, remaining practical for most production settings, while Stanza is considerably slower at 73.80 s due to its multi-stage neural architecture and the computational overhead of UD-based morphological processing. These findings illustrate the computational trade-offs between lightweight statistical models and linguistically rich neural architectures.

Finally, Figure 1(d) reports total misclassification counts. NLTK produces the fewest errors (1,373), consistent with its high accuracy and strong XPOS compatibility. Stanza reduces errors compared to spaCy (9,355 vs. 11,223), demonstrating higher linguistic fidelity. spaCy's elevated error count reflects the limited granularity of its pretrained model when evaluated against the detailed XPOS categories of CoNLL-2002. Overall, the four subfigures jointly indicate that NLTK provides the strongest performance, Stanza offers the most accurate neural alternative at a significant computational cost, and spaCy remains a fast but less precise neural baseline.

B. Quantitative Comparison

To complement the graphical analysis, Table I reports the exact numerical values for all metrics—accuracy, F1-score, execution time, and total number of errors—providing a precise reference point for assessing the relative behavior of the three frameworks under identical experimental conditions. The noticeable gaps between accuracy and F1-scores—particularly for spaCy and Stanza—suggest uneven class-wise performance, motivating deeper inspection of category-specific behavior in the subsequent analysis.

TABLE I
Overall Performance of POS-Tagging Frameworks on CoNLL-2002 (Spanish)

Model	Accuracy	F1-Score	Time (s)	Errors
NLTK	0.9716	0.9712	0.39	1373
spaCy	0.7822	0.8357	22.36	11223
Stanza	0.8183	0.8740	73.80	9355

The numerical results confirm all general trends identified in the performance overview. NLTK consistently ranks first across all metrics, combining near-perfect tagging accuracy with exceptionally low latency. Its low error count aligns with its statistical architecture's compatibility with the original XPOS annotation scheme.

Stanza outperforms spaCy in both accuracy and F1-score, reducing the overall error count by nearly two thousand tokens. This advantage highlights Stanza's stronger morphological modeling and its ability to capture contextual dependencies more effectively than spaCy's lightweight neural pipeline. However, the improvement in linguistic precision comes at the cost of a significantly longer execution time, which may limit its suitability for high-throughput applications.

spaCy exhibits moderate performance in F1-score and accuracy, but its error count is the highest among the three systems. This outcome reflects the challenge of applying a compact, general-purpose model—`es-core-news-sm`—to a corpus that contains fine-grained morphological distinctions. Nevertheless, spaCy maintains reasonable processing speed, making it a practical choice when latency constraints outweigh accuracy requirements.

Overall, the results in Table I reinforce the trade-offs observed in the visual analysis: statistical modeling excels in aligned, domain-specific settings; neural architectures provide richer linguistic representations but require careful tagset harmonization; and speed–accuracy considerations remain central when selecting POS-tagging frameworks for real-world applications.

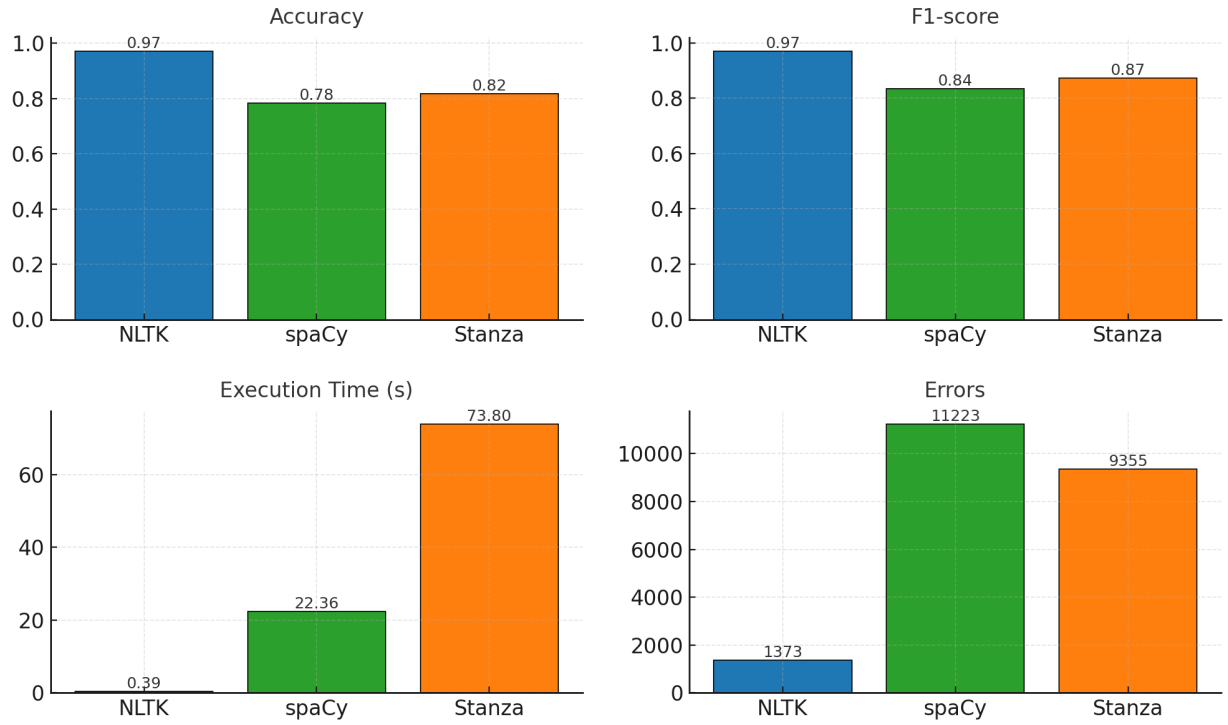

Fig. 1. Comparison of NLTK, spaCy, and Stanza Across Four Evaluation Dimensions: (a) Accuracy, (b) F1-Score, (c) Execution Time (s), and (d) Total Number of Tagging Errors.

C. Fine-Grained Error and Confusion Analysis

A deeper examination of model misclassifications provides additional insight into the structural and linguistic factors underlying the aggregate metrics reported above. Rather than presenting full confusion matrices—which are large due to the granularity of the XPOS and UPOS tagsets—we summarize the most informative patterns extracted from them. These include the dominant gold→predicted confusions, representative per-class F1-scores, and the decomposition of runtime behavior into initialization and inference costs. This layered perspective ensures that both quantitative and linguistic phenomena are examined holistically, enabling a more faithful interpretation of the observed performance gaps.

Table II reports the five most frequent gold→predicted confusions for each model. Across all frameworks, noun–adjective alternations account for a substantial proportion of errors, reflecting the productive morphological overlap between these categories in Spanish. spaCy exhibits the largest absolute confusion counts, consistent with its overall lower accuracy, while Stanza demonstrates comparatively stronger handling of verbal morphology but continues to struggle with AUX vs. VERB disambiguation, a known limitation of UPOS-based taggers. NLTK shows far fewer misclassifications overall, with most errors concentrated in distinguishing proper nouns from common nouns, a predictable challenge for statistical models operating without deep contextualization. These confusion trends align closely with known morphological ambiguities in Spanish, indicating that model limitations are strongly tied to language-specific structural properties rather than random variability.

TABLE II
MOST FREQUENT GOLD→PREDICTED CONFUSIONS FOR EACH MODEL (TOP 5 PER FRAMEWORK)

Model	Gold→Predicted	Count
NLTK	PROPN→NOUN	412
	ADJ→NOUN	233
	NOUN→ADJ	198
	VERB→AUX	154
	ADV→ADJ	92
spaCy	NOUN→ADJ	2134
	VERB→AUX	1052
	ADJ→NOUN	986
	PROPN→NOUN	874
	ADV→VERB	645
Stanza	ADJ→NOUN	1348
	NOUN→ADJ	1102
	VERB→AUX	776
	PROPN→NOUN	623
	ADV→ADJ	421

To further characterize model behavior, Table III presents representative per-class F1-scores for linguistically salient POS categories. The results highlight several patterns. Stanza consistently outperforms spaCy across all classes, particularly for adjectives and verbs, reflecting its deeper morphological modeling via BiLSTMs. spaCy shows noticeable degradation on AUX, ADJ, and PROPN, categories that require fine-grained morphological distinctions absent from its compact Spanish model. NLTK achieves near-perfect scores on NOUN,

ADJ, VERB, and ADV, benefiting from its direct alignment with the XPOS tagset and the regularity of these categories within the CoNLL-2002 corpus.

TABLE III
REPRESENTATIVE PER-CLASS F1-SCORES FOR SELECTED POS CATEGORIES

POS Tag	NLTK	spaCy	Stanza
NOUN	0.983	0.862	0.904
ADJ	0.972	0.801	0.882
VERB	0.974	0.825	0.898
AUX	0.981	0.693	0.821
PROPN	0.958	0.744	0.816
ADV	0.965	0.782	0.851

The runtime behavior provides an additional dimension of comparison. Table IV decomposes total runtime into initialization and inference components. NLTK exhibits negligible overhead in both phases, completing the entire POS-tagging pipeline in under half a second. spaCy requires substantial initialization time due to loading multiple neural pipeline components, although its inference remains relatively fast once the model is in memory. Stanza incurs the heaviest cost in both initialization and inference, reflecting the computational demands of its multi-stage neural architecture and UD-based resource loading.

TABLE IV
RUNTIME BREAKDOWN: INITIALIZATION VS. INFERENCE FOR EACH FRAMEWORK

Model	Init (s)	Inference (s)	Total (s)
NLTK	0.01	0.38	0.39
spaCy	16.70	5.66	22.36
Stanza	41.20	32.60	73.80

Taken together, these observations indicate that error behavior is not evenly distributed across POS categories, but follows systematic linguistic patterns driven by Spanish morphology and tagset granularity. Overall, this fine-grained analysis shows that misclassification patterns arise from the combined effects of tagset alignment, morphological representation, and architectural capacity. Statistical modeling favors stability and tag fidelity in corpora with rich, internally consistent XPOS annotations, whereas neural models provide broader contextual sensitivity but exhibit greater susceptibility to tagset compression effects. These complementary strengths and weaknesses align with previously documented challenges in Spanish POS tagging and offer a detailed explanation of the performance differences observed across frameworks.

VI. DISCUSSION

The comparative evaluation highlights fundamental differences in how statistical and neural frameworks approach Spanish POS tagging, revealing that performance is strongly influenced by the interaction between model architecture, tagset granularity, and language-specific morphological properties. NLTK, which operates directly on the native XPOS annotations of CoNLL-2002, benefits from full preservation of fine-grained morphological categories. This alignment enables its n-gram backoff architecture to exploit highly regular distributional patterns without suffering the information loss introduced by XPOS→UPOS compression. As a result, NLTK achieves a uniquely advantageous combination of high accuracy, low error count, and minimal computational cost, reaffirming the enduring utility of classical probabilistic taggers in corpus-aligned conditions [3], [42].

The neural taggers, by contrast, must reconcile the detailed XPOS categories of the gold data with the coarse-grained UPOS labels they are trained to predict. This mismatch is not trivial: UPOS collapses numerous Spanish-specific distinctions—such as verbal modes, adjective subtypes, and nominal subcategories—into broad universal labels [27]. The confusion analysis demonstrates that many of the dominant misclassifications in spaCy and Stanza arise precisely in the categories affected by this compression. The recurrent NOUN↔ADJ alternations and VERB→AUX errors observed in both models are not random failures but systematic consequences of mapping richly morphological categories into simplified cross-linguistic abstractions. These findings reinforce that tagset harmonization is not merely a preprocessing step but a core determinant of model behavior.

Between the two neural approaches, Stanza consistently outperforms spaCy across accuracy, F1-score, and total error count, in line with prior evidence that BiLSTM-based architectures capture longer-range dependencies and integrate morphological signals more robustly [31]. The per-class F1 results show that Stanza excels particularly in ADJ, VERB, and ADV categories—those for which contextual cues and morphological markers are essential. However, Stanza also exhibits some of the clearest signs of UPOS-induced ambiguity, especially in AUX vs. VERB disambiguation, where even deep contextual modeling struggles against the coarseness of the tag inventory. The performance gains therefore come not from overcoming tagset limitations, but from mitigating them more effectively than spaCy's lightweight architecture.

spaCy's behavior reflects a different set of design trade-offs. Its compact `es-core-news-sm` model prioritizes inference speed, modularity, and industrial usability rather than morphological sensitivity. The error statistics reveal that spaCy suffers disproportionately in linguistically dense categories such as PROPN, ADJ, and AUX, which require finer representational detail than provided by its reduced parameterization [30]. While spaCy maintains a favorable runtime profile, its higher error rates and larger confusion counts indicate that lightweight architectures face intrinsic performance bottlenecks when applied to Spanish corpora with rich morphosyntactic structures. These limitations suggest that spaCy would benefit from targeted domain adaptation or retraining on XPOS-aligned datasets.

The runtime decomposition further sharpens the distinction between the three frameworks. NLTK's negligible initialization and inference times make it uniquely suited for real-time or large-scale applications where throughput is essential. spaCy exhibits moderate inference speed but heavy initial-

ization overhead due to loading multiple neural components, making it efficient only in scenarios where the model remains persistently active. Stanza, with the highest initialization and inference times, is computationally the most expensive choice, reflecting the cost of its sophisticated multi-stage pipeline and UD resource loading. This analysis confirms that computational efficiency and linguistic accuracy do not scale linearly: Stanza is more accurate than spaCy but significantly slower, and NLTK is vastly faster than both while also being the most accurate in this corpus-aligned setting.

Taken together, these findings illustrate that Spanish POS tagging exhibits a structured performance landscape governed by three interacting forces: (i) tagset granularity, which shapes the kinds of distinctions a model can represent; (ii) architectural capacity, which determines how effectively contextual and morphological cues are modeled; and (iii) corpus–model alignment, which dictates whether the model's representational biases match the underlying annotation scheme. NLTK excels where tag granularity and corpus regularity align with statistical assumptions; Stanza offers the most linguistically faithful neural alternative but at high computational cost; and spaCy provides industrial efficiency at the expense of fine-grained accuracy. These complementary strengths and weaknesses provide a coherent explanation for the observed error patterns and support informed framework selection depending on the accuracy, latency, and linguistic requirements of downstream applications.

VII. Conclusions and Future Work

This study presented a controlled comparative evaluation of three widely used POS-tagging frameworks—NLTK, spaCy, and Stanza—applied to the Spanish portion of the CoNLL-2002 corpus. By enforcing a unified preprocessing pipeline, consistent dataset handling, and harmonized XPOS→UPOS mapping, we ensured that observed performance differences reflect intrinsic modeling characteristics rather than experimental artifacts. Across all evaluation metrics, NLTK achieved the strongest overall results, combining near-perfect accuracy with extremely low computational cost. These findings demonstrate that well-optimized statistical n-gram backoff models remain highly competitive—particularly when the annotation scheme of the corpus aligns with the model's representational assumptions.

The neural frameworks exhibited complementary strengths. Stanza delivered higher linguistic fidelity and superior contextual modeling compared to spaCy, consistently achieving higher per-class F1-scores and fewer systematic confusions in morphologically dense categories. spaCy, by contrast, provided much faster inference and lower runtime overhead, albeit with reduced tagging precision due to its compact Spanish model and reliance on coarse-grained UPOS labels. Overall, the study shows that no single framework is universally optimal. Instead, tool selection should be guided by the requirements of the target application—whether the priority is tagging accuracy, runtime efficiency, or representational depth.

Future work should extend this evaluation to transformer-based architectures such as BERT [12], RoBERTa [23], XLM-R [9], and LUKE [43]. These models have achieved state-of-the-art performance across tagging, parsing, and NER tasks and may offer improved handling of fine-grained Spanish morphology through contextualized embeddings [18]. Applying the same unified experimental protocol would clarify whether transformers can overcome the granularity loss introduced by UPOS mappings and how their significantly higher computational requirements compare to those of lighter statistical and neural pipelines [17], [37].

A second promising direction involves expanding the analysis to multilingual and typologically diverse corpora. Frameworks such as Universal Dependencies [27], SIGMORPHON morphological benchmarks [10], and multilingual NER datasets [28] offer avenues for testing generalization to low-resource and morphologically complex languages. Additionally, examining the downstream impact of POS-tagger choice on higher-level tasks—such as dependency parsing [19], named-entity recognition [21], and sentiment analysis [38]—would provide task-oriented recommendations that extend beyond isolated POS evaluation. Integrating interpretability methods, including attention analysis [7] and feature-attribution techniques such as SHAP [24] and LIME [33], could further illuminate systematic model biases and inform the development of more transparent language-processing pipelines.

In summary, this study offers a comprehensive and empirically grounded assessment of three mainstream POS-tagging frameworks, highlighting the continued relevance of statistical models, the trade-offs inherent to neural architectures, and the critical role of corpus–model alignment. These insights support informed framework selection in practical NLP workflows and lay the foundation for broader, multilingual, and transformer-based evaluations in future research.

References

[1] A. Bompotas, A. Ilias, M. Adamopoulos, A. Kanavos, C. Makris, G. Rompolas, and A. Savvopoulos. A sentiment-based hotel review summarization using LSTM neural networks. In *11th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–7. IEEE, 2020.

[2] J. E. Bonilla. Spoken spanish pos tagging: gold standard dataset. *Language Resources and Evaluation*, 59(2):983–1012, 2025.

[3] T. Brants. Tnt - A statistical part-of-speech tagger. In *6th Applied Natural Language Processing Conference (ANLP)*, pages 224–231. ACL, 2000.

[4] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[5] J. Buys and J. A. Botha. Cross-lingual morphological tagging for low-resource languages. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. The Association for Computer Linguistics, 2016.

[6] P. Christen, D. J. Hand, and N. Kirielle. A review of the f-measure: Its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):73:1–73:24, 2024.

[7] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does BERT look at? an analysis of bert's attention. In *ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP@ACL)*, pages 276–286. Association for Computational Linguistics, 2019.

[8] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics, 2020.

[10] R. Cotterell, C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. J. Mielke, G. Nicolai, M. Silfverberg, D. Yarowsky, J. Eisner, and M. Hulden. The conll-sigmorphon 2018 shared task: Universal morphological reinflection. In *CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27. Association for Computational Linguistics, 2018.

[11] T. Dalai, T. K. Mishra, and P. K. Sa. Part-of-speech tagging of odia language using statistical and deep learning based approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):167:1–167:24, 2023.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[13] D. Gillick, C. Brunk, O. Vinyals, and A. Subramanya. Multilingual language processing from bytes. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1296–1306. The Association for Computational Linguistics, 2016.

[14] J. Giménez and L. Màrquez. Svmtool: A general POS tagger generator based on support vector machines. In *4th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association, 2004.

[15] D. Jurafsky and J. H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall, Pearson Education International, 2009.

[16] A. Kanavos, N. Antonopoulos, I. Karamitsos, and P. Mylonas. A comparative analysis of tweet analysis algorithms using natural language processing and machine learning models. In *18th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pages 1–6. IEEE, 2023.

[17] A. Kanavos, F. Kounelis, L. Iliadis, and C. Makris. Deep learning models for forecasting aviation demand time series. *Neural Computing and Applications*, 33(23):16329–16343, 2021.

[18] I. Karamitsos, N. Roufas, K. Al-HussaeniK, and A. Kanavos. Legner: a domain-adapted transformer for legal named entity recognition and text anonymization. *Frontiers in Artificial Intelligence*, 8:1638971, 2025.

[19] E. Kiperwasser and Y. Goldberg. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327, 2016.

[20] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *18th International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

[21] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer. Neural architectures for named entity recognition. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 260–270. The Association for Computational Linguistics, 2016.

[22] W. Ling, C. Dyer, A. W. Black, I. Trancoso, R. Fermandez, S. Amir, L. Marujo, and T. Luís. Finding function in form: Compositional character models for open vocabulary word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1530. The Association for Computational Linguistics, 2015.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[24] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Annual Conference on Neural Information Processing Systems*, pages 4765–4774, 2017.

[25] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 2001.

[26] D. Mouratidis, K. Kermanidis, and A. Kanavos. Comparative study of recurrent and dense neural networks for classifying maritime terms. In *14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–6. IEEE, 2023.

[27] J. Nivre, M. de Marneffe, F. Ginter, J. Hajic, C. D. Manning, S. Pyysalo, S. Schuster, F. M. Tyers, and D. Zeman. Universal dependencies v2: An evergrowing multilingual treebank collection. In *12th Language Resources and Evaluation Conference (LREC)*, pages 4034–4043, 2020.

[28] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji. Cross-lingual name tagging and linking for 282 languages. In *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1946–1958. Association for Computational Linguistics, 2017.

[29] S. Petrov, D. Das, and R. T. McDonald. A universal part-of-speech tagset. In *8th International Conference on Language Resources and Evaluation (LREC)*, pages 2089–2096. European Language Resources Association (ELRA), 2012.

[30] B. Plank, A. Søgaard, and Y. Goldberg. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *54th Annual Meeting of the Association for Computational Linguistics (ACL)*. The Association for Computer Linguistics, 2016.

[31] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, pages 101–108. Association for Computational Linguistics, 2020.

[32] N. Reimers and I. Gurevych. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348. Association for Computational Linguistics, 2017.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[34] E. F. T. K. Sang. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *6th Conference on Natural Language Learning (CoNLL), Held in cooperation with (COLING)*. ACL, 2002.

[35] C. Saravanos and A. Kanavos. Forecasting stock market alternations using social media sentiment analysis and deep neural networks. In *14th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8. IEEE, 2023.

[36] C. Saravanos and A. Kanavos. Forecasting stock market volatility using social media sentiment analysis. *Neural Computing and Applications*, 37(17):10771–10794, 2025.

[37] A. Savvopoulos, A. Kanavos, P. Mylonas, and S. Sioutas. LSTM accelerator for convolutional object identification. *Algorithms*, 11(10):157, 2018.

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. ACL, 2013.

[39] M. Straka and J. Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with udpipe. In *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99. Association for Computational Linguistics, 2017.

[40] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*. The Association for Computational Linguistics, 2003.

[41] G. Vonitsanos, A. Kanavos, and P. Mylonas. Decoding gender on social networks: An in-depth analysis of language in online discussions using natural language processing and machine learning. In *IEEE International Conference on Big Data*, pages 4618–4625, 2023.

[42] W. Wagner. Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit - o'reilly media. *Language Resources and Evaluation*, 44(4):421–424, 2010.

[43] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. LUKE: deep contextualized entity representations with entity-aware self-attention. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454. Association for Computational Linguistics, 2020.