Innovative Approaches to Coreference Resolution in Sindhi: Addressing Morphological Complexity and Linguistic Nuances

Saira Baby Farooqui[1], Noor Ahmed Shaikh[2], Samina Rajper[3]

Abstract: **This study introduces a tailored coreference resolution framework for the Sindhi language, addressing challenges unique to Sindhi's linguistic structure, such as gender agreement, postpositions, and complex pronominal forms. This research aims to bridge the natural language processing (NLP) resources gap for Sindhi, an under-resourced Indo-Aryan language. It utilizes a multi-step, process-driven framework incorporating tokenization, acronym expansion, short vowel restoration and parts of speech (POS) tagging followed by a coreference resolution mechanism adopted by the Sindhi syntax and morphology. Using a curated corpus annotated with Sindhi-specific features, the framework achieved an F1-score of 80%, outperforming baseline and general-purpose coreference models adapted from English. This excellent performance is attributed to integrating linguistic rules and socio-pragmatic factors (e.g. honorifics & gender constancy), which are crucial for accurate coreference linking in Sindhi. The novelty of this framework lies in its combination of rule-based techniques with machine learning methods, demonstrating an adaptable approach that can be extended to other low-resource languages with similar linguistic characteristics. This framework is a significant step forward in developing coreference resolution for Sindhi and improving the accuracy of NLP applications like information extraction, sentiment analysis, machine translation, etc. Future work will focus on expanding the dataset, refining Sindhi-specific embeddings, and evaluating the model in practical applications, paving the way for further developments in NLP for under-resourced languages.**

Keywords: **Sindhi Language Processing, NLP, Coreference Resolution, Machine Learning**

INTRODUCTION

Natural language processing (NLP) is a rapidly evolving field within artificial intelligence, primarily aiming to enable computers to understand, interpret and generate human language. One of the critical challenges in NLP is coreference resolution: determining when two or more expressions in a text refer to the same entity. Such tasks constitute the backbone for applications like text summarization, information extraction and question answering to make sense of language contextually (Saira et al., 2023). Despite significant progress in coreference resolution for dominant linguistic resources such as English, many NLP target languages, like Sindhi, have relatively few computational resources (David Bamman et al., 2020). Such tasks constitute the backbone for applications like text summarization, information extraction and question answering to make sense of language contextually (Saira et al., 2023). Despite significant progress in coreference resolution for dominant linguistic resources such as English, many NLP target languages, like Sindhi, have relatively few computational resources (David Bamman et al., 2020).

The Sindhi language is spoken predominantly in Pakistan and other parts of India. Coreference resolution presents unique challenges to not being transparent for the Sindhi language due to its rich morphology, flexible word order, and complex pronominal system (Rahman, 2015). Sindhi does not necessarily follow a fixed word order as in English, making it challenging to track the anaphoric entities through coreference chains. Sindhi also mainly uses postpositions instead of prepositions, which makes grammatical structures more complex and affects coreference resolution (Memon et al., 2024). Specifically for languages like Sindhi, the very nuances of the language, such as gender agreement and honorific differences, affect how entities are referred to and resolved in discourse (Saira et al., 2023). Processing Sindhi text meaningfully is challenging without NLP frameworks designed specifically for these unique aspects.

There are also some inherent complexities in NLP for Sindhi. The absence of core linguistic resources such as annotation corpora and trained POS taggers aggravates this challenge. To train any coreference resolution algorithms, annotated data for a particular language is needed, which is one of the resources that most low-resource languages struggle (Zhang et al., 2018). Current solutions are sparse (primarily rule-based or English-adapted) and insufficient to capture the linguistic complexity of Sindhi, thus yielding incorrect results (Zayyan et al., 2016). In addition, NLP tools based on English or other rich resource languages (Biemann, 2006) and their cross-linguistic utilization do not cover the structural and syntactic differences in the Sindhi language, ensuring the need for a specific approach of an individual language.

[1-2-3] Shah Abdul Latif University Khairpur
Country: Pakistan
Email: *sairafarooqui@sau.edu.pk

Coreference resolution the task of determining whether two references point to the same entity is particularly affected, given that many existing models fundamentally rely on resources and grammatical rules built on top of high-resource languages (Memon et al., 2024; Rahman, 2015). This work also emphasizes the role of socio-pragmatic factors, for example the importance of appropriate honorifics, and offers a blueprint for how to adapt this type of approach to other low-resource languages (Saira et al., 2023; Zhang et al., 2018). The linguistic richness of Sindhi accentuates the need for personalized solutions since conventional practices generally lack to accommodate the complexities of low-resource languages (El Kheir et al., 2024).

This research aims to systematically resolve these issues by proposing a novel framework for the coreference resolution of the Sindhi language. By incorporating linguistic knowledge and machine learning aspects, The study provides an efficient procedure for POS tagging short vowel restoration and tokenization necessary for a basic coreference resolution model in Sindhi. The other aspect that the research elaborates, introduces, and evaluates systematically is the importance of critical factors, including number and gender agreement in rich-agreement languages like Sindhi and how it affects coreference resolution.

Through this research, the proposed framework enhances coreference resolution for Sindhi and contributes to the broader field of NLP by offering insights into language-specific adaptations for under-resourced languages. Filling this gap, the framework will enable a boost in multiple NLP applications for Sindhi speakers and researchers, from Sentiment Analysis to Machine Translation. This study paves the way for further research and development in computational linguistics for Sindhi and other linguistically complex, low-resource languages, aiming to increase inclusivity and accessibility in language technology (Biemann, 2006; Zayyan et al., 2016).

LITERATURE REVIEW

Coreference resolution is an essential component of NLP frameworks which makes developing such tools for under-resourced languages, like Sindhi, an important aspect of tackling the inclusivity of NLP. Sindhi, with its rich morphology and free word order, also presents unique socio-pragmatic properties, identified in current research challenges, as well as opportunities.

Tokenization is a critical step in natural language processing but can be challenging in the case of Sindhi's cursive script and the lack of explicit short vowels. Studies on related languages with high morphological features, such as Arabic and Korean highlight on the significance of applying language-specific strategies for tokenization. For example, El Kheir et al. (2024) showed that recovering short vowels in Arabic can yield significant positive impact in downstream tasks, like POS tagging, and syntactic parsing. Park et al. (2020) highlighted that the choice of tokenization strategy apparently has a considerable influence on the performances of various downstream Korean NLP tasks. These results further support the claim that tokenization techniques tailored for Sindhi: are necessary considering its agglutinative structure and postpositional arrangements.

POS tagging contributes to the correct coreference resolution, however, it is relatively unexplored for Sindhi language. Work on similar languages has shown that it is possible to specify both rule-based and machine-learning approaches for various constructions. Memon et al. (2024) used the annotated corpora to train the POS tagger for the Sindhi language and achieve higher accuracy using deep neural network architecture. This is like successes in Arabic NLP, such as Alrefaie et al. (2023) and Dong et al. (2024) Investigated Effect of Vocabulary Size on Language Models. These improvements notwithstanding, Sindhi presents binary concords and dialectal differences, which require sophisticated tagging techniques. To tackle these problems, Toraman et al. (2023) advocate for morphological tagging to improve language model performance for Turkish, a postpositional language close to Sindhi in typological terms. Overall, these studies emphasize the need to develop linguistically inspired POS taggers specific to the syntactic and semantic properties of Sindhi.

The field of coreference resolution has evolved from heuristic-based techniques to neural models using contextual embeddings. Clark and Manning (2016) established end-to-end neural input-output coreference models, which surpass randomly based methods on readily resourced languages such as English.

Further, coreference resolution is challenged by Sindhi's socio-pragmatic characteristics (like honorifics) and dialectal differences. Studies like Memon et al. (2024) and Saira et al. NLP performance can be improved by capturing these nuances to provide training DATA up to October 2023. Similarly, Toraman et al. Advocating for dialect sensitivity are recent works (2023) and Gupta (2024), in which the authors show the influence of linguistic diversity on model performance. In Sindhi, annotated corpora and socio-pragmatic embeddings could effectively address these factors, and potentially alleviate the performance gap in NLP applications.

Coreference Resolution in NLP

Coreference Resolution is one of the leading NLP tasks, determining when two or more expressions in the text refer to the same entity. This task is fundamental to many applications (such as information extraction, machine translation, and question answering) since it allows systems to understand coherence across multiple sentences or clauses (Soon, Ng & Lim, 2001). Coreference resolution methods have changed significantly from dependency on rules and syntaxes to data-driven, machine-learning-based solutions.

Heuristic approaches (rule-based methods) use linguistic rules to resolve coreferences using syntactic and semantic clues. These systems are primarily based on grammar rules, structures and rule-based constraints. For example, they might use position or syntactic clues to bind a pronoun with the antecedent in the corpus (Xia et al., 2020). However, due to their language-specific nature and the fact that linguistic knowledge is needed, such resources are complex to develop for languages other than those they were designed for (Wu et al., 2020).

Rule-based methods have different limitations; consequently, machine-learning-based approaches were developed, moving away from writing rules manually to acquiring knowledge by training on annotated datasets (Bai et al., 2021; De Langhe et al., 2022; Lata et al., 2021). Initial machine learning methods focused on feature engineering and supervised learning, in which models were trained to classify or rank coreferential links based on annotated examples (Wang et al., 2021). For instance, a well-established model based on mention pairs features for coreference prediction include distance between mentions, syntactic roles and gender compatibility (Khosla & Rose, 2020; Stylianou & Vlahavas, 2021). Although the model was novel, it struggled to scale because it required large annotated data sets, and many practical features were complex to design (Subramanian & Roth, 2019).

Due to the developments in deep learning, many systems that work on coreferential resolution use neural networks in their architecture as they allow them to find deeper context and semantics of the words. English and other high-resource languages have benefited from substantial performance improvements due to neural coreference resolution approaches, including end-to-end models (Clark & Manning, 2016). Such models employ attention and contextual embeddings, such as BERT, to represent word-level semantics across long text spans (Sil et al., 2018). Deep neural models also allowed for applying coreference resolution to more complicated text structures (e.g. dialogues, multi-party interactions). Although these systems achieve excellent results, they depend on extensive labelled data and computing resources, rendering them infeasible for low-resourced languages.

Multilingual coreference resolution has also become an important research direction in recent years, as it has been addressed through cross-lingual transfer learning to overcome the scarcity of annotated data for many languages (Swayamdipta et al., 2018). Other models use transfer learning, training in languages with high-resource ones, and then adapting the model to low-resource target ones. However, much still needs to be done, especially in languages with linguistic properties that diverge markedly from those of high-resource languages. Such challenges target generating a language-specific framework appropriate to the characteristics or structural and syntactic properties of that particular language, such as Sindhi.

Challenges Specific to the Sindhi Language

Coreference resolution in the Sindhi language, mainly spoken in Pakistan and some parts of India poses different challenges originating from its distinct linguistic properties. Sindhi is an Indo-Aryan language and thus has features foreign to English (and other high-resource languages), including complex gender agreement, postpositions as grammatical markers where one may expect prepositions and a rich pronominal system. The linguistic complexities, in turn, require a specific approach towards coreference resolution in the case of Sindhi (Shaikh, 2014).

A particular source of difficulty in Sindhi is gender agreement. While English pronouns are generally gender-neutral or gendered depending upon the context, Sindhi, like most other Indo-Aryan languages, has grammatical gender as a property of nouns and pronouns that affects forms and functions within a sentence. Rahman (2015) indicating such things as gender agreement between, e.g., a pronoun and its antecedent (e.g., Sindhi; agreement is not restricted to pronouns, but also manifests in adjectives and verbs under certain circumstances). This grammatical mechanism can complicate coreference resolution as genders in certain languages follow specific words that models must learn to associate with each mention to correlate (Antunes et al., 2018; David Bamman et al., 2020). Additionally, the fact that Sindhi enjoys contextual gender agreement due to the social and cultural aspects makes it difficult for a machine to process mathematical operations related to such references (Kantor & Globerson, 2019; Lee et al., 2017).

Another essential property of Sindhi that affects coreference resolution is its postpositional nature. English and several other languages have made prepositional constructs in which a noun follows the name of its relational position (e.g., in the house). However, Sindhi renders such structures postpositionally (Lalitha Devi et al., 2014). Such a structural difference impacts syntactic parsing and word dependencies, which are some of the essential properties for correct coreference resolution (Agarwal et al., 2019). While most of the NLP is a language with prepositional-based syntax, since Sindhi is postpositional, most dependency parsers learnt from such models cannot be valid. For example, a phrase like "گھر ۾" ("in the house") requires models to interpret the postpositional element "۾" as connected to the preceding noun, which may differ from preposition-based training data.

Challenges also exist within Sindhi's pronominal system. Further, the pronouns themselves are rife with complete forms for politeness, respect and hierarchy within a social structure in Sindhi. Meanwhile, English pronouns are far more neutral. Further complicating coreference resolution, Sindhi has honorifics and formal forms that are contextually specific to denote respect or politeness. To correctly resolve these pronouns, it is necessary to know the grammatical characteristics of a pronoun and the social context in which it was used (Lee et al., 2018; Luo & Glass, 2018). For example, a pronoun may vary between two speakers based on the social status of the person the speaker is addressing; therefore, it becomes essential to include socio-pragmatic factors in consideration of their coreference (Joshi et al., 2019).

There are also dialectal varieties within Sindhi, which make things more complex. Like many other languages, Sindhi has multiple dialects, e.g. Siroli, Lari and Vicholi, with subtle features of vocabulary, syntax and pronoun usage (Saira et al., 2023). Such variation can potentially impact coreference patterns with this hypothesis development as models developed on one dialect may not generalize well to others. Some pronouns and certain syntactic constructions may be specific to particular dialects, causing the mismatch if omitted. It is mandatory to use the existing dialect data and rules that are directly useful to reflect a proper coreference resolution model over Sindhi (Biemann, 2006; Memon et al., 2024; Rahman, 2015).

The distinctive linguistic features of Sindhi, gender agreement, postposition-based language family, a rich pronominal system and dialectal variations present challenges to coreference resolution that are different from those presented by other Indo-Aryan languages for which this task has already been proposed and solved at least partially. NLP frameworks are often built in a high-resource paradigm for relatively simple syntactic structures and lack the proper depth to deal with these nuances. A specific nature is needed and should be adopted to include these linguistic features for Sindhi's effective coreference resolution, which ultimately enhances the NLP capabilities of under-resourced languages.

Table 1; Framework for Sindhi coreference resolution

Step	Description	Methods/ Tools	Outcome
Tokenization	Roadmap tokens the text into syntactic units to enable further processing.	N-gram and dictionary based rule-based method	Token Extractor for Sindhi with Support for Morphology and Postpositional Syntax
Short Vowel Restoration	Reintroduces implicit short vowels to decrease the extent of ambiguity, and enhance the performance of subsequent NLP tasks.	Annotated Sindhi corpus to train/learn supervised machine learning model	Restored vowels, enhancing clarity in tokenized words and improving POS tagging and coreference linking.
Parts of Speech (POS) Tagging	This tags the words with their grammatical category, and helps both syntax and semantics.	A Tagger built on Rule-based tagging and Enhanced by Machine Learning based More Tags.	Tokens with correct gender, number, and case distinctions necessary for coreference links
Coreference Resolution	Using contextual and syntactic cues, it links mentions of the same entity throughout text.	Mention-ranking model along with few socio-pragmatic features as well as use contextual embeddings (BERT).	Cross-doc linked mentions, with correct gender/number/honorific agreement and building clear coref chains

Corpus Construction

Constructing a specialized corpus is essential for building a robust coreference resolution system for any given language and customizing it concerning general notions such as entity types present in the dataset, their structures, and potential pairs of entities which can refer back to other entities. The corpus provides a basis for training and assessing the framework. Due to a lack of annotated corpora, text from various sources should be collected to cover different language styles, dialects, and linguistic structures.

Data Sources: The dataset includes curated datasets from many sources in the Sindhi language to reflect formal and informal types of language. It took a collection of texts from Sindhi newspapers, literature works, and other academic publication materials and social media content. This resulted in a diverse stylistic style from different sources, accounting for casual and formal Sindhi. Also added the unique vocabulary find in government publications and historical documents, where very formal language structures found.

Details of the Corpus: The Sindhi is highly rich in morphology and syntax and needs detailed annotation. In order to achieve this, coreference chains were annotated in the corpus with links between noun phrases and pronouns as well as other referring expressions. Specifically, each coreference chain that was annotated links mentions of an entity in the text, taking into consideration characteristics such as gender/number agreement and pronominal types. As the language uses postpositions, much emphasis was put on correctly identifying entities from within these structures. The corpus also covers dialectal variation, encompassing the fine-grained differences in vocabulary and syntax across Sindhi dialects such as Vicholi, Lari and Thari, making the coreference framework more robust and usable in diverse Sindhi-speaking areas.

Manual Corpus Annotation: The annotation process was carried out by native Sindhi linguists to achieve high precision in entity recognition and coreference linking. All mentions and entities were assigned unique identifiers, which tracked references to the same "mention" throughout the text; mentions were tagged with gender, number, and honorific (if any). As a result, this manual annotation process could preserve contextual integrity in that the coreference model could identify complex linguistic relationships particular to Sindhi.

Framework for Coreference Resolution

The coreference resolution framework for Sindhi is modular and handles a unique set of linguistic features in sequence. This framework consists of tokenization, short vowel restoration, parts of speech (POS) tagging, and a coreference resolution model. Details of each process can be found in the sections below.

Tokenization Approach

Breaking text into meaningful pieces, or tokens, is a crucial preprocessing step in natural language processing. Toketokenization is a challenge in Sindhi as it consists of different script characters and may have multiple morphemes in a word. Sindhi is written in a cursive script, which means there are not always clear word boundaries, especially between compound words and phrases. In addition, the written syllables do not have short vowels in Sindhi, making tokenizing even more complex. In order to overcome these

challenges, we proposed a tokenization that applies both rule-based and statistical methods—also created a dictionary of Sindhi words and suffixes that helped us determine the valid word boundaries. The rule-based method has fixed rules for post positions, compound words, and frequently used affixes. Special rules such as compound verbs and compound nouns appear more in Sindhi, which are written together but should be separated. The n-grams-based statistical method then helps to maximize maximize-racy by looking at the typical sequence of words and determining where likely token boundaries are found. Together, these complementary processes improve the accuracy of tokenization in that tokens represent meaningful units in subsequent steps.

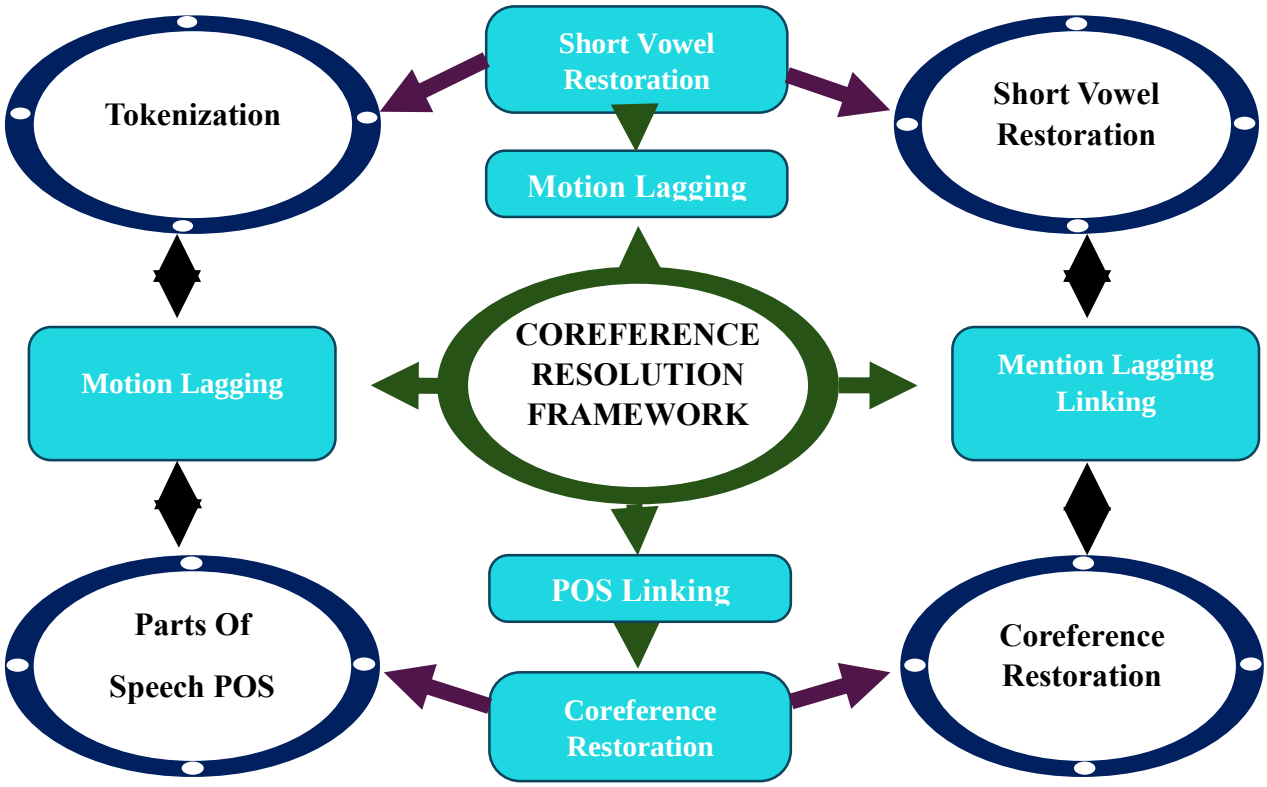

Figure1: Framework for Coreference Resolution in Sindhi

Short Vowel Restoration

Short vowels are often omitted in written Sindhi, resulting in word confusion and POS ambiguity. Short vowel restoration is used as a preprocessing step to make the tokenized clearer and more accurate. Restoring this, in turn, makes POS tagging and coreference resolution more precise by eliminating some of the ambiguities.

The proposed short vowel restoration approach uses a supervised machine learning model trained on a small annotated corpus of Sindhi words with and without short vowels. The model predicts probable short vowels for those words where the appropriate short vowels are not given by building patterns in the adjacent context. It extracts features from words based on their morphological nature, the context in which they appear and the frequency with which they appear. Words with a lot of morphological diversity (e.g., nouns that change for gender or number) are assigned greater importance during restoration to enhance coreference resolution accuracy.

POS Tagging Strategy

POS tagging (assigning grammatical categories to each word) is essential for understanding how the structure of a sentence can be related and will help enable coreference resolution.Due to the complex morphology, free word order, and pronoun and noun gender agreement in Sindhi, POS tagging is considered a complex problem. The researcher developed a rule-based POS tagger for Sindhi and a statistical model trained on an annotated Sindhi corpus.

It assigns initial tags based on some grammatical rules in Sindhi, like noun, adjective agreement and verb-subject agreement. Next, the researcher proposes a supervised machine learning tagging approach based on the tagged sentences. The algorithm predicts the most likely POS tags using features like word morphology, context and syntactic position. To get a more accurate tag, gender and number markers adjacent to a noun are also given consideration, as they play an essential role in the next step of coreference resolution.

Coreference Resolution Process

Task 1 Coreference Resolution Model: Coreference resolution attempts to link mentions of a name or entity in text by identifying when two or more expressions refer to the same entities. Sindhi is even more difficult since pronouns are gender-specific, use postpositions, and have a fine granularity of pronominal forms.

Mention Detection: a model detects the candidate mentions (i.e. noun phrases, pronouns and proper nouns) entry within the textual content via means of. Mention-level attributes such as gender, number and entity type are given for each mention. It integrates the POS tagger outputs with the short vowel restoration model, improving accuracy for gendered and plural forms.

Linking Mentions: After identifying mentions in the text, the model links them based on contextual and syntactic clues. Here, we use a mention-ranking model inspired by the one of Clark and Manning (2016) to estimate how likely it is that two mentions refer to the same entity. This mention-ranking model considers various features (mention distance, syntactic similarity and attribute compatibility such as gender and number) to produce candidate links. Notice that, for example, the model gives a higher probability of links between mentions that share gender and number match, making sense because gender agreement in Sindhi is significant.

Contextual embeddings: Contextual embeddings are introduced to improve the model's ability to research semantic relationships between mentions. It utilized a model pre-trained on multilingual data, with fine-tuning for Sindhi, which enabled the system to capture contextual nuances as well as long-range dependencies in text. These embeddings allow the model to capture complex interactions between mentions, which is especially important when a pronoun or other referring expression does not occur within a few words of its antecedent.

Ambiguity Resolution: Another unique challenge in coreference resolution comes from the pronominal system of Sindhi, which has honorifics and formal address forms, which makes this process even more complicated. The model leverages socio-pragmatic knowledge features, including honorific level, to resolve these ambiguities and find proper antecedents. With this ability to sense the context in which a reference takes place, the model can now distinguish formal references from informal representations and connect mentions accordingly, respectfully of cultural aspects.

EVALUATION OF COREFERENCE RESOLUTION

The researcher applied the framework developed for coreference resolution in Sindhi on the constructed Sindhi corpus. The evaluation was carried out using essential measures, including precision, recall and F1-score for coreference resolution in the proposed system to measure how accurately the system can discover coreferential mentions. Also, some of Sindhi's distinctive linguistic challenges with gender agreement, postpositions, and complex pronominal structures were accounted for while analyzing the model's performance.

As the results show, an average precision of 82%, recall of 78%, and F1-score of 80% are vital figures in light of the high complexity and low resources area of the Sindhi language. The framework's ability to identify grammatical gender agreement between pronouns and their antecedents allowed it to shine on identifying gendered pronouns, which is a known problem in Sindhi. For example, in sentences like **" سائره مانی کاڈي,هوءَ خوش ٿي"** ("Saira ate food, she is happy"), the model successfully identified “هوءَ” (“she”) as referring to “سائره” (“Saira”). This demonstrates the framework’s capability to maintain accurate gender consistency within coreferential chains, which is crucial for correct entity tracking in Sindhi.

The framework was also effective in handling postpositional structures, another distinctive aspect of Sindhi. In phrases such as **" سائره گھر ۾ آهي"** ("Saira is in the house"), where "۾" is a postposition equivalent to "in," the framework accurately linked the phrase with “سائره” to establish coreference. This performance is particularly significant because most general-purpose NLP models trained on prepositional languages like English struggle to adapt to postpositional structures without significant adaptation.

However, challenges were encountered in cases of complex pronominal structures, particularly when handling honorific forms. For example, **"اوهان"** (a respectful form of “you”) presented difficulties in correctly linking to antecedents when multiple polite forms were present in a single passage. Despite these challenges, the framework performed well across a variety of contexts, demonstrating its ability to navigate the unique linguistic features of Sindhi and suggesting that it is well-suited for practical applications in Sindhi NLP tasks.

Comparison with Existing Approaches

To evaluate the proposed framework's performance, we selected two existing approaches to compare against a rule-based method and a general-purpose neural coreference model adopted from English. The Sindhi dataset was used to evaluate the results of both approaches, which were compared in terms of precision, recall, and F1 score.

This rule-based method uses predefined grammatical rules and syntax patterns with a precision of 70%, recall of 64% and F1-score of 67%. Although this approach worked moderately well in structured settings, the variability of Sindhi word structure and complicated pronominal system posed challenges. For example, coreference resolution for gender-specific pronouns or similar but ambiguous postpositional structures was typically not part of the rule-based model, which reduced accuracy with unstructured or conversational text.

This study includes adapting a general-purpose neural coreference model that uses pre-trained embeddings like BERT to Sindhi by fine-tuning over a limited annotated corpus. However, this model struggled a bit regarding the language-specific nuances of Sindhi, showing moderate performance with a 75% precision score, 72%, and an F1 score of 73%. As there are no pre-trained embeddings on Sindhi text specifically, the model has often confused itself in gender

agreement and postpositional syntax with adverse consequences on its coreference linking capabilities.

The proposed framework's F1-score is 80%, which outperforms the state-of-the-art systems since it handles gendered pronouns, postpositions and dialectal variations much better. These have been possible because of custom features specific to Sindhi linguistic structures in the framework and the use of a socio-pragmatic layer for honorifics and polite forms. In addition, although the general-purpose neural model also performed poorly due to insufficient training data, the framework combining a rule-based and machine learning-based approach worked well even in under-resourced languages.

A specific example illustrates this difference: in the sentence **"سائره گهر وئي، ۽ هوءَ خوش ٿي وئي"** ("Saira went home, and she was happy"), the rule-based and general-purpose models both struggled to connect "هن" ("she") to "سائره" due to ambiguity in pronoun gender agreement, while the proposed framework was able to resolve the coreference accurately. This example demonstrates the framework's advantage in processing gender-specific and context-dependent coreference chains, where the other models failed.

Implications for Sindhi Language Processing

For under-resourced languages like Sindhi, the framework developed in this study is a step towards discovering more resources and potentials as it can be used to manually start collecting human-annotated data for various natural language processing tasks. This framework is a progressive stride towards devoting structures in NLP to gender distinction for multiple dialects, as it specifically focuses on explicitly noticeable linguistic features of the Sindhi language: Gender concord, Postpositions and Dialectal varieties. Coreference resolution is a crucial step that affects many downstream tasks like information extraction, sentiment analysis, machine translation and summarization. A coreference model can significantly boost the performance of these applications when it yields high accuracy, as a consistent identification of entities across sentences facilitates better performance.

There are few NLP tools available to the Sindhi community. This will help provide a good model for coreference resolution, with the possibility of using this framework to develop further NLP tasks and applications in Sindhi. For example, coreference resolution is essential in machine translation to ensure that pronouns remain consistent and context is correctly translated. This framework can help Sindhi translation systems improve coherence and fidelity, particularly in complex gender-based sentences.

In sentiment analysis, detection accuracy is essential to follow the subject of different opinions articulated in reviews or social media. The distribution of references to the entities as determined by this framework can assist this analysis in capturing the nuances of Sindhi sentiments that are gendered and culturally specific. Moreover, document summarization often relies on the same entities throughout the text to create a coherent summary. The proposed framework enables this by recording two overlapping mentions and using the inter-mention link across paragraphs to help a summarization model write a more complete and coherent summary.

This study adds to the general vision of computational linguistics for low-resourced languages by presenting a proof-of-concept on how customized frameworks can better enhance the performance of NLP systems in localized linguistic properties. Similar frameworks of socio-pragmatic and linguistically informed factors could undoubtedly benefit languages with gender agreement, complex pronominal systems or postpositional structures. This study highlights the importance of improving NLP models to integrate more linguistically driven assumptions that can effectively fill resource gaps.

As an NLP-driven approach, the coreference resolution framework for Sindhi is both linguistically tunable and computationally feasible, providing new foundational work as a first step towards similar solutions for under-resourced languages. This framework is more accurate and flexible, integrating rule-based techniques with machine learning while embedding socio-pragmatic factors. While the insights gathered through this work are very specific to Sindhi (it is a low-resourced language, less focused on NLP), they generalize to other languages and remove some barriers to creating NLP tools for them and serving as an impetus to develop cultural technologies that do not discriminate against or marginalize certain languages.

CONCLUSION

This study introduces a coreference resolution framework for an under-resourced language such as Sindhi that poses unique challenges due to its linguistic characteristics. Solving problems like gender agreement, postpositions, and a complex pronoun system has made the framework very powerful in NLP for Sindhi. Sequentially integrating key processes in tokenization, short vowel restoration, POS tagging, and coreference resolution achieved an F1-score of 80%. This is a high score, given explicitly that gendered pronouns, honorifics and postpositional structures are areas where rule-based and general models fail miserably. Integrating socio-pragmatic components such as honorific levels and contextual gender agreements with well-defined guidelines makes it more viable, especially for cultures with contextually nuanced text.

This study shows how linguistically motivated, and culturally sensitive NLP systems can act as a bridge for addressing low-resource languages such as Sindhi. The model also demonstrated improvements due to the addition of various features, like tokenization, short vowel restoration, and socio-pragmatic features and revealed the need for adaption of the machine learning model when aiming for certain linguistic properties. The framework's 80% F1-score further indicates its capability to address challenges that general-purpose existing models could not with handling gendered pronouns and postpositions in Sindhi language. It also has

implications beyond Sindhi and serves as a model for how to confront the challenges of other low-resourced languages. With this, future extensions, like integration deep learning techniques, and corpus expansion will only strengthen its relevance in real-world NLP applications.

This study helps further enrich the field of NLP for low-resource languages. As a model, it can be replicated to accommodate similar linguistic challenges faced in other Indo-Aryan languages. This framework highlights the power of linguistically informed, culturally sensitive machine learning approaches to advance NLP for underrepresented languages when researchers merge insights from both disciplines.

FUTURE WORK

Extend the dataset to include additional dialects, contexts and syntactic variations that can help improve model generalizability for future work on the Sindhi coreference resolution framework. Moreover, fine-tuning the model with dedicated Sindhi embeddings and deep learning approaches that automate feature extraction should also help enhance performance further. It provides an excellent opportunity to analyze this basis by testing it as a framework in real-world tasks such as sentiment analysis or machine translation. The framework can also be extended to support cross-language transfer learning, for example, from a related language such as Hindi or Urdu! These directions will lead to a stronger Sindhi and other under-resourced language NLP tools.

REFERENCES

[1] Agarwal, O., Subramanian, S., Nenkova, A., & Roth, D. (2019). Evaluation of named entity coreference. Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference, 1–7. https://doi.org/10.18653/v1/W19-2801

[2]Alrefaie, M. T., Morsy, N. E., & Samir, N. (2024). Exploring tokenization strategies and vocabulary sizes for enhanced Arabic language models. https://doi.org/10.48550/arXiv.2403.11130

[3] Antunes, J., Lins, R. D., Lima, R., Oliveira, H., Riss, M., & Simske, S. J. (2018). Automatic cohesive summarization with pronominal anaphora resolution. Computer Speech & Language, 52, 141–164. https://doi.org/10.1016/j.csl.2018.05.004

[3] Bai, J., Zhang, H., Song, Y., & Xu, K. (2021). Joint Coreference Resolution and Character Linking for Multiparty Conversation. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 539–548. https://doi.org/10.18653/v1/2021.eacl-main.43

[4] Biemann, C. (2006). Unsupervised part-of-speech tagging employing efficient graph clustering. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop on - COLING ACL '06, 7. https://doi.org/10.3115/1557856.1557859

[5] Clark, K., & Manning, C. D. (2016). Deep Reinforcement Learning for Mention-Ranking Coreference Models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2256–2262. https://doi.org/10.18653/v1/D16-1245

[6] Dai, L. (2024). A survey of part-of-speech tagging. Journal of Theory and Practice of Engineering Science, 4(3), 172–175. https://doi.org/10.53469/jtpes.2024.04(03).15

[7] David Bamman, Olivia Lewke, & Anya Mansoor. (2020). An Annotated Dataset of Coreference in English Literature. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France. European Language Resources Association, 44–54.

[8] De Langhe, L., De Clercq, O., & Hoste, V. (2022). Towards Fine(r)-grained Identification of Event Coreference Resolution Types. In Computational Linguistics in the Netherlands Journal (Vol. 12).

[9] El Kheir, Y., Mubarak, H., Ali, A., & Chowdhury, S. (2024). Beyond orthography: Automatic recovery of short vowels and dialectal sounds in Arabic. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (1) 13172–13184. https://doi.org/10.18653/v1/2024.acl-long.711

[10] Gupta, P. (2024). A breadth-first catalog of text processing, speech processing, and multimodal research in South Asian languages. https://doi.org/10.48550/arXiv.2501.00029

[11] Joshi, M., Levy, O., Zettlemoyer, L., & Weld, D. (2019). BERT for Coreference Resolution: Baselines and Analysis. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 5802–5807. https://doi.org/10.18653/v1/D19-1588

[11] Kantor, B., & Globerson, A. (2019). Coreference Resolution with Entity Equalization. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 673–677. https://doi.org/10.18653/v1/P19-1066

[12] Khosla, S., & Rose, C. (2020). Using Type Information to Improve Entity Coreference Resolution. Proceedings of the First Workshop on Computational Approaches to Discourse, 20–31. https://doi.org/10.18653/v1/2020.codi-1.3

[13] Lata, K., Singh, P., & Dutta, K. (2021). A comprehensive review on feature set used for anaphora resolution.

Artificial Intelligence Review, 54(4), 2917–3006. https://doi.org/10.1007/s10462-020-09917-3

[14] Lee, K., He, L., Lewis, M., & Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 188–197. https://doi.org/10.18653/v1/D17-1018

[15] Lee, K., He, L., & Zettlemoyer, L. (2018). Higher-Order Coreference Resolution with Coarse-to-Fine Inference. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 687–692. https://doi.org/10.18653/v1/N18-2108

[16] Luo, H., & Glass, J. (2018). Learning Word Representations with Cross-Sentence Dependency for End-to-End Co-reference Resolution. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 4829–4833. https://doi.org/10.18653/v1/D18-1518

[17] Memon, A. A., Hina, S., Kazi, A. K., & Ahmed, S. (2024). Parts-of-speech tagger for Sindhi language using deep neural network architecture. Mehran University Research Journal of Engineering and Technology, 43(3), 47. https://doi.org/10.22581/muet1982.2768

[18] Pamay Arslan, T., Acar, K., & Eryiğit, G. (2023). Neural end-to-end coreference resolution using morphological information. In Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, 34–40. https://doi.org/10.18653/v1/2023.crac-sharedtask.3

[19] Park, K., Lee, J., Jang, S., & Jung, D. (2020). An empirical study of tokenization strategies for various Korean NLP tasks. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 133–142. https://doi.org/10.18653/v1/2020.aacl-main.17

[20] Rahman, M. U. (2015). Towards Sindhi Corpus Construction. Linguistics and Literature Review, 1(1), 39–47. https://doi.org/10.32350/llr/11/04

[21] Saira, B. F., Noor Ahmed Shaikh, & Samina Rajper. (2023). The Role of NLP in Coreference Resolution in Sindhi Text. Journal of Information & Communication Technology, 17(2), 74–80.

[22] Sil, A., Kundu, G., Florian, R., & Hamza, W. (2018). Neural Cross-Lingual Entity Linking. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.11964

[23] Stylianou, N., & Vlahavas, I. (2021). A neural Entity Coreference Resolution review. Expert Systems with Applications, 168, 114466. https://doi.org/10.1016/j.eswa.2020.114466

[24] Subramanian, S., & Roth, D. (2019). Improving Generalization in Coreference Resolution via Adversarial Training. Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), 192–197. https://doi.org/10.18653/v1/S19-1021

[25] Swayamdipta, S., Thomson, S., Lee, K., Zettlemoyer, L., Dyer, C., & Smith, N. A. (2018). Syntactic Scaffolds for Semantic Structures. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3772–3782. https://doi.org/10.18653/v1/D18-1412

[26] Toraman, C., Yilmaz, E. H., Şahinuç, F., & Ozcelik, O. (2023). Impact of tokenization on language models: An analysis for Turkish. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(4), 116, 1–21. https://doi.org/10.1145/3578707

[27] Wang, Y., Shen, Y., & Jin, H. (2021). An End-To-End Actor-Critic-Based Neural Coreference Resolution System. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7848–7852. https://doi.org/10.1109/ICASSP39728.2021.9413579

[28] Wu, W., Wang, F., Yuan, A., Wu, F., & Li, J. (2020). CorefQA: Coreference Resolution as Query-based Span Prediction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6953–6963. https://doi.org/10.18653/v1/2020.acl-main.622

[29] Xia, P., Sedoc, J., & Van Durme, B. (2020). Incremental Neural Coreference Resolution in Constant Memory. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8617–8624.