Development of Annotated Corpus Resources of Sindhi

Mutee U Rahman[1], Tafseer Ahmed[2], and Muhammad Shaheer Memon[3]
[1,3] Isra University, Hyderabad, [2]Mohammad Ali Jinnah University, Karachi
muteeurahman@gmail.com, tafseer@gmail.com, shaheer.memon@isra.edu.pk

Abstract

We present ongoing work on the development of an annotated corpus resources project for Sindhi. A multi-layer annotation model is presented and experimentally applied on a subset of an existing plaintext Sindhi corpus. The multilayer model may possibly include different annotation layers like part-of-speech, morphological features, phrase structure, and dependency structure, etc. A compact POS tagset based on universal pos tags is considered for the POS annotations layer. Initially, a gold standard of 0.1 million words balanced corpus is created by using manual tagging tools with inter-annotator agreement considerations. A model is also trained with this gold standard corpus. Testing and evaluation show precision, recall, and F-measure accuracies with 97%, 96.7%, and 96.9% respectively.

1. Introduction

Annotated corpus is an important language resource used in theoretical and computational linguistics to reveal the deep linguistic structures and capture the computational properties of a natural language. Modern language technologies use these insights to develop high performance software systems with natural language processing and understanding capabilities [1]. Being under resourced language, annotated corpus resources for Sindhi are rarely available. This work presents an initiative of annotated corpus resources development project for Sindhi. Main objective is to lay down the foundations of multipurpose annotated corpus development model. A corpus development model with possibility of multiple annotation layers is presented. The proposed model is based on James Pustejovsky & Amber Stubbs model [2] with some changes. Initially this model is used to develop part-of-speech (POS) tagged corpus of Sindhi. Subset of an existing Sindhi corpus [3] is used for experimental development of pos-tagged corpus. At the outset first layer is annotated with part-of-speech tags. An obligatory POS tagset based on universal POS tags is used for annotations. Webanno [4] was initially used for manual annotations to create a gold standard for machine learning. Later on, Stanford tagger [5] was used for machine learning and automatic pos tagging. Gold standard is incrementally developed by automatic tagging and manual tweaking of wrongly tagged words in different sub-sets of corpus under consideration.

subsequent sections discuss the existing work, proposed multilayer annotation model, development of pos-tagged corpus, results, future work, and conclusion.

2. Existing Work

Only few corpus development studies for Sindhi are there which include Rahman (2010) [3], Mazhar, et., al. [6], and Syed & Bhatti (2018) [7]. In first study Rahman (2010) presented Sindhi corpus construction project. The corpus collection cleaning and organization process is discussed with plain text corpus analysis results including unigram, bi-gram, and tri-gram frequencies. This work lacks the annotation model and its implementation. In second study Mazhar, et. al., (2019) presented the development and analysis of Sindhi corpus for feature attributes and sentiment analysis. This corpus is made available as a dataset with around seven thousand (7000) entries annotated with universal POS tagset. Entries mostly include discrete sentences without any continuity of topic. Dataset includes universal pos-tags, with morphological (number, gender, and person) information, negative, positive sentiment and polarity values. The third study Syed & Bhatti (2018) presented an XML based document structure for development of Sindhi corpus. However, only document structure model is presented, and linguistic annotations are not discussed in this study.

Other related studies are mostly about pos tagger development and training, and development of tagsets for Sindhi Language [8]. [9] and [10] present POS taggers with reasonable accuracy results, however, there is no publicly available annotated corpus except [6] discussed above.

3. Annotated Corpus Development Model

As discussed above, particularly for this corpus development project a subset of an existing plain text corpus [3] is selected for experiments and final annotations. However, the overall corpus development model is shown in Figure 1. Various phases of corpus development process are summarized in the figure. Guidelines include the necessary documentation regarding what annotators need to know about the corpus and its overall design including the corpus subset selection criteria, annotations, and annotation process guidelines. Selected corpus segments, and tagset alongwith guidelines are given to annotators for manual annotations. Different phases of the presented model are discussed in subsequent sections.

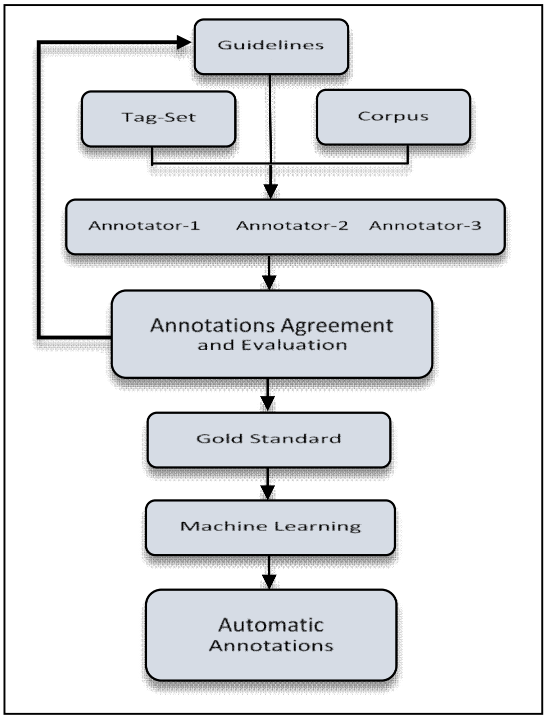


Figure 3. Annotated Corpus Development Model

3.1 Annotation Structure & Representation

The annotation model is designed as a multilayer model where each annotation layer is independent of other layers. This model is inspired by Stand-off annotation by Character Location [11]. This not only solves the white-space tokenization problems but allows simultaneously different layers on same text token/entity with possibility of links between them. Table 1 shows a three-layer sample of layered annotation model with part-of-speech tag, morphological feature tag, and syntactic function tag layers.

Table 6. Layered Annotations

Text	ڪيو	پروسو	تي	مون	چوڪريءَ
POS Tags :	VERB	NN	ADPP	PRON	NN
Morph Tags:	SMPAST	SMNOM	OBL	SGOBL	SFNOM
FUNC-Tags:	VC	NP-POF	PP-OBL		NP-SUB

XML representation of above model are as given below:

```
<TEXT> چوڪريءَ مون تي پروسو ڪيو</TEXT>
<POSTAGS>
 <NN id="N0" start="1" end="7" text="چوڪريءَ" />
 <PRON id="P0" start="9" end="11" text="مون" />
 <ADPP id="A0" start="13" end="14" text="تي" />
 <NN id="N1" start="16" end="21" text="پروسو"/>
 <VERB id="V0" start="23" end="25" text="ڪيو"/>
</POSTAGS>
<MORPHTAGS>
 <SFNOM id="SFO0" start="1" end="7" text="چوڪريءَ"/>
 <SOBL id="SO0" start="9" end="11" text="مون" />
 <OBL id="O0" start="13" end="14" text="تي" />
 <SMNOM id="SMN0" start="16" end="21" text="پروسو" />
 <SMPAST id="SMP0" start="23" end="25" text="ڪيو" />
</MORPHTAGS>
<FUNCTIONALTAGS>
 <NPSUB id="NS0" start="1" end="7" text="چوڪريءَ"/>
 <PPOBL id="PO0" start="9" end="14" text="مون تي"/>
 <NPPOF id="NPF0" start="16" end="21" text="پروسو"/>
 <VC id="VC0" start="23" end="25" text="ڪيو" />
</FUNCTIONALTAGS>
```

Layers (<POSTAGS>, <MORPHTAGS>, <FUNCTIONALTAGS>) contain tags of different categories. For example, <POSTAGS> layer contains NN (Common Noun), PRON (Pronoun), ADPP (Postposition), and VERB tags. Multiple tags within same category have unique id attributes followed by starting and ending position of a token being annotated in the text. It can be seen that multiple layers can mark same location (token) with different tags without disturbing each other. For example, in case of token "چوڪريءَ" ("girl" a common noun with singular, feminine, nominative features) pos tag layer marks it as a common noun tag "NN", morphtags layer marks it with singular feminine and nominative features (SFNOM), and functional tags layer marks the same token as noun phrase subject (NPSUB) function. Overlapping can also be observed where multiple tags of one layer are part of single tag of another layer. This can be seen in functional tags layer where PPOBL (Postpositional Oblique Phrase) spans over the start position 9 to ending position 14 marking single token at functional layer, whereas other layers have two different tags within the same span.

3.2 Tag-Set Considerations

Sindhi has rich morphological constructions as compared to its neighboring languages. Along-with various sub-classes of different parts of speech morphological feature include number, gender, and case in nouns. Morphology also includes rich pronominal suffixation system with nouns, verbs, postpositions, and adverbs. Verbs also have complex morphological causative system. To avoid extra granularity levels initial experimental design of tag-set includes only major parts of speech categories. Morphological features are considered as a separate layer and are not discussed in this paper. POS tagset considered for tagging is based on Universal POS tags [12] and is shown in Table 2.

Table 7. Obligatory Tagset Based on Universal POS Tags

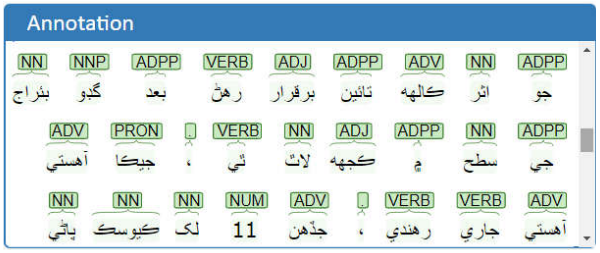
S.No.	POS	POS-Tag
1.	Common Noun	NN
2.	Proper Noun	NNP
3.	Pronoun	PRON
4.	Adjective	ADJ
5.	Adverb	ADV
6.	Preposition	ADP
7.	Postposition	ADPP
8.	Conjunction	CONJ
9.	Interjection	INTJ
10.	Particle	PRT
11.	Negation	NEG
12.	Punctuation	.
13.	Number	NUM
14.	Other Symbols / Unknown	X

3.3 Corpus Selection for Annotations

Two sections (representing two different genres of text) of existing corpus [3] are selected for annotations. Selected corpus sections include news and folk stories Reason behind the selection of these two genres is that news section contains written language with well-formed sentences and folk stories contain vocabulary used by common people in everyday life. Together these two genres represent the Sindhi language of everyday use. 0.1 million words corpus from these two genres (approximately half from each genre) is annotated and used as gold standard for machine learning to automate the pos-tagging process.

3.4 POS Tagging Process

As discussed earlier that selected corpus is annotated with parts of speech tags. Three different annotators were given segments of text for manual POS tagging. WebAnno [4] tool was used for manual POS tagging. Figure 2 shows the snapshot of pos tagging screen in WebAnno.



جو ADPP اثر NN ڪالهه ADV تائين ADPP برقرار ADJ رهڻ VERB بعد ADPP گدو NNP بئراج NN

جي ADPP سطح NN ۾ ADPP ڪجهه ADJ لاٿ NN ٿي VERB ، . جيڪا PRON آهستي ADV

آهستي ADV جاري VERB رهندي VERB ، . جڏهن ADV 11 NUM لک NN ڪيوسڪ NN پاڻي NN

Figure 2. Screenshot of Webanno Tagging Window

Manually annotated segments were then discussed among three annotators to sort out the differences in annotations. These three agreed upon tagged segments were finally combined to have an initial gold standard for machine learning. Stanford pos-tagger was trained on this data and training model was used to tag text segments automatically. These automatically tagged segments were again given to annotators for review and corrections. Correct segments were incrementally added to gold standard. This process is shown in Figure 3. During this process the usability of compact POS-

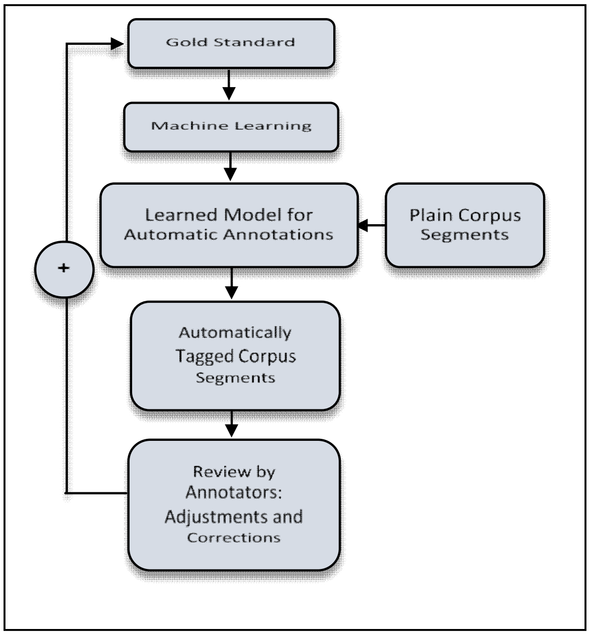

Gold Standard
Machine Learning
Learned Model for Automatic Annotations
Plain Corpus Segments
+
Automatically Tagged Corpus Segments
Review by Annotators: Adjustments and Corrections

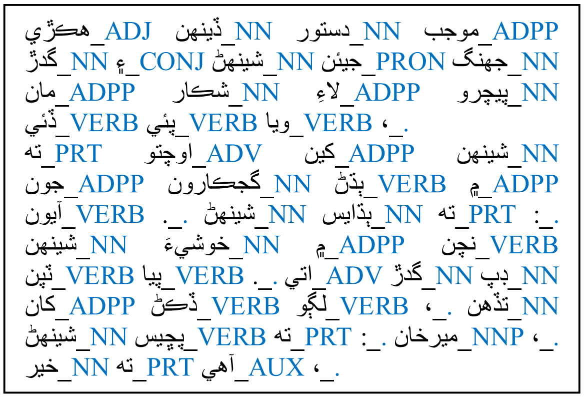
Figure 3. Gold Standard: Incremental Development Process

Tagset being considered was also discussed and evaluated by annotators.

4. Discussion, Results, and Evaluation

Selected subsets of Sindhi corpus include newspaper and folk stories genres. The major reason behind selecting these two genres was their representativeness of language. During corpus analysis it was found that newspaper corpus represents well formed written language which does not necessarily include the spoken language flavor. Folk stories on the other hand are transcriptions of stories narrated by folk storytellers and include rich linguistic features of language used in everyday life. It was found that most interesting linguistic features including the pronominal suffixes, causatives, and inflectional variations were more frequent in folk stories. In contrast newspaper corpus rarely included those features and is mostly comprised of formal written language.

As discussed earlier, internal structure of developed corpus is represented as XML based standoff annotation by character location. This notation is LAF (Linguistic Annotation Framework an ISO standard) [13] compliant. Despite of internal XML based representation, the annotated corpus is easily representable by using common inline tagged notation format. Screenshot of an automatic annotation result generated by Stanford Tagger is shown in Figure 4. It may be noted that this output shows the POS layer in inline format where "_" underscore is used as a tag separator.



```
هڪڙي_ADJ ڏينهن_NN دستور_NN موجب_ADPP
گدڙ_NN ۽_CONJ شينهڻ_NN جيئن_PRON جهنگ_NN
مان_ADPP شڪار_NN لاءِ_ADPP پيچرو_NN
ڏئي_VERB پئي_VERB ويا_VERB ،_.
ته_PRT اوچتو_ADV ڪين_ADPP شينهن_NN
جون_ADPP گجڪارون_NN ٻڌڻ_VERB ۾_ADPP
آيون_VERB ._. شينهڻ_NN ٻڌايس_NN ته_PRT :_.
شينهن_NN خوشيءَ_NN ۾_ADPP نچن_VERB
ٽپن_VERB پيا_VERB ._. اتي_ADV گدڙ_NN دپ_NN
ڪان_ADPP ڏڪڻ_VERB لڳو_VERB ،_. تڏهن_NN
شينهڻ_NN پڇيس_VERB ته_PRT :_. ميرخان_NNP ،_.
خير_NN ته_PRT آهي_AUX ،_.
```

Figure 4. Output of Trained Stanford POS Tagger

By using incremental approach shown in Figure 3 and discussed in section 3.4, 0.1-million-word corpus is tagged and verified as a gold standard. Model trained by using this gold standard is used for automatic POS tagging. Sample output of such tagging is shown in Figure 4. Tagger produces results with 97.0% and 96.7% precision and recall respectively. The given F-measure results are 96.9%. Reasonable accuracy is achieved by trained POS tagger. Most of the error patterns are with the tokens where same token form has more than one POS tags. For example, token "بند" can either be a common NN (stanza) or a VERB (close). Few errors are due to probabilities of noun clusters where inner proper noun NNP or adjective ADJ is tagged as common noun NN. For example, in cluster "صورتحال مڪمل طور" the inner token "مڪمل" is tagged as common noun NN instead of adjective ADJ due to higher probability of three common noun clusters in training data.

5. Conclusion and Future Work

Linguistic resources for Sindhi are rarely available, this work provide the basis for annotated Sindhi corpus resources development with different kinds of annotations. As an experiment compact version of POS tags based on Universal POS tags is used to annotate the selected segments of existing pre-processed and cleaned corpus. First the usability of the compact POS tagset was analyzed and annotators did not find any major problem while annotating the corpus using compact POS tagset. Second, the tagged corpus was used to develop gold standard for machine learning this helped the annotators to speed up the annotation process. Gold standard is evaluated by using machine learning model of Stanford POS tagger and results show reasonable precision and recall accuracy between 96 – 97%. These experimental results are encouraging, and the corpus development is being extended to other layers incrementally. Corpus distribution model is also being worked out to share the corpus resources. This will help to build more robust computational resources and models for Sindhi language processing. We also plan to use this model to develop annotated corpus resources for other Pakistani languages.

9. References

[1] P., James, and A., Stubbs. "*Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*." O'Reilly Media, Inc.", 2017. pp 1 – 23.

[2] P., James, and A., Stubbs. "*Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*." O'Reilly Media, Inc.", 2017. pp 106.

[3] M. URahman ., Towards Sindhi Corpus Construction, *Linguistics and Literature Review* 1(1): UMT. 2015., 39- 48.

[4] S. M., Yimam, I. Gurevych,., R. E., de Castilho, & C. Biemann, "WebAnno: A flexible, web-based and visually supported system for distributed annotations." in *Proc. of the*

51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations August 2013, pp. 1-6.

[5] C., Manning, M., Surdeanu, J., Bauer, J., Finkel, S., Bethard, & D., McClosky. “The Stanford CoreNLP natural language processing toolkit.” In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* June, 2014. pp. 55-60.

[6] M. A. Dootio, & A. I. Wagan. “*Unicode-8 based linguistics data set of annotated Sindhi text*”. Elsevier Data in brief, *19*, 2018. pp.1504-1514.

[7] Z. Bhatti, and M. Shah. "Sindhi Text Corpus using XML and Custom Tags" *Sukkur IBA Journal of Computing and Mathematical Sciences* 2.2, 2018. pp.30-37.

[8] M., URahman. “Developing a Part of Speech Tagset for Sindhi”. In proc. Of the Conference on Language and Technology UET Lahore. 2012.

[9] J. A., Mahar., and G. Q., Memon. "Probabilistic Analysis of Sindhi Word Prediction using N-Grams." *Australian Journal of Basic and Applied Sciences* 5.5, 2011. pp. 1137-1143.

[10] R., Motlani, H., Lalwani, M., Shrivastava, & D. M. Sharma. “Developing part-of-speech tagger for a resource poor language: Sindhi.” In *Proc. of the 7th Language and Technology Conference LTC 2015, Poznan, Poland*.

[11] P., James, and A., Stubbs. “*Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*.” O'Reilly Media, Inc.", 2017. pp 87 – 103.

[12] S., Petrov, D., Das, R., McDonald. "A universal part-of-speech tagset" arXiv preprint arXiv:1104.2086 2011.

[13] N. Ide, and R. Laurent. "International standard for a linguistic annotation framework." *Natural language engineering* 10.3-4. 2004. pp. 211-225.