Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS

Maitrey Mehta
Kahlert School of Computing
University of Utah
maitrey@cs.utah.edu

Vivek Srikumar
Kahlert School of Computing
University of Utah
svivek@cs.utah.edu

Abstract

Good datasets are a foundation of NLP research, and form the basis for training and evaluating models of language use. While creating datasets, the standard practice is to verify the annotation consistency using a committee of human annotators. This norm assumes that multiple annotators are available, which is not the case for highly specialized tasks or low-resource languages. In this paper, we ask: Can we evaluate the quality of a dataset constructed by a *single* human annotator? To address this question, we propose four weak verifiers to help estimate dataset quality, and outline when each may be employed. We instantiate these strategies for the task of semantic analysis of adpositions in Gujarati, a low-resource language, and show that our weak verifiers concur with a double-annotation study. As an added contribution, we also release the first dataset with semantic annotations in Gujarati along with several model baselines.

1 Introduction

Most NLP research focuses only on a few languages: a small fraction of the about 7,000 languages of the world have datasets or linguistic tools. For example, the Universal Dependencies project (Nivre et al., 2016), perhaps the most linguistically diverse community effort in recent years, only covers about 130 languages. Even widely spoken languages like Gujarati and Hausa—both with about 50 million native speakers (Eberhard et al., 2022), almost equaling the population of England—are considered low-resource in the NLP context. In a "rich-get-richer" effect, these trends can increase disparities across the world's languages.

Progress in NLP for high-resource languages like English has been powered not only by advances in modeling techniques, but also by high quality linguistic datasets like the Penn Treebank (Marcus et al., 1994), PropBank (Palmer et al., 2005; Pradhan et al., 2022), OntoNotes (Hovy et al., 2006), and TimeBank (Pustejovsky et al., 2003). Making and tracking progress in low-resource languages requires building similarly large high quality datasets. But what defines a good dataset? The standard way to measure the quality of a manually annotated dataset involves computing an inter-annotator agreement metric such as the ubiquitous kappa score (Artstein and Poesio, 2008).[1]

The very notion of inter-annotator agreement hinges on the availability of at least two annotators. But multiple annotators might not be available, e.g., for a low-resource language. In this paper, we ask: *Can we verify annotation quality when only a single expert annotator is available?* To address this open question, we observe that evaluating the quality of annotated data involves measuring the compatibility between labels proposed by one annotator with a second source of information about the labels. For inter-annotator agreement, the second source is another human trained on the same task. We take the position that for low-resource languages where multiple annotators are unavailable, we can relax the requirement that the second source needs to be a human expert in the same language. To illustrate this position, we propose four weak annotation verifiers and outline the scenarios in which they can be helpful.

We evaluate the efficacy of our verifiers via a semantic annotation effort in Gujarati, a low-resource Indic language. We consider the task of supersense disambiguation of adpositions, which are known to be extremely polysemous. We use an inventory of adpositional supersenses (SNACS, Schneider et al., 2015, 2018) to construct an annotated corpus of a Gujarati version of the book *The Little Prince*. Experiments with this data indicate that our verifiers can be successfully employed in scenarios where

[1] Annotation quality assessment is multi-faceted. Here, we focus only on one such facet, i.e, human agreement and subsequent references to 'dataset or annotation quality' refer to this aspect alone. See §7 for a detailed discussion.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10941–10958
July 9-14, 2023 ©2023 Association for Computational Linguistics

multiple annotators are unavailable.

In summary, the contributions of this work are:

1. We introduce the notion of weak verification of singly annotated corpora and propose four weak verification methods.

2. We evaluate these methods with a dataset annotated by a single annotator in a low-resource language—Gujarati. We show that these methods concur with a double-annotation study performed on the dataset.

3. We release a new adposition supersense dataset for Gujarati based on the SNACS formalism. Notably, this is the first instance of semantic annotation dataset in Gujarati.[2]

2 The Problem of the Single Annotator

Dataset creation in NLP typically involves multiple annotators (experts or crowd workers) labeling a corpus using the task definition and annotation guidelines, followed by a manual or heuristic adjudication step to construct the aggregated ground truth. Datasets form the backbone of computational linguistics and NLP research; ensuring their quality is of paramount importance. Their quality is commonly measured using annotator agreement, and metrics such as Cohen's kappa (Cohen, 1960) reflect consensus (Artstein and Poesio, 2008). A good inter-annotator agreement (IAA) score—over 0.6, per Landis and Koch (1977)—implies a better agreed-upon dataset, whereas a poor one may indicate gaps in the task definition or annotation guidelines. Interesting insights can be drawn by viewing IAA scores alongside model performances. For instance, a dataset with high human consensus and a poor model score suggests a task seemingly simple for humans like common sense reasoning, but difficult for our models (Talmor et al., 2019).

However, human agreement is undefined when we have only one annotator. This could happen when: (1) The task requires specialized expertise, like biomedical named entity tagging (Sazzed, 2022), or, (2) The language does not have readily accessible NLP expertise (Hedderich et al., 2021), such as the Universal Dependencies annotation for the K'iche' (Tyers and Henderson, 2021), and Breton (Tyers and Ravishankar, 2018) languages. In this paper, we study the question of evaluating annotation quality of such singly annotated datasets.

The Principle. Measuring agreement requires two separate sources of annotation, which we will call the primary and secondary sources. In a multiple annotator setup, both are human. When annotation by multiple experts is not possible, we should consider other available resources. In this work, we suggest several resources that can serve as the secondary annotation source. These can be in the form of pre-trained contextualized embeddings, parallel corpora, human expertise in a cognate language, or native speakers of the target language who are not linguistically inclined. In §4, we propose verifiers that use these resources as secondary annotation sources, and evaluate their effectiveness in §5.

3 Gujarati SNACS: A Case Study

This section introduces a new semantic annotation dataset that will serve as a testbed for our verifiers.

3.1 Background

Adpositions (pre-, post-, and inpositions) and case markers are ubiquitous grammatical components that bear diverse semantic relations and are extremely polysemous (Litkowski and Hargraves, 2006; Müller et al., 2010, and others). Schneider et al. (2015) categorized their semantic behavior into coarse-grained categories called **supersenses**.

Hwang et al. (2017) argued that a single supersense label is insufficient to capture the semantic nuances of adpositional usage. They theorized the idea of *construal*, where adpositions are labeled for: a) their meaning in the broader scene, i.e, **scene role**, and b) the meaning coded by the adposition alone, i.e., **function**. Schneider et al. (2018) defined a hierarchy of fifty supersenses called SNACS (Semantic Network of Adposition and Case Supersenses) and annotated a corpus of English prepositions with construals. SNACS has since been extended to multiple languages; annotated corpora exist in Korean (Hwang et al., 2020), Hindi (Arora et al., 2022), among other languages.

In this work, we extend the SNACS project to Gujarati, an Indic language spoken in western India, with about 56 million L1 speakers (Eberhard et al., 2022). Yet, in NLP research, it remains impoverished. Gujarati grammars (Tisdall, 1892; Doctor, 2004) discuss the syntactic usage and diversity of Gujarati adpositions and case markers but their semantic versatility is hitherto unstudied. Gujarati is closely related to its somewhat higher resource

[2] The dataset, and associated code, is available at: https://github.com/utahnlp/weak-verifiers.

cousin Hindi, especially in adposition usage.[3]

3.2 Dataset and Annotation

A bilingual speaker of Gujarati and Hindi annotated all adpositions in *Nānakado Rājakumār*, a Gujarati translation of the novella *Le Petit Prince* (The Little Prince, de Saint-Exupéry, 1943) following its use for SNACS annotation in other languages. Since Gujarati has similar adpositional usages as Hindi, the annotator followed the Hindi-Urdu guidelines v1.0 (Arora et al., 2021), while referring to the English guidelines v2.5 (Schneider et al., 2017) for definitions.

We show some examples below with the annotations for the adposition highlighted in **bold**. In example (1-a), the ergative marker conveys the agency of the action of giving to Sam in the phrase. Hence, the marker gets an Agent scene role which is the same as the function since the ergative marker is prototypically used to describe agency. However, in example (1-b)., the locative adposition *par* is used to convey an instrumentative relation between the phone and the action of talking, meriting different scene role and function annotations.

(1) a. Sam-**e** dīdhuṁ
Sam.**ERG** give.PRF
"Sam gave"
Scene Role: Agent
Function: Agent

b. phona **para** vāt karī
phone **LOC** talk do.PRF
"Talked **on** the phone"
Scene Role: Instrument
Function: Locus

Gujarati has certain unique cases which do not exist in Hindi. We list these details and the target selection heuristics in Appendix C.2.

Table 1 shows descriptive dataset statistics. Appendix C.1 contains additional statistics such as the most prevalent supersenses and construals, and adposition-wise label entropies. While model building is not the focus of this work, we present the results of several baseline models in Appendix C.4. Multilingual models (Indic and general) produce F1 scores between 56-69% for scene roles, and 56-75% for functions with gold adpositions.[4]

Chapters	27
Sentences	1488
Tokens	18516
Targets	3765
Scene Role	3765 (47)
Function	3765 (39)
Construals	
Scene Role = Function	2555 (39)
Scene Role ≠ Function	1210 (110)

Table 1: Dataset Statistics. The numbers in the parentheses denote the number of distinct targets/construals.

Inter-annotator agreement study. A second expert annotator, also a native Gujarati speaker, labeled Chapters 4 and 5 (253 targets) giving a Cohen's kappa score of 0.893 for scene roles, and 0.940 for functions. Besides attesting the quality of our new dataset, this IAA study can help validate the verifiers introduced in §4.

4 Weak Verifiers of Annotation Quality

This section presents four weak verifiers that assess annotation quality. We introduce each one by first stating the prerequisite resources that are needed to use it. We refer to the low-resource language of interest as the *target language*. We note that the verifiers are all *weak*: they are not meant to replace a second annotator, but can help gauge dataset quality in the absence of multiple annotators.

4.1 Using Contextualized Representations

Prerequisites. Pre-trained contextualized representations in the target language.

The research program of training contextualized embeddings (e.g., Devlin et al., 2019) using massive amounts of *unlabeled* text now extends to multiple languages (e.g., Conneau et al., 2020). We propose that, besides their use for model building, these embeddings can also help verify annotations.

For this purpose, we use DirectProbe (Zhou and Srikumar, 2021), a heuristic that probes embeddings using their geometric properties. It clusters labeled instances in an embedding space such that each cluster contains examples with the same label. The number of clusters indicates the linear separability of labels in that space: if the number of clusters equals the number of labels, then the labeled points are linearly separable by a classifier.

Given a singly annotated dataset, we can project the annotation targets into an embedding space us-

[3] The distinction between case markers and adpositions in Hindi (Spencer, 2005) applies to Gujarati, but is irrelevant to this work. We refer to both collectively as 'adpositions'.

[4] Note that merely obtaining high predictive accuracy does not imply high annotation quality; the model might rely on dataset artifacts to achieve its performance.

C_{org}	C_{rand}	CRA	Interpretation
Low	High	High	Good Affinity
High	High	Low	Poor Affinity
High	Low	Low	Poor Affinity
Low	Low	Low	Method Unreliable

Table 2: Trends and Interpretation for CRA.

ing a pre-trained model.[5] If the model representation agrees with the annotations, it should dedicate separate convex hulls (i.e., clusters or regions in the space) for most labels, if not all. Conceptually, this property characterizes a form of agreement between the representation and the human annotator. If the representation disagrees with the human, it would place an example within a cluster associated with the "wrong" label, breaking the cluster into sub-clusters. Consequently, the number of clusters would increase as the disagreement increases.

Merely verifying a one-to-one mapping between labels and clusters is insufficient. We need to compare to how random label assignments behave. If examples were randomly labeled, we should obtain a large number of clusters—in the worst case, almost as many as the number of examples.[6]

Two factors determine the affinity of an annotation with an embedding: (1) each label in the annotation should occupy a separate region (i.e., a distinct cluster) in the embedding space, and (2) if the labels are randomly shuffled across examples, the number of clusters should increase. The latter accounts for the possibility of labels being grouped in the embedding space by chance. We define a metric, CONTEXTUAL REPRESENTATION AFFINITY (CRA), that takes both factors into account to assess the chance-corrected affinity (as in the kappa score) between an annotation and an embedding:

$$\text{CRA} = 1 - \frac{C_{org}}{C_{rand}} \tag{1}$$

Here, C_{org} is the number of clusters produced by DIRECTPROBE with an annotated dataset, and C_{rand} is the number of clusters obtained when its labels are randomly shuffled while ensuring that the label distribution is conserved.[7]

The design of the CRA metric is inspired by Cohen's kappa, and can be interpreted as quantifying the regularity introduced into the embedded points beyond a random labeling. When labels are grouped into a small number of clusters (i.e., low C_{org}), but random labeling leads to a large number of clusters (i.e., high C_{rand}), then the CRA will be high. This means that the representation agrees and the annotations bear information that goes beyond chance. However, a low CRA score does not guarantee disagreement. With a low CRA score, we need to look at the C_{org} and C_{rand} values. When C_{org} is high, labels occupy overlapping regions of the embedding space and the labeled data has low affinity with the embedding. However, when both counts are low, and close to the number of labels, both label sets occupy distinct regions of the embedding space. In such a case, CRA is not conclusive. Table 2 summarizes these four scenarios.

4.2 Using Cognate Language Annotation

Prerequisites. Annotated corpus in a cognate language, Bilingual or multilingual expert annotator for the target language.

Some annotation projects (e.g., our case study) involve parallel annotated corpora. *Existing* annotation in a cognate language can be used by manually or automatically aligning sentences and comparing annotations (manual alignments require a bilingual annotator). We can then measure agreement of labels assigned to the aligned components. A similar approach had been undertaken by Daza and Frank (2020) for semantic role labeling. They use mBERT (Devlin et al., 2019) embeddings to align predicate and arguments from English to various other target languages, and project gold annotations in English to the target languages.

Two points are worth noting. First, the cognate language need not be a high-resource language. Second, this approach is inapplicable when the labels are not preserved across translations. For example, for the task of grammatical gender classification, we can use this verification strategy only if both languages follow the same gender classes, and carry the same gender for translations of nouns.

4.3 Translate and Verify

Prerequisites. Bilingual or multilingual translator between the target and a cognate language, and an expert annotator in the cognate language.

This approach, like the previous one, requires that labels be preserved across the target-cognate translation. However, instead of relying on existing

[5] Note that there is no fine-tuning of the model for the task.

[6] In practice, this worst case is highly unlikely as even a random assignment will exhibit some grouping of labels.

[7] In practice, C_{rand} is averaged over multiple runs.

annotation and alignment tools, it requires an expert annotator in the cognate language. A bilingual speaker is required to translate the text in the target language to the cognate language conserving the intricacies of the task. The annotator can then label the translated corpus and the labels can be compared for agreement.

4.4 Verification Using Non-expert Annotators

Prerequisites. A pool of non-expert annotators in the target language.

Certain tasks, by design, are not amenable for crowd-sourcing due to their complexity. Much work has been dedicated in making the annotation easier by methods like enforcing an annotation curriculum (Lee et al., 2022), and iterative feedback (Nangia et al., 2021), to name a few. He et al. (2015) propose querying annotators for question-answer pairs for the Semantic Role Labeling task which might not be straight-forward for a non-expert. Wein and Schneider (2022) propose a *worker priming* approach where a proxy task primes a crowd worker to a subsequent downstream task for which annotated data is required.

At a high level, verifying with non-expert annotators involves casting the target task into task(s) more favorable for annotation by non-experts (who may possibly be anonymous crowd workers). Naturally, the simplification process would vary from task to task. In §5.4, we provide a concrete instantiation of this idea for our target task.

5 Evaluating the Verification Strategies

In this section, we instantiate the verification strategies from §4 for the Gujarati SNACS annotation task to empirically evaluate them. We show experiments on additional datasets wherever possible.

5.1 Using Contextualized Representations

We instantiated the strategy of using contextual representations (§4.1) with six pre-trained language models: IndicBERT (Kakwani et al., 2020), MuRIL (base & large) (Khanuja et al., 2021), mBERT (Devlin et al., 2019), and XLM-R (base & large) (Conneau et al., 2020). We average the contextual embedding of all tokens (for multi-word adpositions) to obtain adposition embeddings.

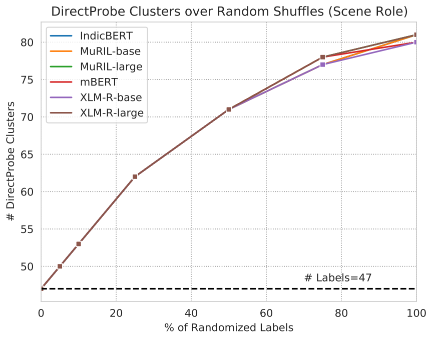


Figure 1: DirectProbe clustering results for Scene Roles over varying randomization.

Since the CRA score is a novel contribution of this work, in addition to presenting the scores for the new Gujarati dataset, we show examples illustrating the different regimes of the score, and also show its high correlation with Cohen's kappa.

Number of DirectProbe clusters. We applied Zhou and Srikumar (2021)'s implementation of DirectProbe[8] to our dataset for the six embeddings. In all cases, and for both scene role and function, the number of clusters C_{org} obtained from the singly annotated dataset is the minimum, namely the number of labels. In other words, *for both tasks, across all embeddings, each label is allocated a separate region of the embedding space.*

Next, to confirm that the embeddings can recognize bad annotations, we shuffled $q\%$ of the labels for $q = \{5, 10, 25, 50, 75, 100\}$. (Note that the number clusters for $q = 100\%$ is C_{rand}.) Recall from Table 2 that if randomized labels do not correspond to an increased number of clusters, we cannot draw any conclusions. Figure 1 shows the trend for scene roles. We average the results over five random runs for each value of q. *We observe that the number of clusters increases with increased randomization of labels for all embeddings.*

Gujarati CRA scores. Table 3 shows the CRA scores for the scene role and function tasks with the various embeddings. We see that all representations are similar in how they handle random label assignments. Consequently, their CRA scores, which measure the affinity of the annotation beyond chance, are in a similar numeric ranges for both tasks. Figure 2 shows the behavior of CRA scores with increasing randomization of labels.

We see that the CRA scores are negatively correlated with the amount of randomization in the labels. In other words, noisier annotations (via

[8] www.github.com/utahnlp/DirectProbe

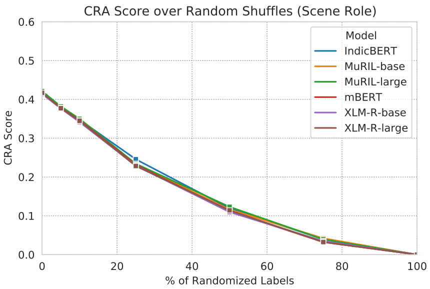


Figure 2: CRA scores for Scene Roles over varying randomization in data. The C_{org} values are the ones as shown in Figure 1 for various randomizations.

Task	**Model**	C_{org}	C_{rand}	**CRA**
Gujarati SNACS Scene Role	IndicBERT	47	81	0.417
	MuRIL$_{\text{large}}$	47	81	0.421
	mBERT	47	80	0.415
	XLM-R$_{\text{large}}$	47	81	0.417
Gujarati SNACS Function	IndicBERT	39	73	0.469
	MuRIL$_{\text{large}}$	39	74	0.470
	mBERT	39	73	0.469
	XLM-R$_{\text{large}}$	39	74	0.470

Table 3: CRA scores for Gujarati SNACS. We use the entire dataset (3765 targets) for this study. C_{rand} values are rounded to the closest integer. Due to space constraints, we moved base variant results to the appendix.

Task	L	C_{org}	C_{rand}	**CRA**
SNLI$_{10k}$	3	6	10	0.375
STREUSLE SR	46	46	48	0.034
STREUSLE Fx	38	38	40	0.050
Estonian UPoS	17	17	120	0.859

Table 4: CONTEXTUAL REPRESENTATION AFFINITY scores for SNLI, English SNACS and Estonian UPoS. C_{rand} values are rounded off to the closest integer. **SR**: Scene Role, **Fx**: Function, **L**: # labels.

randomization) have lower CRA scores, thus validating the definition of the CRA as a verifier.

Appendix A.1 shows the similar behavior of the number of clusters and CRA score across label randomizations for the adposition functions.

CRA behavior. To better understand the numeric ranges of the scores, and to illustrate a failure case, we apply the approach to several existing datasets: a 10k subsample of the SNLI dataset (Bowman et al., 2015), the English SNACS STREUSLE corpus (Schneider et al., 2018), and Estonian EWT Universal Part of Speech (UPoS) dataset (Muischnek et al., 2014). We used the XLM-R$_{\text{large}}$ in all cases. Table 4 shows their scores.

With SNLI, we see that C_{org} is more than the number of labels, namely three, suggesting we have a minor disagreement between the representation and the annotation. However, we also see that the random annotation fares much worse (thrice the number of labels). The CRA score suggests that its affinity to the embeddings *beyond chance* is slightly less than the case of Gujarati SNACS .

On the Estonian UPoS data, C_{org} is equal to the number of labels while the C_{rand} is about seven times more. Hence, this yields a high CRA score.

We observe low CRA scores with the English SNACS datasets. We also see that C_{rand} values are small and close to their respective C_{org}, placing us in the last row of Table 2. The verifier is unsuitable for this embedding-dataset combination. We conjecture that this might be due to a wider spread of the data in the embedding space which allows even a random labeling set to show clustering behavior.

Kappa vs CRA Correlation Analysis. To show that the CRA score behaves like the an agreement score, we conduct an experiment using the TweetNLP data (Gimpel et al., 2011) to show how Cohen's kappa and CRA vary with different amounts of annotation noise. Hovy et al. (2014) supplemented the original labels by crowd sourcing five annotations per instance. We use the majority crowd label for this experiment, and add noise to it by shuffling $q \in \{0, 5, 10, 25, 50, 75, 100\}$ percent of the labels. For each case, we compute the kappa against the original gold labels and also the CRA score using XLM-R$_{\text{large}}$. We compute these scores with five random shuffles for each q. Figure 3 shows a scatter plot between the scores. We observe that as noise increases, both scores decrease, and we have a high Pearson correlation of 0.915 between the two scores. This gives additional validation to the CRA metric as a measure of agreement.

5.2 Using Cognate Language Annotation

To instantiate the verifier in §4.2, we compared our Gujarati annotation with the adjudicated Hindi annotation of *The Little Prince* of Arora et al. (2022) by aligning sentences between the translations. A target is aligned if, in the parallel sentence, the object of the adposition and its governor match. We used chapters 4, 5, 6, 17, and 21 for this task. A bilingual Hindi-Gujarati speaker performed the manual alignment.

Of the 757 adpositions annotated in the selected

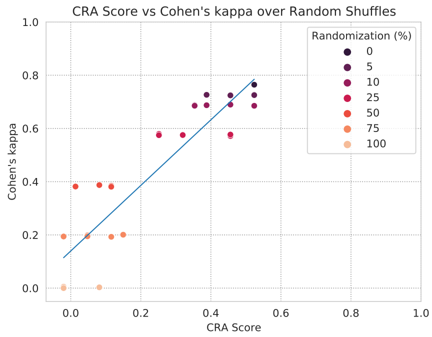


Figure 3: Correlation plot between CRA and Cohen's kappa on TweetNLP dataset. The Pearson correlation between CRA and kappa is 0.915.

Gujarati corpus, 526 tokens could not be aligned to the Hindi corpus as semantically equivalent sentences can be written using a different adposition (i.e. that is not a direct translation), or even without using one. For the remaining 231 aligned tokens, we computed the Cohen's kappa to obtain the agreement between the Hindi and Gujarati annotations. We observe high agreement scores of 0.781 and 0.886 for scene roles and functions respectively.

To verify if the alignment introduced any bias, we compare the kappa scores from the double annotated study for the aligned and the unaligned tokens for chapters 4 and 5. The IAA for the aligned tokens were 0.932 for the scene roles and 0.959 for the functions, while for the unaligned ones they were 0.875 and 0.930 respectively. This suggests negligible bias due to alignment, if any. On the aligned tokens of the same chapters (i.e., chapters 4 and 5), the kappa score between the Hindi and Gujarati aligned annotations is 0.824 and 0.876 for scene roles and functions respectively. We observe that the Hindi alignment recovers over 90% of the kappa scores from the relevant subset of the double annotation study.

As a second evaluation of this verifier, we assessed the Portuguese part-of-speech annotations with respect to Spanish annotations using a 181-token subset of PUD datasets (McDonald et al., 2013) and found a high kappa score of 0.881.

5.3 Translate and Verify

A bilingual Hindi-Gujarati annotator translated chapters 4 and 5 of the Gujarati version of *The Little Prince* into Hindi. The translations were generated such that adpositions were conserved and mapped to their respective counterparts in Hindi. (This is unlike the setting of §5.2 that used an existing translation, and due to which many adpositions were lost in translation.) The translated sentences were subsequently annotated by a Hindi SNACS expert annotator. We ask the Hindi expert to flag any ungrammatical translations, which are avoided. We observe that the targets selected by the Gujarati and Hindi annotators matched 83.0% of the times. That is, the two annotators largely assigned labels to the same tokens (up to translation). We observe a kappa score of 0.802 and 0.837 for scene roles and functions respectively for the tokens identified both in Gujarati and Hindi annotations as targets.

5.4 Verification Using Non-expert Annotators

Our experiments for this strategy (§4.4) focus on the scene role identification task.[9] Our goal is to set up an annotation task that a native speaker of the target language (Gujarati) can perform *without* having to read the annotation guidelines. To do so, we provide the annotator with an instance of a sentence with a highlighted adposition, and ask two questions: (1) Given four sentence choices that use the same adposition, which sentence employs the adposition to convey the relation that is most similar to the one conveyed by the highlighted adposition? (Task 1). (2) Given four supersense definitions for the adposition (attested in the annotated corpus), which one most closely resembles the sense in which the highlighted adposition is used? (Task 2). Appendix B shows the task instructions and example questions. We consider the answers provided to the first question (Task 1) as the priming task for scene role selection (Task 2).

We emphasize that this setup is different from crowd-sourcing of labels: the singly annotated data dictates the choice of the four sentence options and definition choices. The task is closer to annotation verification rather than a fresh round of annotation.

A non-expert native Gujarati speaker performed these annotations. As mentioned earlier, one annotation instance consists of a sentence with a highlighted adposition which serves as the query sentence for the Task 1 and Task 2 questions. To construct the data for this task, we choose one representative sentence from every adposition-supersense pair to act as query sentences.[10] Adpositions that

[9] We consider only scene roles as they are prone to ambiguity and hence more interesting. Functions are more rule-based, and take some amount of orientation with guidelines.

[10] This implies that rare instances are equally weighted as the more frequent ones. This is done for comparison against a diverse set of examples. Consequently, this task would be harder and a lower score would be expected.

have at least four different supersense annotations were considered. We measured agreement between the original annotations and the annotation from the non-expert speaker. We found that the exact agreement for Task 1 and kappa score for Task 2 were 51.4% and 0.547 respectively.[11]

We repeat this experiment with the original Gujarati expert to verify intra-annotator agreement. This shows the internal consistency of an annotator with the same task. To emulate non-expert conditions, no external resources were referred. We observe an accuracy score of 87.2% on Task 1 and a kappa score of 0.867 for Task 2 which shows reasonable consistency of the expert on a harder split and under stricter conditions.

6 Commentary on Results

In this section, we compare our double annotation study with the results from the verifiers.

Using Contextual Representations. As discussed in §5.1, we find favorable number of clusters in the original and random settings for the Gujarati SNACS experiment. Subsequently, we get good CRA scores across all representations indicating high agreement. We argue that obtaining high (and identical) affinity with one representation may be a statistical accident, but obtaining high affinity over multiple representations is unlikely to be a mere coincidence (subject, of course, to the caveat that many representations are pretrained on similar datasets). This lends credence to the hypothesis that the annotations are linguistically meaningful.

In CRA, we propose a new verification score. Note that this score is not intended to replace metrics like Cohen's kappa, nor is it a panacea for the difficulties of single-annotator settings. Instead, it provides a new dimension for annotation verification. To validate the metric itself, we show results on several datasets and present an additional study between kappa and CRA. Admittedly, it requires monolingual or multilingual embedding. If a language is not yet represented in such embeddings, one can augment existing pre-trained embeddings via continued pre-training with a small amount of unlabeled text (Ebrahimi and Kann, 2021).

Using Cognate Language Annotation & Translate and Verify. Both methods rely on two sets of expert annotations, albeit, across two cognate languages. We see high kappa scores on both these experiments, comparable to the double-annotation scores. These methods can be useful for high resource-low resource language pairs.

Verification Using Non-expert Annotators. Our results suggest that this method underperforms with respect to double annotation. This is expected because we use non-experts. The method also relies on the ability to simplify a task for non-experts, which might not be straightforward. However, we note that our observed kappa scores still fall in the higher end of the 'moderate category' agreement according to Landis and Koch (1977).

7 Discussion and Related Work

Dataset quality. Datasets have been be evaluated along various dimensions, e.g., annotation artifacts (e.g., Gururangan et al., 2018) and demographic biases (Bordia and Bowman, 2019; Barikeri et al., 2021). These efforts often use external resources to define evaluation techniques. We can draw parallels between such work and ours in that we use external resources to validate a dimension of dataset quality, i.e., human agreement. Swayamdipta et al. (2020) offer a relevant viewpoint by using training dynamics to analyze the difficulty of learning individual examples in a dataset.

Verifiers for Prescriptive Annotation. Rottger et al. (2022) point out two annotation paradigms: (i) *prescriptive:* where annotators adhere to a prescribed set of guidelines, and (ii) *descriptive:* which encourages annotators to follow their subjective opinions for annotation. Both IAA and our verifiers are applicable only in the 'prescriptive' scenario.

Labels in a Low-Resource Scenario. The problem of data collection in low-resource settings is not new. Recently, Hedderich et al. (2021) presented a survey on low-resource data collection and discuss a range of methods ranging from data augmentation (Feng et al., 2021), distant supervision (Mintz et al., 2009; Ratner et al., 2017), to cross-lingual annotation projection (Plank and Agić, 2018). Active learning (Settles, 2009) can also be useful for efficient annotation. Such methods are meant to facilitate annotation. While these methods are important, we seek to answer the question: *How does one measure annotation quality when you have exactly one expert annotator?*

[11] Non-expert who did *not* see the SNACS guidelines attempted these tasks. A relatively lower score is to be expected.

Gujarati Adpositions. Only a small body of work exists on Gujarati postpositions. Tessitori (1913) trace the origins of the dative and genitive markers in Gujarati to Old Western Rajasthani. They also attempt to explain the use of prototypical dative markers in agentive roles. Turner (1914) argues against the theory. Tisdall (1892) and Doctor (2004) provide an extensive list of postpositions along with conditions for valid syntactic usage.

Our methods should not be seen as replacing multi-annotator efforts, although in such setups, our methods can act as supplementary verifiers.

8 Conclusions

An inter-annotator agreement study is an essential checklist item before the release of human curated datasets. But inter-annotator agreement cannot be computed when we only have one annotator. We address the open question of verifying singly annotated datasets with a new paradigm of exploiting ancillary resources that can serve as weak surrogates for other annotators. The intuition is that each such verifier provides hints about the dataset quality, and their *cumulative* success is more likely to point to a good dataset.

We presented four verification strategies that operate in this paradigm, and a new agreement metric (CONTEXTUAL REPRESENTATION AFFINITY). We also created the first semantically focused dataset of adpositions in Gujarati, a low-resource language, in the single-annotator setting. We showed that our verification strategies, when instantiated to the new dataset, are promising and concur with a traditional double-annotation study.

9 Limitations

We do not propose a solution for extremely low-resource languages, where neither unlabeled text for building language models, nor native speakers are readily available. Examples of such languages include Muscogee, with about 4500 native speakers, and 325 articles in the Muscogee language Wikipedia, and Arapaho with about 1000 speakers and no Wikipedia articles. In such cases, finding even a single expert annotator might be difficult. The development of resources in such languages, however, do not necessarily rest purely on technological factors.

On the technical side, DIRECTPROBE relies on the fact that a representation can be generated for the instance to be annotated. However, obtaining an effective representation for structured annotations (e.g., frames, dialogue states, tables, etc) is non-trivial. While this is a problem, this is orthogonal to our contributions.

10 Acknowledgements

This work would not have been possible without the selfless volunteer work by Mayank Jobanputra and Aryaman Arora. We extend our sincere thanks for their invaluable time and feedback during the process. We would like to thank members of the Utah NLP and CARMLS groups for their valuable insights, and the reviewers for their helpful feedback. We also acknowledge the support of National Science Foundation Grants #1801446, #1822877, #2007398 and #2217154.

References

Aryaman Arora, Nitin Venkateswaran, and Nathan Schneider. 2021. Hindi-Urdu Adposition and Case Supersenses v1. 0. *arXiv preprint arXiv:2103.01399*.

Aryaman Arora, Nitin Venkateswaran, and Nathan Schneider. 2022. MASALA: Modelling and Analysing the Semantics of Adpositions in Linguistic Annotation of Hindi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5696–5704, Marseille, France. European Language Resources Association.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1941–1955, Online. Association for Computational Linguistics.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Angel Daza and Anette Frank. 2020. X-SRL: A Parallel Cross-Lingual Semantic Role Labeling Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3904–3914, Online. Association for Computational Linguistics.

Antoine de Saint-Exupéry. 1943. *Le Petit Prince [The Little Prince]*. Reynal & Hitchkock (US), Gallimard (FR).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Raimond Doctor. 2004. *A Grammar of Gujarati*, volume 28. Lincom Europa.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World. Twenty-fifth edition.* SIL International, Dallas, Texas.

Abteen Ebrahimi and Katharina Kann. 2021. How to Adapt Your Pretrained Multilingual Model to 1600 Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, Oregon, USA. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-Answer Driven Semantic Role Labeling: Using Natural Language to Annotate Natural Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. Experiments with Crowdsourced Re-annotation of a POS Tagging Data Set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, Baltimore, Maryland. Association for Computational Linguistics.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double Trouble: The Problem of Construal in Semantic Annotation of Adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 178–188, Vancouver, Canada. Association for Computational Linguistics.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean Adposition Semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona Spain (online). Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M.

Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual Representations for Indian Languages. *arXiv preprint arXiv:2103.10730*.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174.

Ji-Ung Lee, Jan-Christoph Klie, and Iryna Gurevych. 2022. Annotation Curricula to Implicitly Train Non-Expert Annotators. *Computational Linguistics*, 48(2):343–373.

Kenneth C. Litkowski and Orin Hargraves. 2006. Coverage and Inheritance in The Preposition Project. In *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant Supervision for Relation Extraction without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, and Dage Särg. 2014. Estonian Dependency Treebank and its Annotation Scheme. In *Proceedings of 13th Workshop on Treebanks and Linguistic Theories (TLT13)*, pages 285–291.

Antje Müller, Olaf Hülscher, Claudia Roch, Katja Keßelmeier, Tobias Stadtfeld, Jan Strunk, and Tibor Kiss. 2010. An Annotation Schema for Preposition Senses in German. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 177–181, Uppsala, Sweden. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Barbara Plank and Željko Agić. 2018. Distant Supervision from Disparate Sources for Low-Resource Part-of-Speech Tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 614–620, Brussels, Belgium. Association for Computational Linguistics.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'gorman, James Gung, Kristin Wright-bettner, and Martha Palmer. 2022. PropBank Comes of Age—Larger, Smarter, and more Diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The Timebank Corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Salim Sazzed. 2022. BanglaBioMed: A Biomedical Named-Entity Annotated Corpus for Bangla (Bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329, Dublin, Ireland. Association for Computational Linguistics.

Nathan Schneider, Jena D Hwang, Archna Bhatia, Vivek Srikumar, Na-Rae Han, Tim O'Gorman, Sarah R Moeller, Omri Abend, Adi Shalev, Austin Blodgett, et al. 2017. Adposition and Case Supersenses v2. 5: Guidelines for English. *arXiv preprint arXiv:1704.02134*.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive Supersense Disambiguation of English Prepositions and Possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A Hierarchy with, of, and for Preposition Supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 112–123, Denver, Colorado, USA. Association for Computational Linguistics.

Burr Settles. 2009. Active learning Literature Survey.

Andrew Spencer. 2005. Case in Hindi. In *Proceedings of the LFG05 Conference*, pages 429–446. CSLI Publications Stanford, CA.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Luigi Pio Tessitori. 1913. XVIII On the Origin of the Dative and Genitive Postpositions in Gujarati and Marwari. *Journal of the Royal Asiatic society*, 45(3):553–567.

WS Tisdall. 1892. *A Simplified Grammar of the Gujarati Language*, volume 22. Sagwan Press.

Ralph Lilley Turner. 1914. The Suffixes-ne and-no in Gujarati. *The Journal of the Royal Asiatic Society of Great Britain and Ireland*, pages 1053–1058.

Francis Tyers and Robert Henderson. 2021. A Corpus of K'iche' Annotated for Morphosyntactic Structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20, Online. Association for Computational Linguistics.

Francis M Tyers and Vinit Ravishankar. 2018. A Prototype Dependency Treebank for Breton. In *Actes de la Conférence TALN. Volume 1 - Articles longs, articles courts de TALN*, pages 197–204, Rennes, France. ATALA.

Shira Wein and Nathan Schneider. 2022. Crowdsourcing Preposition Sense Disambiguation with High Precision via a Priming Task. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 15–22, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2021. DirectProbe: Studying Representations without Classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5070–5083, Online. Association for Computational Linguistics.

A Additional Verifier Experiments

A.1 CRA Scores across Randomization

In Figure 4, we show the change in the number of clusters with increasing randomization. Furthermore, we show that the CRA scores decrease with an increasing amount of noise in Gujarati function annotations in Figure 5.

A.2 Complete CRA Results on Gujarati SNACS

The complete set of results on all models and their variants are given in Table 5.

B Native Speaker Verification Instructions and Examples

We show the screenshots of the instructions in Figure 6 and an example question in Figure 7.

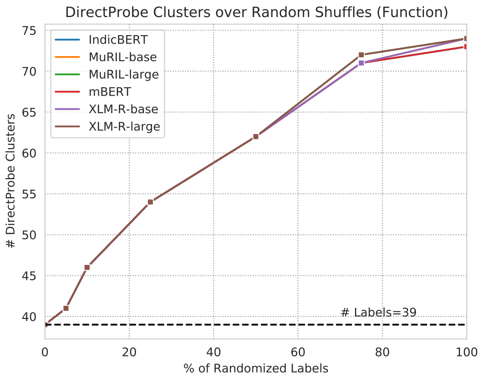


Figure 4: DIRECTPROBE clustering results for Functions over varying randomization.

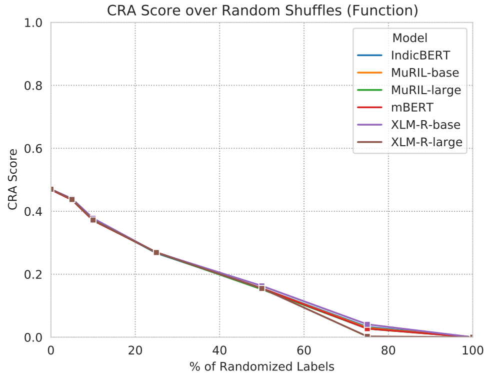


Figure 5: CRA scores for Functions over varying randomization.

Task	Model	C_{org}	C_{rand}	CRA
Gujarati SNACS Scene Role	IndicBERT	47	81	0.417
	MuRIL$_{\text{base}}$	47	81	0.418
	MuRIL$_{\text{large}}$	47	81	0.421
	mBERT	47	80	0.415
	XLM-R$_{\text{base}}$	47	80	0.414
	XLM-R$_{\text{large}}$	47	81	0.417
Gujarati SNACS Function	IndicBERT	39	73	0.469
	MuRIL$_{\text{base}}$	39	74	0.470
	MuRIL$_{\text{large}}$	39	74	0.470
	mBERT	39	73	0.469
	XLM-R$_{\text{base}}$	39	74	0.470
	XLM-R$_{\text{large}}$	39	74	0.470

Table 5: CRA scores for Gujarati SNACS. We use the entire dataset (3765 targets) for this study. C_{rand} values are rounded to the closest integer.

C Gujarati SNACS Statistics, Examples, Assumptions and Baseline Models

C.1 Dataset Statistics

We show the most frequent supersense assignments and the most frequently occurring construals in Table 6. Table 7 shows label entropies of frequently occurring adpositions.

C.2 Notable Examples and Target Assumptions

Here, we discuss certain examples which are present in Gujarati but are not seen in Hindi. Also, we would point to some linguistic issues and their consequent assumptions.

New Prototypical Adpositions. Gujarati houses certain specialized adpositions that are used prototypically for certain semantic relations. Take for instance the adpositions *vade* and *kartāṁ*, which are prototypically used for INSTRUMENT and COMPARISONREF respectively. See example (2). They can optionally be preceeded by the genitive marker/adposition which, in turn, attaches to a non-pronoun complement. See corresponding examples in (3).

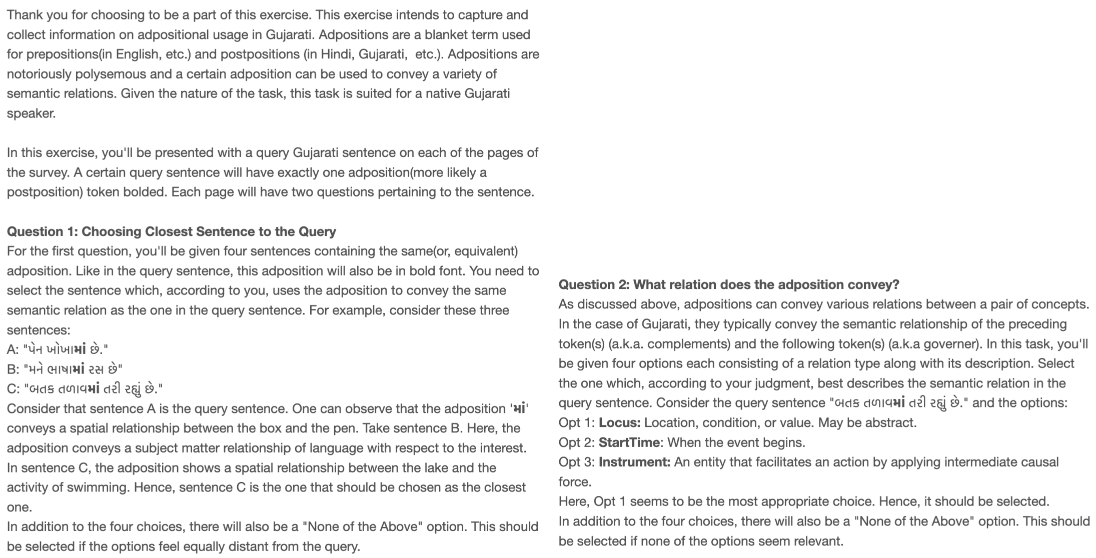
(2) a. cāvī **vade** darvājo kholyo
key INS door open.PRF
"Opened door **with** the key"
b. Sam **kartāṁ** ūṁco Mark
Sam COMP tall Mark
"Mark taller **than** Sam"

(3) a. cāvī-**nī** **vade** darvājo kholyo
key.GEN INS door open.PRF
"Opened door **with** the key"
b. Sam-**nā** **kartāṁ** ūṁco Mark
Sam.GEN COMP tall Mark
"Mark taller **than** Sam"

Hindi, in such cases, uses the *se* adposition which is fairly polysemous and can be used in an Ablative, Comitative, or Instrumentative case (Arora et al., 2022). Corresponding Hindi translations would be *"chābī **se** darvāzā kholnā"* ((2-a) and (3-a)) and *"Sam **se** ūṁcā Mark"* ((2-b) and (3-b)).

Target Selection Assumptions.

1. Certain tokens like *vīshe* (about) and *kāraṇe* (due_to) are used to convey TOPIC and EXPLANATION relation. However, etymologically, these can be broken down into



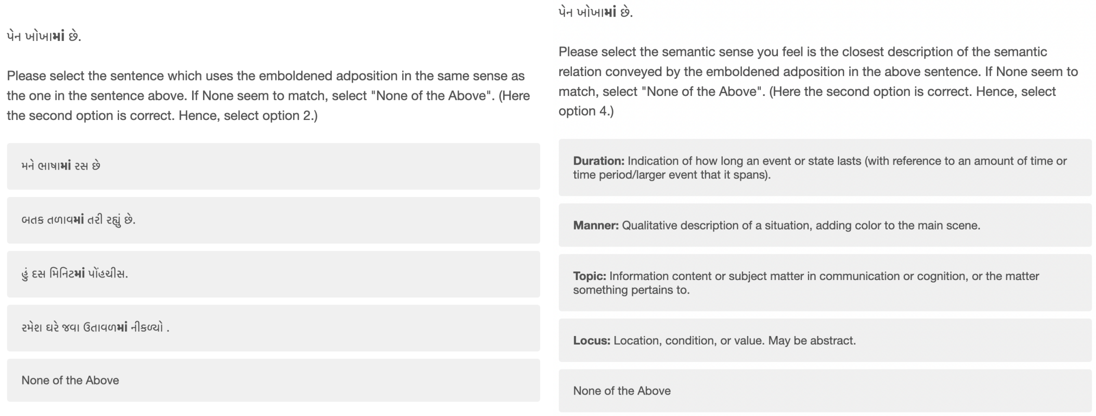
Thank you for choosing to be a part of this exercise. This exercise intends to capture and collect information on adpositional usage in Gujarati. Adpositions are a blanket term used for prepositions(in English, etc.) and postpositions (in Hindi, Gujarati, etc.). Adpositions are notoriously polysemous and a certain adposition can be used to convey a variety of semantic relations. Given the nature of the task, this task is suited for a native Gujarati speaker.

In this exercise, you'll be presented with a query Gujarati sentence on each of the pages of the survey. A certain query sentence will have exactly one adposition(more likely a postposition) token bolded. Each page will have two questions pertaining to the sentence.

Question 1: Choosing Closest Sentence to the Query
For the first question, you'll be given four sentences containing the same(or, equivalent) adposition. Like in the query sentence, this adposition will also be in bold font. You need to select the sentence which, according to you, uses the adposition to convey the same semantic relation as the one in the query sentence. For example, consider these three sentences:
A: "પેન ખોખા**માં** છે."
B: "મને ભાષા**માં** રસ છે"
C: "બતક તળાવ**માં** તરી રહ્યું છે."
Consider that sentence A is the query sentence. One can observe that the adposition '**માં**' conveys a spatial relationship between the box and the pen. Take sentence B. Here, the adposition conveys a subject matter relationship of language with respect to the interest. In sentence C, the adposition shows a spatial relationship between the lake and the activity of swimming. Hence, sentence C is the one that should be chosen as the closest one.
In addition to the four choices, there will also be a "None of the Above" option. This should be selected if the options feel equally distant from the query.

Question 2: What relation does the adposition convey?
As discussed above, adpositions can convey various relations between a pair of concepts. In the case of Gujarati, they typically convey the semantic relationship of the preceding token(s) (a.k.a. complements) and the following token(s) (a.k.a governer). In this task, you'll be given four options each consisting of a relation type along with its description. Select the one which, according to your judgment, best describes the semantic relation in the query sentence. Consider the query sentence "બતક તળાવ**માં** તરી રહ્યું છે." and the options:
Opt 1: **Locus:** Location, condition, or value. May be abstract.
Opt 2: **StartTime**: When the event begins.
Opt 3: **Instrument:** An entity that facilitates an action by applying intermediate causal force.
Here, Opt 1 seems to be the most appropriate choice. Hence, it should be selected.
In addition to the four choices, there will also be a "None of the Above" option. This should be selected if none of the options seem relevant.

Figure 6: Native Speaker Verification Task Instructions. Rough translations of sentences A: *“The pen is **in** the box.”*, B: *“I am interested **in** languages.”*, and C: *“The duck is swimming **in** the lake.”*



પેન ખોખા**માં** છે.

Please select the sentence which uses the emboldened adposition in the same sense as the one in the sentence above. If None seem to match, select "None of the Above". (Here the second option is correct. Hence, select option 2.)

મને ભાષા**માં** રસ છે

બતક તળાવ**માં** તરી રહ્યું છે.

હું દસ મિનિટ**માં** પોંહચીસ.

રમેશ ઘરે જવા ઉતાવળ**માં** નીકળ્યો .

None of the Above

પેન ખોખા**માં** છે.

Please select the semantic sense you feel is the closest description of the semantic relation conveyed by the emboldened adposition in the above sentence. If None seem to match, select "None of the Above". (Here the second option is correct. Hence, select option 4.)

Duration: Indication of how long an event or state lasts (with reference to an amount of time or time period/larger event that it spans).

Manner: Qualitative description of a situation, adding color to the main scene.

Topic: Information content or subject matter in communication or cognition, or the matter something pertains to.

Locus: Location, condition, or value. May be abstract.

None of the Above

Figure 7: Native Speaker Verification Task Practice Question. Rough translations of query sentence: *“The pen is **in** the box.”*, Option 1: *“I am interested **in** languages.”*, Option 2: *“The duck is swimming **in** the lake.”*, Option 3: *“I will reach **in** ten minutes”*, and Option 4: *“Ramesh left for home **in** a hurry.”*

Scene Role	**%**	**Function**	**%**	**SR = Fx**	**%**	**SR ≠ Fx**	**%**
Focus	17.13	Focus	17.13	Focus	17.13	Orig.↝Agent	8.23
Orig.	9.75	Agent	13.39	Theme	6.16	Exp.↝Recipient	6.67
Exp.	8.84	Recipient	10.46	Locus	4.54	Stimulus↝Theme	1.73
Theme	7.20	Theme	8.15	Topic	3.88	Exp.↝Agent	1.49
Locus	5.23	Locus	7.89	Agent	3.61	Orig.↝Gestalt	1.46
Agent	4.14	Gestalt	7.65	Gestalt	3.61	Goal↝Locus	1.33
Topic	3.96	Topic	4.01	Whole	2.87	Soc.Rel↝Gestalt	1.22
Gestalt	3.67	Source	3.27	Recipient	2.84	Char.↝Identity	0.98
(a)		(b)		(c)		(d)	

Table 6: Supersense statistics. Each subfigure shows the most prevalent - (a) scene role supersenses, (b) function supersenses , (c) non-construals, and (d) construals in the data. Exp.-Experiencer, Orig.-Originator, Soc.Rel-SocialRel. Char.-Characteristic

Adposition	Entropy	Counts
nā/nī/nuṁ/nā/nāṁ (GEN)	3.80	856
thī (ABL/INS/COM)	3.34	212
maṁ (LOC-in)	3.11	200
ne (DAT/ACC)	2.37	616
(nī) sāthe ("with")	2.42	35
(nī / ne) māṭe ("for")	2.33	82
e (ERG)	1.96	552
(nā) par (LOC-on)	1.73	121
(nā) viṣhe / vishe ("about")	0.53	33
ja (EMP)	0	332
to (EMP)	0	167
paṇ (EMP)	0	78
ya (EMP)	0	66

Table 7: Entropy of labels by adpositions. Adpositions with a minimum count of 30 were considered

viṣhe = *viṣhay*(subject) + *-e*, and *kāraṇe* = *kāraṇ*(reason) + *-e*. This presents a dilemma about annotating the entire token or just the *-e*. We choose to annotate the entire token given that they exist on the list of postpositions mentioned in Tisdall (1892).

2. Gujarati is also notorious for compound adposition constructions. In certain cases, it is possible to separate out the semantic contribution of the constituent adpositions. Take the instance in example (4). The bolded adposition *taraph-thī* contains two adpositions *taraph* and the ablative *thī*. Hence, a DIRECTION and SOURCE annotation would be appropriate for the respective adpositions. On the other hand, take the sentence in example (5). Here, the compound postposition *-nī_bājumāṁthī* contains three component adpositions - the locatives *-nī_bāju* and *māṁ*, and the ablative *thī*. However, making distinctions between the semantic contributions of each of these adpositions is not staright-forward. Hence, we avoid the complexities of breaking compound postpositions and assign a single set of labels for the entire expression.

(4) darīyā **taraph-thī** āyo
sea **towards.ABL** come.PRF
"came **from** the sea"

(5) darīyā-**nī** **bāju-māṁ-thī** gayo
sea.**GEN** **beside.LOC.ABL** go.PRF
"went **alongside** the sea"

C.3 Double Annotation Disagreements

In this section, we highlight a few examples on which the experts disagreed. Of the 254 adpositions annotated for the double annotation study, only 33 of them had disagreements on either or both scene role and function.

(6) vārtā parīkathā**nī** **māfaq** sharū qarvānuṁ
story fairytale.**GEN COMP** start
"start the story **like/as** a fairytale"

In (6), the annotators disagreed on the scene role between COMPARISONREF and MANNER. One can make an argument that the former is appropriate as the fairytale acts as a reference point. On the other hand, it can be seen as describing a particular way the story has been started and hence MANNER.

(7) ghetā**ne** laine chālyo
sheep.**ACC** take.CONJ walk.PRF "took the sheep and walked"

In (7), the annotators differ with THEME and ANCILLARY annotated for scene roles. If the sheep is seen as the accompanier in the action of walking, it qualifies as an ANCILLARY. If not, the sheep can be seen as the undergoer of the action and hence a THEME.

C.4 Baseline Models

As done for Hindi (Arora et al., 2022), we frame the task as a sequence tagging problem with the supersenses decorated by the BIO scheme. All adpositions are pre-tokenized (separated from the objects) using an existing list of adpositions curated from Gujarati grammar books.[12] This is to ensure that object embeddings are not utilized during classification. The models are trained in an end-to-end fashion without gold adpositions being provided; that is, they have to both identify and label adpositions. All models use token representations from pre-trained contextualized embeddings as inputs to a linear layer followed by a CRF layer that predicts the BIO labels. Additionally, we trained a classifier to predict supersenses given gold adpositions. This linear classifier uses the mean-pooled target adposition embeddings to predict a probability distribution over the label set. Appendix C.5 gives additional details about the experimental setup.

We conduct experiments using six multilingual models, three of which focus on Indic languages.

[12] This list will be released on publication

	Scene Role		**Function**	
Model	**F1**	**F1$_{\text{gold}}$**	**F1**	**F1$_{\text{gold}}$**
Majority	-	18.74	-	23.37
IndicBERT	52.09	56.34	63.37	67.58
MuRIL$_{\text{base}}$	57.27	62.02	69.14	69.68
MuRIL$_{\text{large}}$	**66.23**	**68.66**	**73.17**	**74.80**
mBERT	53.79	56.60	63.05	68.02
XLM-R$_{\text{base}}$	57.08	60.22	69.63	69.53
XLM-R$_{\text{large}}$	62.03	64.05	70.32	73.54

Table 8: Baseline model performance (in %) for Scene Roles and Functions. The column F1 indicates the macro-F1 score for the end-to-end BIO tagger system while the column F1$_{\text{gold}}$ is the F1 score for the token classifier when the gold adpositions are provided. The majority baseline assigns the most frequent label for a target in the training data. If a target is not in the training data, the most frequent label of the corpus is assigned. The models considered are mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), IndicBERT (Kakwani et al., 2020), and MuRIL (Khanuja et al., 2021).

Table 8 shows performance of the CRF models (column F1) and token classifiers (column F1$_{\text{gold}}$). For the former, a correct prediction requires both the span and the label to be correct. All results are averaged across five random seeds.

We observe that MuRIL$_{\text{large}}$ performs the best across all categories. The performances improve by about two points when gold adpositions are known. We also find that the models are near-perfect at identifying adposition spans, with the F1 scores in the 96-97% range for all the models.

C.5 Experimental Setup and Resources for Baseline Models

We use PyTorch v1.10 for the implementation of our baseline models. We use HuggingFace's transformers library for the pre-trained language models. We use the CRF implementation from the pytorchcrf library (`https://pytorch-crf.readthedocs.io/en/stable/`). We choose the best learning rate from {0.0005, 0.0001, 0.00005, 0.00001} based on a small development set. All models are trained till 100 epochs with an early stopping of 5 epochs. The random seeds which we use for our experiments are 11, 20, 42, 1984, and 1996. All computations were conducted on an Nvidia Titan RTX 24 GB GPU.

ACL 2023 Responsible NLP Checklist

A For every submission:

□ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Left blank.

□ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

□ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B □ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C □ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.