Strengthening the WiC: New polysemy dataset in Hindi and lack of cross lingual transfer

Farheen Dairkee[1], **Haim Dubossarsky**[1,2,3]
[1]Queen Mary University of London, [2]University of Cambridge, [3]The Alan Turing Institute
h.dubossarsky@qmul.ac.uk

Abstract

This study addresses the critical issue of Natural Language Processing in low-resource languages such as Hindi, which, despite having substantial number of speakers, is limited in linguistic resources. The paper focuses on Word Sense Disambiguation, a fundamental NLP task that deals with polysemous words. It introduces a novel Hindi WSD dataset in the modern WiC format, enabling the training and testing of contextualized models. The primary contributions of this work lie in testing the efficacy of multilingual models to transfer across languages and hence to handle polysemy in low-resource languages, and in providing insights into the minimum training data required for a viable solution. Experiments compare different contextualized models on the WiC task via transfer learning from English to Hindi. Models purely transferred from English yield poor 55% accuracy, while fine-tuning on Hindi dramatically improves performance to 90% accuracy. This demonstrates the need for language-specific tuning and resources like the introduced Hindi WiC dataset to drive advances in Hindi NLP. The findings offer valuable insights into addressing the NLP needs of widely spoken yet low-resourced languages, shedding light on the problem of transfer learning in these contexts.

Keywords: Less-Resourced/Endangered Languages, Multilinguality, Word Sense Disambiguation, Transfer Learning, Cross Lingual Transfer

1. Introduction

The growing field of Natural Language Processing is marked by its rapid expansion, focused on developing algorithms and methodologies that enable computers to comprehend and manipulate human language. NLP has diverse applications, ranging from machine translation and sentiment analysis to text generation and chatbots that garnered global recognition. However, the creation of such transformative applications necessitates extensive training data. Consequently, model development on a range of NLP tasks becomes more challenging, especially for supervised approaches, without high-quality annotated datasets. State-of-the-art models thrive on vast text corpora, an essential prerequisite for their efficacy.

While accumulating such voluminous data is a formidable task, it is notably more manageable for English. In contrast, the situation becomes more challenging for other languages, giving rise to a predicament. The scarcity of high quality linguistic resources like annotated textual corpora, characterizes the realm of low-resource languages, even if these are spoken by substantial populations. This scarcity acts as a barrier to NLP progress in these languages.

Hindi, spoken by around 580 million individuals globally, serves as a vivid illustration. Despite its prevalence, NLP advancements remain sluggish due to these resource constraints. Yet, utilizing models trained on languages other than Hindi using transfer learning approaches yields suboptimal performance, highlighting the language-specific nuances that demand tailored training data. This problem is not limited to Hindi of course, and many low resource languages suffer from poor performance when transfer is made via English as the source language (Senel et al., 2024). Therefore, addressing the NLP needs of such widely spoken, yet low-resourced languages, such as Hindi, emerges as a pivotal endeavor. By undertaking this pursuit, we want to demonstrate if and how necessary are quality resources in a target language to make sure NLP performance is kept at high level in low-resource but widely-spoken languages.

Within the realm of NLP tasks, Word Sense Disambiguation (WSD) has a prominent role as a core NLP task that has been challenging NLP models for decades (Lesk, 1986; Navigli, 2009; Raganato et al., 2017). WSD targets polysemous words, focusing on detecting their distinct senses according to their specific usage context. Polysemous words introduce a range of complexities due to their contextually versatile nature and inherent ambiguity, and even state-of-the-art NLP systems struggle to capture subtle nuances apparent to human understanding further complicating the matter, as reflected by their performance which is below human level (Raganato et al., 2020).

The role of WSD extends far beyond mere linguistic analysis. Think of the word *bank* for example, its varied senses influence many domains

LREC-COLING 2024, pages 15341–15349
20-25 May, 2024.

like machine translation, information retrieval, text classification, and even Named Entity Recognition (NER) (Chandra and Dwivedi, 2014). WSD's applications stretch even beyond the boundaries of NLP. For instance, it enhances image analysis by refining image captions and object recognition through precise word disambiguation (Raganato et al., 2023). All of this makes it paramount for NLP systems to properly handle polysemous words due to its effects on a range of tasks.

Notably, a word that displays ambiguity in one language might not necessarily exhibit the same multifaceted behavior in another language. These divergent cross-lingual polysemy patterns make it an ideal case to test ”the promise” of transfer learning in solving WSD.

Although the task of WSD has garnered extensive research attention across many languages, the majority of these are Western languages (English, Italian, German, etc.), and low-resource languages are lagging in progress. This disparity becomes increasingly evident when considering languages like Hindi, which enjoys widespread usage but suffers from an under explored NLP landscape. In particular, access to relevant corpora and datasets with sense annotations remains limited in Hindi. This discrepancy assumes significance given the considerable global population that communicates in Hindi. All of the above makes Hindi a prime candidate for in-depth NLP exploration of the necessity of high-quality resources in low-resource languages and an excellent testbed for assessing the ability of standard approaches like transfer learning to address this gap.

This work has three major contributions. It develops a Hindi dataset for WSD, the first of its kind in the modern WiC format (Pilehvar and Camacho-Collados, 2019), enabling the training and testing of contextualized models for polysemy in Hindi. It tests the prevailing assumption that a combination of multilingual models and transfer learning is sufficient to handle polysemy in languages lacking quality resources for model training. Lastly, recognizing that providing quality training data is expensive, we provide a thorough analysis of the minimal amount of training data necessary for a viable solution.

2. Related work

Sinha et al. (2004) and (Yusuf et al., 2022) used a Lesk-like algorithm with Hindi WordNet Bhattacharyya et al. (2008) to disambiguate word senses, and evaluate their models on a small set of 20 items. Using supervised approach, Singh et al. (2014) and Singh and Siddiqui (2015) explored semantic relations through sense definition vectors and context vectors using a sense annotated Hindi corpus (Singh and Siddiqui, 2016). Kumari and Lobiyal (2022) used unsupervised approaches that do not rely on lexical resources, but achieve low performance. All the above studies are problematic with respect to at least their evaluation, scalability (WordNet and annotated-corpus are difficult to scale), or performance.

A recent contribution in WSD is the Word-in-Context (WiC) dataset (Pilehvar and Camacho-Collados, 2019). This dataset was introduced as a benchmark to evaluate the contextual sensitivity of word representations in English, and serves as a standard evaluation tool for various techniques, including contextualized word and sense representations, as well as word sense disambiguation. Its key novelty lies in its ability to transform inventory-based resources, like sense annotated corpus, to an inventory-independent task, which can scale effectively.

The ubiquitous use of contextual word embeddings in the broader realm of NLP gave rise to the development of a multilingual extension of the WiC dataset, known as the XL-WiC dataset (Raganato et al., 2020). This dataset encompasses development and test sets across 12 diverse languages, supplemented by training sets in German, French, and Italian. The inception of the XL-WiC dataset underscores the NLP community's recognition of linguistic diversity as a driving force behind various NLP applications and the advancement of the field. Raganato et al. (2020) demonstrated the competitive performance of multilingual models trained on English, when transferred (i.e., tested) on other languages within the dataset. However, when applied to languages linguistically distant from English, this transfer learning showcased reduced effectiveness. The XL-WiC dataset was meticulously curated using WordNet resources and expert annotators who categorized word pairs into positive and negative contexts. Notably, while this dataset encompasses a multitude of languages, it regrettably excludes Hindi.

Despite the commendable efforts towards inclusively within the realm of NLP, a significant gap remains evident in the representation of Hindi, one of the world's most widely spoken languages. Although strides have been made to address language diversity, the scarcity of Hindi-specific resources, such as the sense-annotated Hindi corpus for constructing datasets, is evident. This research paper serves as an attempt to spotlight the inadequacies of incomplete efforts toward inclusively in NLP, and provides the first Hindi variant of WiC. This new resource will enable to support Hindi-focused NLP research in WSD and polysemy handling, hopefully improving the performance of Hindi in general NLP tasks.

3. Our approach

WSD is a a challenging task even for humans. Given a sentence with a polysemous word, the aim is to focus only on the relevant meaning of that word, and ignore the other sense(s) the word may have in different contexts. For instance, in the example below, we need to consider only the monetary sense of the word *bank*, and ignore all other senses. While the context given in this sentence is sufficient to easily disambiguate the relevant meaning, WSD can quickly become much more complicated. Consider for example the same example, ignoring the underlined last three words, that could be fairly interpreted also as related to the river bank sense.

The woman went to the bank with her dog to deposit money.

In the context of NLP, the task of WSD is often defined as classifying each occurrence of polysemous word in a sentence to its correct sense. For that purpose, a dedicated corpus with annotated senses of polysemous words is used. Clearly, from a ML point of view, this task is rather cumbersome, as it is basically a multi-class classification problem. This problem formulation would require many examples of each class to train a classifier, as well as to know the specific sense inventory beforehand. Unfortunately, these requirements are not feasible for annotated corpora which are typically quite small, and contains a few dozens of example for each sense, due to the huge human effort that is required for annotation (see for example SemCor Miller et al. (1993)). This is especially true for contextualized models, that require large numbers of examples for fine tuning.

To circumvent these difficulties, the problem is reduced into a binary classification task. In this task, sentences with the same polysemous target word are paired together, and the task is to classify if the target word has the same meaning in the two sentences or a different one. This task eliminates the problem of small number of examples per each category (i.e., a specific sense), because the number of potential pairs is much larger, making it suitable for training contextualized models. See the below three sentences with the polysemous word *sage*, as they would have appeared in a sense-annotated corpus, that can either refer to a scholar or a herb, and compare these sentences to how they would appear in a WiC-style dataset in Table 1.

The old sage^1 has spoken.
She acted on a sage^1 advice.
I fry the sage^2 leaves in oil.

The formulation of the WSD task into a binary classification task has became very popular through the WiC dataset and task (Pilehvar and Camacho-Collados, 2019). This task has become the de-facto standard for testing models' performance on WSD task, suitable for the hungry data demands of contextualized models. Nowadays WiC datasets exist for more than twelve languages which allows to test transfer learning scenarios between them.

In this work we develop the first Hindi WiC dataset based on a small sense-annotated corpus in Hindi, according to the protocol used in the original English WiC (Pilehvar and Camacho-Collados, 2019). This allows us to train and test models directly on binary classification of polysemous word pairs in Hindi. We specifically focus on the ability of multilingual models that support Hindi but were trained on WiC in English to transfer to Hindi. We compare their results to the ideal upper-bound case where a model was trained on the full WiC dataset in Hindi in order to demonstrate how critical these resources are to maintain good NLP support for a key task in low resource languages.[1]

4. Creating Hindi WiC

We used the sense annotated Hindi Corpus (SAHC) (Singh and Siddiqui, 2016). SAHC contains 60 unique polysemous nouns, each appearing in dozens of different sentences for each sense in their natural context, and provides sense-annotated instances across different meaning of the target word. The corpus statistics are given in Table 2.

We followed the protocol developed by Pilehvar and Camacho-Collados (2019) to transform the inventory-based SAHC to inventory-independent dataset in the WiC format. For each target word, all possible sentence pairs were computed, and then pruned to ensure an equal number of instances of the same sense (positive pairs) as well as of different senses (negative pairs). The pairs were then randomly divided into training, validation, and test set, with 7000, 3000 and 2000 pairs, respectively.

We ensured there is no overlap of individual contextual sentences across the splits (i.e., the same sentences will never appear in more than one split). However, some target words do overlap between the different sets (i.e., different instances of *bank* can appear in the training and test splits). In our test set, only 67% of the target words overlap with the training set. This 67%-33% split allows testing the in two difficulty conditions, known and novel polysemous words.

[1] The Hindi-WiC dataset is available on `https://github.com/haimdub/HindiWiC.git`

Lang.	Target word	Sentence 1	Sentence 2	Label
English	sage	The old sage has spoken	She acted on a sage advice	T
English	sage	I fry the sage leaves in oil	She was one of the few sages remaining	F
Hindi	तिल	मुझे रोटी पर तिल पसंद है *I like sesame on bread*	तिल के तेल का उपयोग खाना पकाने में किया जाता है *Sesame oil is used in cooking*	T
Hindi	तिल	मेरी नाक पर एक तिल है *I have a mole on the nose*	दुकान से तिल खरीदो *Buy sesame from the shop*	F

Table 1: Positive and negative pairs of the same target words in English and Hindi (*with translation*)

Number of Words	60
Total Number of Senses	151
Total Number of Instances	6248
Avg. Number of Senses/Word	2.52
Avg. Number of Instances/Word	29.6

Table 2: SAHC descriptive statistics

The above steps transform SAHC, an inventory-based WSD dataset to a unique inventory-independent dataset, stripping all specific sense annotated information. As a result, the problem we now have is a binary classification task. Each instance of a target word *w* is provided with two contexts, *c1* and *c2*, and the task is to identify if the occurrences of *w* in *c1* and *c2* correspond to the same meaning or not.

5. Experiments and Results

5.1. Setup and research questions

In this work we study the importance of quality resources in handling polysemy in Hindi. We start by finetuning all models on the full training set of English WiC. We then test the viability of the standard approach in NLP to this problem that uses transfer learning under several experimental conditions, the comparison of which shed light on these research questions.

We define several finetuning conditions with respect to the Hindi Wic, from zero-shot to few-shot to a complete finetuning on the entire Hindi WiC dataset. We also define two testing conditions, KNOWN for target words the models were trained on during finetuning, and NOVEL for unseen polysemous words. We use four models that are pretrained on Hindi, three of which are multilingual to facilitate cross lingual transfer. We use two general-purpose cross-lingual models, a Hindi-focused multilingual model, and a monolingual Hindi model. Comparing the performances of these models across the different experimental conditions yield insights into the influence of language specificity, cross-lingual transfer, and tailored finetuning on WiC task performance in Hindi. These comparisons help us understand the importance of developing quality datasets in the target language, and how to make the best out of them for effective transfer learning.

To address this problem statement, we tested the performance on the Hindi WiC dataset in a few scenarios using the different models. We used multilingual models (mBERT and XLM-R) finetuned on English WiC and tested them on Hindi WiC. This scenario enabled us to test transfer learning under zero-shot conditions which was previously reported to work for other languages (Raganato et al., 2020). We also tested these models under graded few-shot conditions (from 1% to 30%), and also after the models were finetuned on the full training set in Hindi to establish their upper bound performance. We expected these models would show increasing improvements which is associated with the amount of examples in Hindi they were trained on. Lastly, we tested the ability of the models to generalize beyond polysemous words they were trained on.

Critically, the comparison of few-shot condition to the zero-shot condition help us evaluate the merits of using high quality dataset in Hindi, but also the amount of finetuning needed. We further used Hindi-focused multilingual model (MuRIL) as well as monolingual model (HindiBERT) and tested them under similar conditions. Comparing their performance to the commonly-used multilingual models (mBERT and XLM-R) enable us to evaluate the importance of richer language specific information in handling polysemy in Hindi in general. Lastly, the above experimental setup allows us to test generalization ability of these models to unseen polysemous words as a function of the specific model and the amount of finetuning they had.

5.2. Models and training

mBERT is a multilingual variant of the BERT model created by Kenton and Toutanova (2019). It is pre-trained on a diverse collections of 104 languages including Hindi, making it a versatile choice for cross-lingual applications and transfer.

XLM-RoBERTa stands for Cross-lingual Machine learning Robust Bert (Conneau et al., 2020), is pretrained on 100 languages. It was trained on 2.5 TB of filtered CommonCrawl data as opposed to mBERT which was trained on Wikipedia data, and reported to outperform mBERT in a variety of cross-lingual benchmarks like XNLI, MLQA and NER.

HindiBERT is a monolingual model trained on the Hindi CommonCrawl and Wikipedia texts. The model when tested on the tasks BBC Hindi news classification and on Hindi movie reviews / sentiment analysis gave comparable results to mBERT.

MuRIL stands for Multilingual Representations for Indian Languages (Khanuja et al., 2021), and designed to address the linguistic diversity and complexity of Indian languages. It is trained on 17 Indian languages that includes Assamese, Bengali, Gujarati, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu and Urdu, as well as Hindi and English. MuRIL has been trained on approximately 16 billion tokens.

Fine Tuning Training splits of Hindi WiC were used for finetuning (all models were first finetuned on the English WiC). The amount of finetuning depended on the specific experiment, and ranged from 0% (zero shot), to 1%, 5%, 10%, 20%, 30% (few shots), to 100% (full shot). Following Pilehvar and Camacho-Collados (2019) we passed the vectors of the two sentences though a MLP and then though fully connected layer followed by softmax activation which computed probabilities for each label. The model was trained for up to 10 epochs with early stopping based on validation accuracy. The AdamW optimizer was used with a learning rate of 1e-5. To prevent overfitting, different weight decay rates of 0.0 and 0.01 were applied to bias/batch norm parameters and other parameters respectively. The training batch loss and accuracy were tracked and model weights updated each epoch. Validation loss and accuracy were also monitored after each epoch. The weights with the best validation accuracy were saved. The training loop iterated through batches from the Hindi WiC train set. The model computed the logits and loss for each batch. Loss values were backpropagated to update model weights. Validation was performed each epoch on the dev set. The loss and accuracy were computed for each batch and aggregated. This fine-tuning approach allowed efficient training of the models on the English or Hindi WiC dataset depending upon the experiment. The best model weights were retained for final evaluation.

5.3. Results

Table 3 shows the performance on the Hindi WiC test set for zero-shot, monolingual training, and multilingual training (finetuning on English and Hindi WiC's). The results clearly show that mBERT, XLM-R, HindiBert and MuRIL, when trained solely on English WiC (zero-shot), perform almost at chance level, showing no transfer capacities. In contrast, when these models, with the exception of HindiBert, are finetuned on the full Hindi WiC training set (either with or without training on the English WiC), their performance improve dramatically, achieving accuracy scores between 83% and 90%. Lastly, it seems that for some models (mBERT, XLM-R) the best results are obtained for Hindi only finetuning (Mono), as models that were trained in multilingual settings have slightly worse performance, showing the lack of utility of transfer even under full-shot conditions. Overall, this pattern of results demonstrate the lack of transfer learning in both zero-shot and even full-shot settings on the one hand, and the immense benefit Hindi dataset brings to solving WiC in Hindi. HindiBert model shows far worse performance in comparison to the other models, and is excluded from further analysis (but discussed in the next section).

	ZS	Mono	Multi
mBERT	52%	87%	83%
XLM-R	55%	88%	87%
HindiBert	48%	70%	60%
MuRIL	55%	89%	90%

Table 3: Model accuracy for zero-shot transfer (ZS), monolingual training (Mono), and multilingual training (Multi).

As the benefit that our WiC dataset provides is so clear, we wanted to test how extensive the WiC dataset needs to be in order to obtain the above gains. Figure 1 shows the performance of mBERT, XLM-R, and MuRIL in different few-shot settings, using increasing number of training examples (up to full finetuning). The results demonstrate that even relatively modest few-shot conditions lead to significant improvements. At only 10% (700 pairs) of the training examples, there is a big performance gain, and by 30% (2100 pairs) results are close to the maximum achieved with full finetuning. These results further demonstrate the usefulness of quality dataset in the target language to facilitate polysemy resolution. Importantly, it shows that the dataset does not have to be particularly large to lead to substantial gains.

We compared the monolingual and multilingual training conditions of the above analysis to further

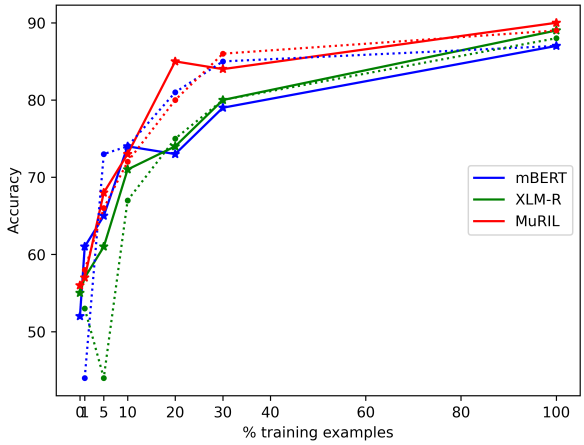

Figure 1: Model accuracy at different training ratios of the Hindi WiC, for the multilingual (solid) and monolingual (dotted) conditions.

investigate the above pattern from Table 3, that finetuning models in multilingual setting slightly hinder their performance. The results obtained in figure 1 provide mixed results. While MuRIL clearly benefit from multilingual training as evidence by its higher accuracy throughout the different finetuning ratios (when compared to monolingual setting), the pattern for mBERT is the exact opposite, and XLM-R performance lies in between, showing no clear preference to either monolingual or multilingual training. MuRIL also stands out as the model that peaks the earliest (for both monolingual and multilingual training) at 20% relative to the other models that peak at 30% . This clear advantage may suggest that language specificity of a model plays a key role as well, as MuRIL was trained on 17 Indian languages. Overall, the results of the few shot analysis above hardly provide any evidence to support that models could benefit from transfer learning in the case of multilingual WiC training in both English and Hindi over Hindi alone.

To conclude our analysis, we wanted to test the abilities of the finetuned models to successfully classify examples of unseen polysemous words (polysemous words that did no appear in any sentence pair in the training set). In other words, the capacity of these models to generalize. Table 4 shows the performance of the three models, at different few-shot ratios, split into two conditions, known words that appear in the training set and comprise of 67% of the examples, and novel words that comprise the remaining 33%. Although known words perform better than novel words throughout the conditions, as expected, the performance levels for novel words is still quite high, and follow closely the pattern of increase performance with increasing few-shot ratios. The performance on novel polysemous words attest to that the models did not rely on pattern matching strategies (which could be argued for the known words (Raganato et al., 2020)), and stratify that even a small set of WiC examples can significantly contribute to the training of WiC models and their generalization on ”out of vocabulary” polysemous words that are not part of the training set.

% Training set	mBERT		XLMR		MuRIL	
	Known	Novel	Known	Novel	Known	Novel
0	46	52	54	49	57	55
1	52	55	49	55	58	56
5	70	62	65	59	71	63
10	68	62	70	64	74	67
20	82	69	77	70	85	77
30	85	71	85	76	89	77
100	95	72	95	78	96	82

Table 4: Accuracy for Known and Novel polysemous words for the 3 models, as a function of ratio of training examples.

6. Discussion

This work provides the first WiC dataset in Hindi, and demonstrates its absolute necessity in training models for polysemy in Hindi. Our experiments provide converging evidence that such a dataset in Hindi is vital to enable high performance in this task, and that reliance on zero-shot transfer, as proposed by prior work (Raganato et al., 2020), simply does not work. This study further identifies the minimum amount of training data required for effective Word Sense Disambiguation in Hindi, offering a viable solution for similar low-resource languages. Overall, the results provide a cautionary counter-evidence to the over reliance on transfer learning in NLP where task-specific resources are lacking. Our results show that harvesting linguistic data from human assessments produces high quality results when time and funds are invested.

Comparing the performance of the different models across various analyses and experimental conditions reveals remarkably similar pattern among the multilingual models. While mBERT, XLM-R, and MuRIL do not exhibit any transfer in the zero-shot condition, they quickly achieve near-optimal performance even with a small set of examples (20%-30%), and demonstrate effective generalization to unseen polysemous words. It appears that the languages on which these models were trained have little influence on their overall performance. The exception is MuRIL, which seems to require slightly fewer training examples (20%) before showing substantial gains and achieving the best performance (90% compared to 83% for mBERT, while XLM-R falls in between). This outcome is not surprising given MuRIL's unique multilingual pretraining setup that uses 17 Indian languages. On the other hand, HindiBERT, does not exhibit good transfer even when trained on the full dataset (multiiIngual condition), probably due to a lack of English support. However, it also fails to demonstrate improved performance in the monolingual condition compared to the multilingual models, where the multilingual models do not hold any advantage.

It is both interesting and relevant to compare our results to those reported in the XL-WiC paper (Raganato et al., 2020). Similar to their work, we developed a new dataset and benchmark and made it available for the research community. However, unlike Raganato et al. (2020) who showed that language models are effective performers in the zero-shot cross-lingual setting, our results show the opposite. It is worth getting to the bottom of this discrepancy, as it may influence the methods developed to support other low resource languages in the future.

Raganato et al. (2020) developed WiC datasets for twelve languages (Bulgarian, Chinese, Croatian, Danish, Dutch, Estonian, Farsi, French, German, Italian, Japanese, and Korean), and similarly to our experiments, tested the performance on these in zero-shot, monolingual, and multilingual settings. Contrary to our finding, their results demonstrate clear zero-shot transfer, albeit modest, in all twelve languages. Their monolingual and multilingual finetuning conditions showed mixed results, with some languages having better performance in monolingual setting and some in multilingual setting. These findings are similar to our own results that show multilingual training may not help, and can even worsen performance compared to monolingual setting.

One potential explanation for the absence of zero-shot transfer in our study could be attributed to the dissimilarity between English as the source language and Hindi as the target language. It was suggested that zero-shot transfer is more effective for linguistically similar languages, and that English and Hindi are simply too dissimilar to transfer effectively (Lauscher et al., 2020; Senel et al., 2024). However, it is worth noting that Hindi and English both belong to the Indo-European language family, which makes them considerably closer when compared to some other language pairs that have been reported to facilitate zero-shot transfer by Raganato et al. (2020). Notably, their study demonstrated the most successful zero-shot transfer outcomes for Farsi, another Indo-European language, as well as for Chinese, which belongs to a different language family altogether. This challenges the hypothesis of language similarity as a factor influencing zero-shot transfer success.

Another potential explanation for the absence of the transfer results observed in this study may be related to different methodologies employed in constructing the WiC datasets in our study compared to previous research, as well as the varying levels of granularity in sense distinction. Raganato et al. (2020) utilized WordNet and Wiktionary for constructing their WiC datasets, reporting larger zero-shot transfer for the former than the latter. WordNet offers highly refined sense distinctions, often surpassing distinctions made by native speakers themselves. In contrast, our Hindi WiC dataset relies on human annotations of text, similar to SemCor Miller et al. (1993), which yields more natural but coarser sense distinctions for polysemous words. Consequently, our dataset aligns more closely with Wiktionary in terms of sense distinction granularity. Following this line of reasoning, it would be expected to exhibit even less zero-shot transfer, as our results confirm. Importantly, we are not asserting that transfer between English and Hindi is unattainable, or that finer WordNet-based sense distinctions are inferior to coarser distinctions. Instead, we emphasize the significance of considering the different levels of granularity between the source and target languages when assessing transferability.

7. Future work and Limitations

In future research we propose to investigate these two potential factors, language similarity and sense granularity, in order to understand how new resources in other languages should be curated. We emphasize that these factors, which may explain the discrepancy between our results and these of Raganato et al. (2020), do not undermine or weaken the validity or our empirical findings that such transfer is not found in the Hindi WiC.

Curating the Hindi WiC was greatly facilitated by the availability of the above mentioned sense annotated Hindi Corpus (SAHC) (Singh and Siddiqui, 2016). SAHC provides a list of polysemous words, and the sentences in which these words are used in. This allowed us to transform SAHC to a WiC-style dataset rather straightforwardly (see § 4). However, many low resource languages may lack similar sense annotation corpora, a glossary of polysemous words, or even a quality and extensive digitized corpora, as per their definition of being low resource. We thus expect that curating a WiC-style dataset in other low resource languages would require more efforts.

Given the time and costs involved in curating human annotated polysemous dataset like WiC, our findings are able to mitigate at least some of these concerns. Our study demonstrates that the number of training examples (i.e., sentences) required to obtain significant gains in the WiC setup need not be high, and that models can generalize quite well to ”out of vocabulary” polysemous words (i.e., number of unique polysemous words can be kept quite low). Understanding that the size and scope of the required resources is within their reach, researchers may be more inclined to take the challenge and invest in their development. In an era where AI is rapidly evolving and its impact on our daily lives is steadily growing, it is crucial to ensure that modern Machine Learning and NLP models offer effective support to as many languages as possible.

8. Conclusions

In this paper we presented the first WiC dataset in Hindi with the objective to advance NLP research in Hindi. The study has evaluated the efficacy of transfer learning using the English WiC dataset in zero shot and few shot settings. The outcomes of this study provide strong and converging evidence that in order to handle polysemy models must be trained on data in the relevant language. This underscores the importance of creating high quality language-specific datasets.

9. Ethical considerations

In the realm of NLP, there has been notable progress in embracing linguistic diversity. However, a substantial gap remains when it comes to low-resource languages, which continue to be underrepresented. Hindi serves as a prime example, being one of the world’s most widely spoken languages. Despite efforts to promote inclusivity, the lack of Hindi-specific resources, including a sense-annotated Hindi corpus for constructing datasets, is glaring, which has an adverse effect on basic NLP capabilities in Hindi. This research paper aims to shed light on the deficiencies in ongoing inclusivity efforts within NLP. It introduces the inaugural Hindi variant of WiC, an innovative resource poised to bolster Hindi-focused NLP research, particularly in Word Sense Disambiguation and polysemy handling. This resource has the potential to enhance Hindi’s overall performance in various NLP tasks that require word sense disambiguation as part of pre-processing stage, or that are sensitive to polysemous words. The creation of such resources underscores the NLP community’s recognition of the critical role that linguistic diversity plays in driving various NLP applications and advancing the field, thereby promoting digital and AI equality worldwide.

10. Acknowledgements

This research was partly funded by the research program Change is Key! (under ref. num. M21-0021) supported by Riksbankens Jubileumsfond.

11. Bibliographical References

References

Ganesh Chandra and Snajay K. Dwivedi. 2014. A literature survey on various approaches of word sense disambiguation. In *2014 2nd International Symposium on Computational and Business Intelligence*, pages 106–109.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Archana Kumari and DK Lobiyal. 2022. Efficient estimation of hindi wsd with distributed word representation in vector space. *Journal of King Saud University-Computer and Information Sciences*, 34(8):6092–6103.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Michael E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *ACM International Conference on Design of Communication*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2023. SemEval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2227–2234, Toronto, Canada. Association for Computational Linguistics.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with big brother's help – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Satyendr Singh and Tanveer J Siddiqui. 2015. Role of semantic relations in hindi word sense disambiguation. *Procedia Computer Science*, 46:240–248.

Satyendr Singh, Tanveer J. Siddiqui, and Sunil K. Sharma. 2014. Naïve bayes classifier for hindi word sense disambiguation. In *Compute*.

Manish Sinha, Mahesh Kumar Reddy, Pushpak Bhattacharyya, Prabhakar Pandey, and Laxmi Kashyap. 2004. Hindi word sense disambiguation.

Mirza Yusuf, Praatibh Surana, and Chethan Sharma. 2022. Hindiwsd: A package for word sense disambiguation in hinglish & hindi. In *WILDRE*.

12. Language Resource References

Bhattacharyya, Pushpak and Pande, Prabhakar and Lupu, Laxmi. 2008. *Hindi WordNet*. [link].

Miller, George A. and Leacock, Claudia and Tengi, Randee and Bunker, Ross T. 1993. *A Semantic Concordance*. [link].

Satyendr Singh and Tanveer J. Siddiqui. 2016. *Sense Annotated Hindi Corpus*. IEEE. [link].