

Building Low-Resource NER Models Using Non-Speaker Annotations

Tatiana Tsygankova[†], Francesca Marini[†], Stephen Mayhew[‡], Dan Roth[†]

[†]University of Pennsylvania, Philadelphia, PA, 19104

[‡]Duolingo, Pittsburgh, PA, 15206

ttasya@seas.upenn.edu, fmarini@sas.upenn.edu
stephen@duolingo.com, danroth@seas.upenn.edu

Abstract

In low-resource natural language processing (NLP), the key problem is a lack of training data in the target language. Cross-lingual methods have had notable success in addressing this concern, but in certain common circumstances, such as insufficient pre-training corpora or languages far from the source language, their performance suffers. In this work we propose an alternative approach to building low-resource Named Entity Recognition (NER) models using “non-speaker” (NS) annotations, provided by annotators with no prior experience in the target language. We recruit 30 participants to annotate unfamiliar languages in a carefully controlled annotation experiment, using Indonesian, Russian, and Hindi as target languages. Our results show that use of non-speaker annotators produces results that approach or match performance of fluent speakers. NS results are also consistently on par or better than cross-lingual methods built on modern contextual representations, and have the potential to further outperform with additional effort. We conclude with observations of common annotation practices and recommendations for maximizing non-speaker annotator performance.

1 Introduction

Work in low-resource languages is not only academically compelling, breaking from popular use of massive compute power on unlimited English data, but also useful, resulting in improved digital tools for under-resourced communities. Two common strategies for low-resource NLP include (a) building cross-lingual models, and (b) annotating data in the target language.

Cross-lingual approaches – in which models are trained on some high-resource language, and applied to the target language – have been shown to be surprisingly effective (Wu and Dredze, 2019;

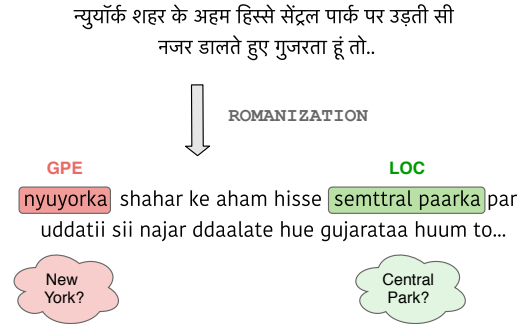


Figure 1: An example of how romanized Hindi text can be annotated without prior language knowledge.

Lample and Conneau, 2019). However, in certain common circumstances, such as when working with languages with insufficient training corpora or those far from the available source languages, cross-lingual methods suffer (Wu and Dredze, 2020). Absent sufficient cross-lingual methods, conventional wisdom suggests that only native (or fluent) speakers of a language are able to provide useful data to train NLP models. But in low-resource scenarios, fluent speakers may not be readily available.

To address this limitation, we hypothesize that the search for annotators can be extended beyond fluent speakers. In this work, we propose an unconventional approach for low-resource named entity recognition (NER) by getting annotations from annotators with no familiarity in the target language, referred to as “non-speaker” annotation. Research in human word recognition (Dijkstra, 2007) suggests that when encountering words of different languages, annotators are able to use phonetic, syntactic, and even semantic information from their languages of fluency to inform recognition. One example of how phonetic information can be used for NER annotation is shown in Figure 1.

We test our hypothesis in a carefully controlled annotation experiment, comparing the performance

of non-speakers (NS) annotators to that of fluent speakers (FS) in Indonesian, Russian, and Hindi.

Our findings are summarized in three key take-aways: (1) the performance of NS models approaches FS models over time; (2) non-speaker annotations are on par or better than cross-lingual methods built on modern contextual representations; and finally (3) individual annotators improve over time, demonstrating that models built on NS annotations have the potential to significantly outperform current cross-lingual methods, given more time spent annotating. We conclude our analysis with observations over factors that can influence NS annotation quality, such as availability of a good romanization system, or presence of capitalization in the target language.

2 Related Work

Named Entity Recognition (NER) has been studied for many years (Ratinov and Roth, 2009; Lample et al., 2016; Ma and Hovy, 2016), with most focus on English and a few other European languages (Tjong Kim Sang and De Meulder, 2003).

In recent years, there has been growing interest in low-resource NLP, with work in part-of-speech tagging (Agić et al., 2015; Fang and Cohn, 2016; Plank and Agić, 2018), dependency parsing (McDonald et al., 2011; Rasooli and Collins, 2017), machine translation (Gu et al., 2018; Xia et al., 2019), and other fields. In particular, low-resource NER has seen work using Wikipedia (Tsai et al., 2016), bilingual dictionaries (Mayhew et al., 2017), self attention (Xie et al., 2018), parallel projection (Enghoff et al., 2018), and zero-shot transfer with multilingual contextual representations (Wu and Dredze, 2019).

In the past, others have studied the effect of having *non-expert* annotators (Snow et al., 2008; Novotney and Callison-Burch, 2010), where annotators speak the target language, but are not formally trained in the target task.

A preliminary investigation of non-speaker annotations was conducted by Mayhew et al. (2019), in the context of partial annotations. In that work, non-speakers were shown (romanized) Bengali text, and were instructed to annotate only those entities for which they had high confidence, leading to high-precision and low-recall annotations. Their proposed model, which we use later, gave large improvements over a standard NER model.

Several interfaces have been developed for non-

speaker annotations in NER, including ELISA IE, Dragonfly (Lin et al., 2018), and TALEN (Mayhew, 2018), which we use.

A similar approach has been proposed for machine translation (Hermjakob et al., 2018b) and speech recognition (Jyothi and Hasegawa-Johnson, 2015; Chen et al., 2016). In the former case (assuming the translation direction is Foreign-to-English), it is often sufficient to translate several of the most important content words, then reconstruct the most likely sentence that uses these, an observation also made in (Pourdamghani et al., 2019). In the case of speech recognition, it is possible to listen to a language one does not speak, and produce a phonetic transcriptions that can be aggregated with others into a reasonable transcription, a process referred to as *mismatched crowdsourcing*.

3 Experimental Setup

Our experiment consisted of a series of trials, typically attended by 1-5 participants and spread out over the course of a few weeks to accommodate individual schedules. Each trial ran for four hours and consisted of three tasks: (1) one-hour instructional training, (2) 20-minute English annotation exercise and (3) series of five 30-minute sessions annotating documents in the target language, with corresponding breaks in between. An overview of a typical annotation session is shown in Figure 2.

In the study, three target languages are used: Indonesian, Russian and Hindi. These languages, arguably high- or mid-resource, were chosen based on availability of gold-annotated data and varying dissimilarity to English, both in terms of entity overlap and script, with Indonesian being the most similar and Hindi being the least similar.

Participant Selection

In total, there were 30 participants involved in the study, selected largely through a network of friends and acquaintances at the University of Pennsylvania, and partially recruited from an introductory NLP class. All participants were uniformly paid \$10/hour for their time. While participants were allowed to participate in multiple experiment sessions, preliminary screening ensured that they did not work with languages they had any prior exposure to, or attended multiple sessions for the same language. Top NS annotators of each language were rewarded with a university mug as an additional incentive to do well. Finally, we chose

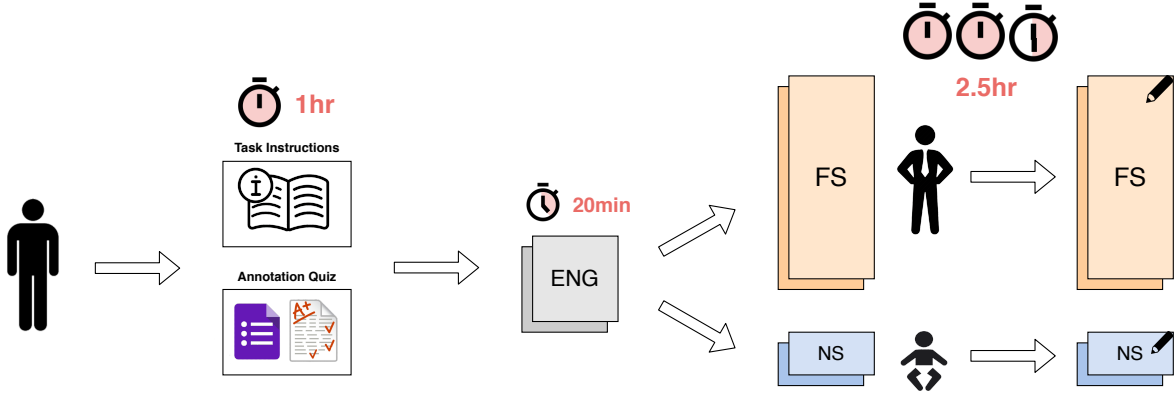


Figure 2: Timeline of a typical annotation trial for a participant, including (1) one-hour instructional training, (2) 20-minute English annotation exercise and (3) series of five 30-minute sessions annotating documents in the target language. The size of the document set that the participant was tasked with annotating depended on their role as either a fluent speaker (FS) or non-speaker (NS), with NS annotators each assigned only a quarter of the train set.

Language	Train	Dev	Test
Indonesian	76K	18K	16K
Russian	59K	16K	16K
Hindi	72K	18K	20K

Table 1: Size of LORELEI datasets for each language, measured in tokens. Splits were created by the authors.

Language	FS	NS
Indonesian	19K	38K
Russian	28K	38K
Hindi	18K	45K

Table 2: Size of datasets produced by fluent speaker (FS) and non-speaker (NS) annotators, in tokens.

not to use Mechanical Turk for this task to allow flexibility in administration format and recruitment strategy. The methodology for the study was approved by the Institutional Review Board at the University of Pennsylvania.

Data

For our experiments, we used gold-annotated NER data from the LORELEI project (Strassel and Tracey, 2016; Tracey et al., 2019). This data uses 4 entity tags: Person (PER), Organization (ORG), Location (LOC), and Geo-political Entity (GPE). We created train, dev, test splits of these datasets ourselves, statistics of which can be seen in Table 1.

To account for annotation speed differences, FS and NS annotators were given document sets of different sizes to annotate during the same time frame. The train set was divided into 4 equally-sized disjoint subfolders, with each NS annotator working on only one subfolder per trial to maximize annotator coverage. Each document set used in the experiment was annotated by at least two participants, resulting in a minimum of 2 FS and 8 NS annotators recruited for each language (a visual reference can be found in Figure 3).

Task 1: Instructional Training

Much of our experimental design was motivated by the need for a controlled, reproducible environment, which resulted in training resources prepared ahead of time to ensure minimal variation in instruction between annotation trials. In total, two instructional documents were used – one providing an overview of the task goals and annotation software, and the other outlining key annotation principles in the form of an interactive annotation guideline quiz. The annotation software used was TALEN (Mayhew, 2018), a tool specifically designed for annotating named entities when the annotators don’t speak the target language. The annotation quiz consisted of 23 questions and provided participants with detailed feedback upon submission.

Task 2: English Annotation Exercise

Following the quiz, participants were asked to annotate English LORELEI data for 20 minutes. The goal of this exercise was both to familiarize the participants with the software interface and provide an indicator for their annotator potential and understanding of the annotation guidelines. We used this indicator later in the preliminary data analysis to filter out low-quality annotators.

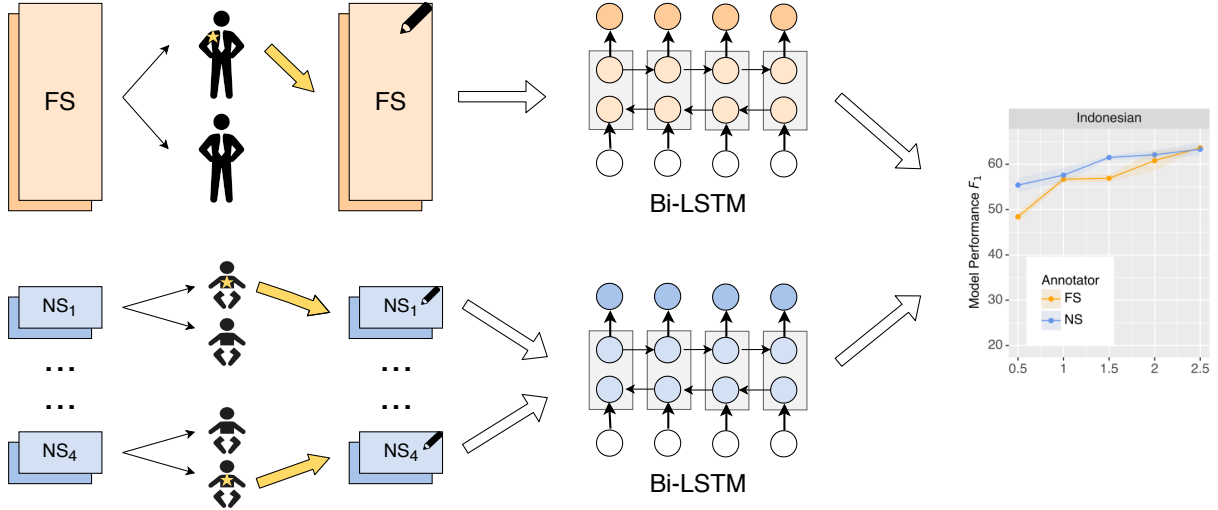


Figure 3: An overview of the data selection process involved in training models on the FS (fluent speaker) and NS (non-speaker) annotations. In each document set, the stars refer to annotators with the higher English exercise score, whose data is used in training. Details on model performance for each language are shown in Figure 4.

Task 3: Target Language Annotation Sessions

Participants completed their 2.5 hours of annotation in 5 sessions of 30 minutes each. All FS annotators spent their time annotating documents in their native language, while NS annotators worked with foreign languages that they had no prior exposure to. Given that all of the languages used in the study were high- to mid-resource, annotators were given explicit instructions not to use external model resources such as Google Translate, but were allowed to use internet search to find any relevant maps, pictures or Wikipedia articles to determine the nature of the entities. For Russian and Hindi, which do not use a Latin script, we provided uroman (Hermjakob et al., 2018a) romanization, so that the script was not a barrier to successful annotation. In general, annotators rarely finished the entire set of documents assigned to them during the five sessions, and those that did were encouraged to go through the same documents again until the trial was over. Often times, this type of “speeding” resulted in many empty documents which were removed in a later post-processing of the data. A summary of the annotated documents can be found in Table 2.

4 Experiments & Analysis

Once we had finished gathering annotations from FS and NS annotators, we ran several experiments on them. In this section, we describe the setup of our models and metrics used, and then move on to discuss key experimental takeaways.

Language	FS			NS		
	P	R	F_1	P	R	F_1
Indonesian	80.6	75.6	78.0	59.8	55.7	57.7
Russian	69.0	67.3	68.1	57.0	45.9	50.9
Hindi	85.5	80.4	82.9	59.8	33.4	42.8

Table 3: Annotation quality of annotations collected from fluent speaker (FS) and non-speaker (NS) annotators against the gold data.

4.1 Models & Metrics

Two Performance Measures In this work, we report two distinct F_1 performance measures, calculated in different ways. We refer to the two metrics as *Annotation Quality* and *Model Performance*.

Annotation Quality refers to the results of participant annotation compared to the existing gold annotations on the same documents. In this evaluation, no model is trained, and we simply calculate the F_1 scores by treating NS annotations as predictions themselves. The annotation quality scores of FS and NS data across all languages are shown in Table 3. Unsurprisingly, the F_1 scores of FS data vastly exceed scores of NS data. There is also an expected decrease in NS scores correlated with perceived language difficulty, with Indonesian having the highest quality, and Hindi having the lowest.

In contrast, *Model Performance* refers to the more traditional NER setup, in which we train a model over obtained annotations, and predict on some held out test set. The following sections outline the results of this performance metric.

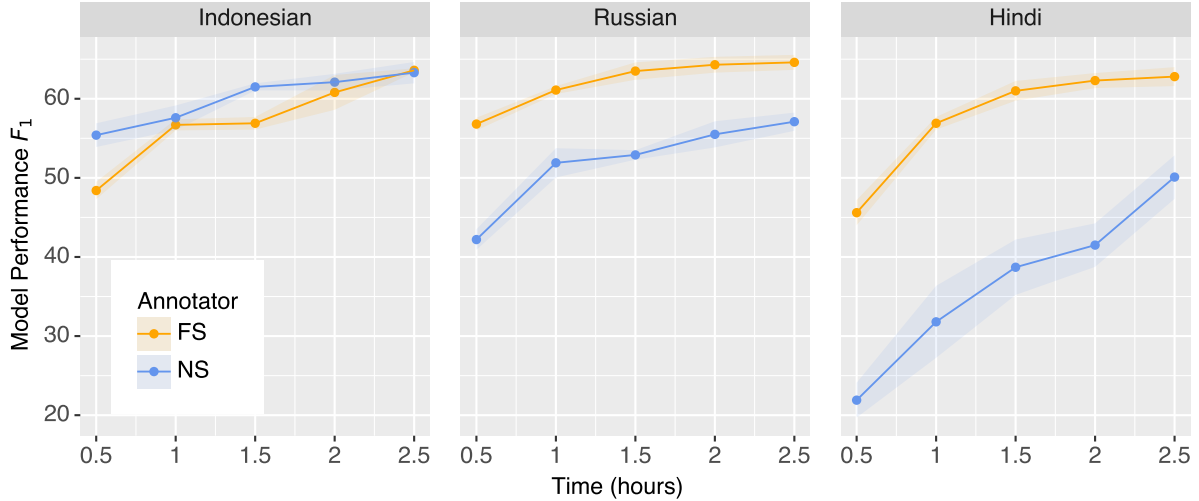


Figure 4: Performance of models trained on fluent speaker (FS) and non-speaker (NS) annotations on Indonesian, Russian and Hindi test data, showing a surprisingly small gap, widening with language difficulty. The shaded regions depict range of model performance over 5 trials. The FS annotations are made by a single annotator, while the NS annotations are compiled from four different annotators, working on partitions of the FS set. As a result, the x -axis reports wall-clock time, and not human effort time.

Data Preparation To account for random errors, we prioritized recruiting at least two participants to annotate each document set used in the experiment. We then used English exercises scores to choose between the resulting conflicting annotations for the same document sets. A summary of the data selection process is shown in Figure 3.

In order to ensure that documents with no annotations were considered to be NS annotator mistakes rather than negative training examples, we removed all empty documents from the NS data before training. No other pre-processing was done.

Machine Learning Models For all experiments, we used a standard BiLSTM-CRF model (Ma and Hovy, 2016) implemented in AllenNLP (Gardner et al., 2018), and used multilingual BERT embeddings (Devlin et al., 2019), which have been shown to exhibit surprising cross-lingual properties (Wu and Dredze, 2019). For the sake of speed and simplicity, we use BERT embeddings as features, and do not fine-tune the model. For each dataset, we train with 5 random seeds (Reimers and Gurevych, 2017) and report the average.

4.2 Main Results: Key Takeaways

Our experimental analysis on the obtained annotations points to three main takeaways:

1. Performance of models trained on NS annotations approaches the performance of models trained on FS annotations (Figure 4).

2. Performance of models trained on NS annotations is on par with current cross-lingual methods, consisting of models trained without target language data (Figure 5).
3. NS annotation quality improves over time, as annotators get more comfortable and familiar with the target language (Figure 6).

Takeaway 1: NS Approaches FS Annotations

Figure 4 compares the performance of models trained on NS annotations to models trained on FS annotation over time. The time increments on the horizontal axis of the figure correspond to the five annotation sessions done by the annotators. Models are trained cumulatively on each session, meaning that training data size consistently increases.

As seen in the figure, across all languages there is a clear upwards trend in performance for models trained on NS annotations, and a stable plateau of FS scores around 60 F_1 . While a gap persists between NS and FS model performance, the size of the gap depends on the perceived language difficulty. As a result, the NS performance for Indonesian is closest to the corresponding FS scores, while the NS performance for Hindi is farthest. Although the duration of annotation collection only spanned 2.5 hours, the upwards trajectory of NS model performances suggests that over time, NS scores will improve and bridge the gap to FS scores.

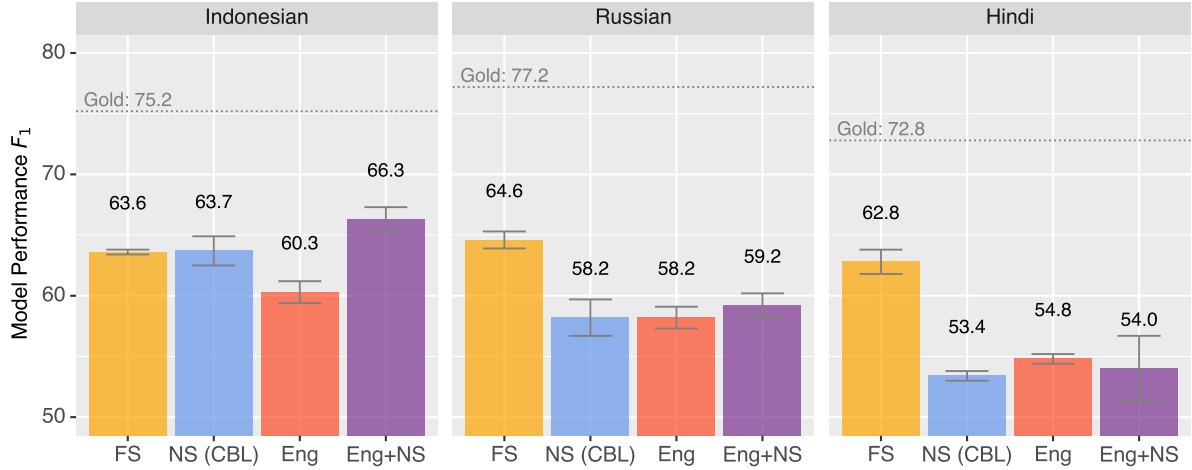


Figure 5: Comparison of models trained on fluent speaker (FS) and non-speaker (NS) annotations to cross-lingual models, showing comparable or improved performance across all languages. Error bars show one standard deviation calculated over five trials. The NS (CBL) results refer to those found in Table 4, representing the best available NS scores. The Eng+NS model is trained on the concatenation of English and NS data. The dashed lines refer to the performance of models trained on the gold annotated training set.

Language	F_1	F_1 (CBL)	δ
Ind NS	63.3 ± 1.4	63.7 ± 1.5	0.4
Rus NS	57.1 ± 1.2	58.2 ± 0.6	1.1
Hin NS	50.1 ± 2.7	53.4 ± 1.8	3.3

Table 4: Model performance improvement (δ) for model trained on non-speaker (NS) annotations using Constrained Binary Learning (CBL) methods for partial annotations. Results shown with \pm one standard deviation calculated over five runs.

The FS performance plateau seems lower than what one might expect from gold annotations for several reasons. First, these annotations are created in 2.5 hours, which makes it an extraordinarily small amount of data for gold annotations. Second, these FS were given a relatively small amount of training, and are subject to different guideline interpretations than the annotated test documents. In a dedicated corpus-building effort, one might provide more focused training to each annotator.

While it is tempting to assume that the lower results for Hindi point to shortcomings of NS annotations on more difficult languages, that is not necessarily the case. The existing positive trend in Hindi NS scores suggests that more time is required to obtain comparable results to NS scores of other languages. As a result, we predict that with additional annotation time and training, NS performance on more difficult languages will not be significantly lower than that of other languages.

We recognize that these annotations are missing many entities. Following recent work in the literature on partial annotations, we use an iterative method from (Mayhew et al., 2019) called Constrained Binary Learning (CBL) that detects tokens likely to be entities and down-weights them in training. Results are shown in Table 4.

Takeaway 2: NS Remains On Par With Cross-Lingual Baselines

As a strong language-independent baseline for existing cross-lingual methods, we trained models on English NER data and evaluated on the target language test data. For the English data, we manually re-annotated half of the CoNLL 2003 data to align with the LORELEI tagset, removing MISC tags and converting LOC tags to GPE tags where appropriate. Experiments on cross-lingual models trained on related languages showed similar results to English, and are included in the supplementary materials. A comparison against models trained on gold-annotations is included as well.

The results are shown in Figure 5. Across all languages, adding NS annotations to English training data consistently matches or exceeds scores for cross-lingual only models, within a margin of error. For Indonesian especially, there is a notable benefit to including NS data, while for the other languages, the improvements are more modest.

Note that performance of cross-lingual models depends on the resources available for the target

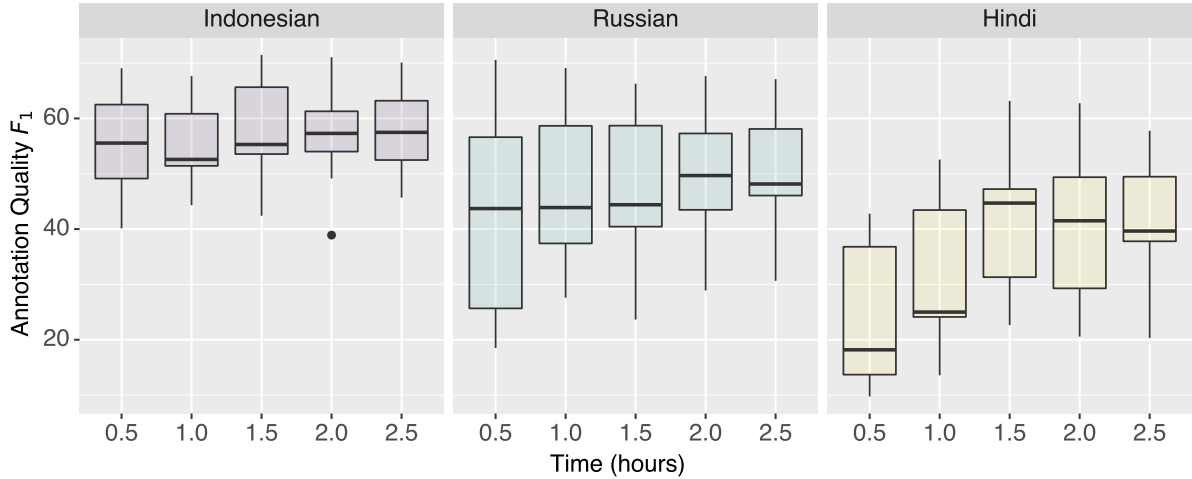


Figure 6: Changes in annotation quality of aggregate non-speaker (NS) annotations over time show an upwards trajectory that steepens with language difficulty. Horizontal lines represent means; points outside of the boxplots denote outliers.

language (i.e. strength of cross-lingual representations), so their performance on high-resource languages tested here is artificially high – one could expect cross-lingual performance to decrease on lower resource languages (Wu and Dredze, 2020). In those cases, the use of NS annotations could prove to be especially useful.

Another benefit to using NS annotations comes from their growth potential. For cross-lingual models, additional training data is unlikely to boost performance. In contrast, an upwards trend of NS model scores to approach performance of FS models suggests that adding more (and better) NS annotations could push results beyond those of cross-lingual methods, and be helpful in the long run.

While an unexpected observation shows that FS scores are always 15–20 points below models trained on gold-annotated data, we hypothesize that this difference can be mainly attributed to training level and not language ability, in addition to a domain shift between annotators (Geva et al., 2019).

Takeaway 3: Annotators Improve Over Time

The earlier observed performance increase of models trained on NS annotations over time can be attributed to two key factors. One contributing factor is data size, which is clearly increasing since NS models are trained on cumulative annotations at each time interval. Another, more interesting, contributing factor to the performance increase is data quality, measured using annotation quality F_1 . In Figure 6, we summarize trends in annotation quality for individual annotators over time.

Across the three languages, we can observe different trends in annotation quality. In Indonesian, there is no significant change in annotation quality over time, suggesting that at the start of the experiment, NS annotators already produce high quality annotations. In contrast, in Russian and Hindi, there is an evident improvement, indicating that NS annotators become better at annotating these languages over time.

The degree to which annotators improve over time is likely correlated with the perceived difficulty of the language. For languages with higher entity overlap with English, such as Indonesian, there is less language-specific information for annotators to learn over time. Model performance improvement for such languages is attributable mainly to increase in data size. In languages less similar to English, such as Russian or Hindi, there are more nuances to the language which must be noticed by annotators over time, resulting in a more overt learning curve. Model performance improvement in these cases is attributable to both increase in data size and annotation quality improvement over time. As NS annotators become more familiar with both the task and the intricacies of the language, their ability to accurately annotate increases, as demonstrated by our data.

For languages with a steeper learning curve, one possible tactic to improve NS model performance would be to disregard low-quality data produced earlier in the annotation process. Following our goal of depicting FS/NS model capabilities as they are, we chose not to take this approach.

5 Discussion

While Section 4 showed quantitative outcomes of experimental processes, this section explores the many factors that can contribute to obtaining good quality NS annotations.

NS Annotation Practices & Strengths

When capitalization is available in the target language, it is a strong indicator for named entities. Analyzing NS annotations over languages which have capitalization – Indonesian and Russian – shows that over 90% of annotated tokens are capitalized, a rate similar to what we would expect in English. Though capitalization is a valuable clue, over-reliance on it can result in both spanning issues and the mislabelling of entities at the beginning of sentences.

For languages with non-Latin scripts – Russian and Hindi – NS annotators often relied on phonetic clues and always annotated on romanized versions of the text. Having access to well-romanized text is critical, as it helps NS annotators make connections between English cognates or previously tagged entities. Some real examples of phonetically recognizable entities from Hindi are:

*paakistaan, amariikaa, biibiisii hindii,
baamglaadesh, ddonaldda ttrampa*

A majority of entities tagged in languages with no capitalization are either geo-political entities (i.e. Pakistan, America) or well-known Western names (i.e. Obama, Twitter, BBC). Once an annotator learns a word representation in the target language, they tend to tag every instance as an entity. We found that NS annotators tend to tag a proportionally less diverse set of entities than FS annotators. But even repeated entities show up in diverse contexts, useful for model training and generalization. Familiarizing annotators with local geography and political climate would also improve the diversity of entities tagged.

One strength of human non-speaker annotators to annotate NER is that — unlike an automatic system — they are able to make inferences over common sense world knowledge. For example, they are familiar with structural elements of different documents, and could leverage a header in a news article to pick out the basic location and scope of the report. In addition, they are also able to use neighboring entities to inform their decisions – in Figure 1, finding New York informs the existence of Central Park as a taggable entity.

What makes a good annotator?

In examining individual annotator performance, we find that there are not many quantitative prerequisites for what makes a good annotator. Analyzing participant language familiarity and instructional quiz scores shows that neither multilingualism nor preliminary guideline understanding present a clear predictor for good annotators. Instead, the characteristics are a lot more subjective: participants who performed best were detail-oriented, patient, and often proactively vocalized their interest in the task or the top annotator award incentive.

How does this generalize to other tasks and languages?

In early stages of this project, we tried annotation with Chinese (Mandarin) and Arabic (modern standard). In both cases, the romanization didn't contain enough phonetic information for the annotators to make it useful. However, prior projects have seen success in annotating NER for such languages as Kinyarwanda, Sinhalese, Ilocano, and Odiya. In each case, annotators were trained more thoroughly, and exhibited a much more focused effort.

Looking to other NLP tasks, it seems clear that non-speaker annotations of conceptually in-depth tasks such as dependency parsing or textual entailment are unlikely to have usable quality. However, for tasks such as part of speech tagging, it could be possible, especially with the help of a tag lexicon and an elementary grammar.

6 Conclusion

In this work, we demonstrate the effectiveness of using non-speaker annotations as an alternative to cross-lingual methods for building low-resource NER models. A qualitative exploration of the resulting data provides insights about what makes NS annotators so unintuitively successful. One avenue for future exploration is with active learning (Settles, 2009), which has been shown to help in low-resource situations (Chaudhary et al., 2019). Further work may also explore optimal ways to combine NS annotators with FS annotators, should they be available. Finally, we encourage others to extend this work to additional languages, including those with poor romanization tools.

References

- Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. [If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272, Beijing, China. Association for Computational Linguistics.
- Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime Carbonell. 2019. [A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5164–5174, Hong Kong, China. Association for Computational Linguistics.
- Wenda Chen, Mark Hasegawa-Johnson, and Nancy F Chen. 2016. Mismatched crowdsourcing based language perception for under-resourced languages. *Procedia Computer Science*, 81:23–29.
- 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.
- Ton Dijkstra. 2007. The multilingual lexicon. *Handbook of psycholinguistics*, pages 251–265.
- Jan Vium Enghoff, Søren Harrison, and Željko Agić. 2018. [Low-resource named entity recognition via multi-source projection: Not quite there yet?](#) In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 195–201, Brussels, Belgium. Association for Computational Linguistics.
- Meng Fang and Trevor Cohn. 2016. [Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection](#). In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 178–186, Berlin, Germany. Association for Computational Linguistics.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Taffjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. [AllenNLP: A deep semantic natural language processing platform](#). In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.
- Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. [Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.
- Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. [Universal neural machine translation for extremely low resource languages](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.
- Ulf Hermjakob, Jonathan May, and Kevin Knight. 2018a. [Out-of-the-box universal Romanization tool uroman](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 13–18, Melbourne, Australia. Association for Computational Linguistics.
- Ulf Hermjakob, Jonathan May, Michael Pust, and Kevin Knight. 2018b. [Translating a language you don’t know in the Chinese room](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 62–67, Melbourne, Australia. Association for Computational Linguistics.
- Preethi Jyothi and Mark Hasegawa-Johnson. 2015. Acquiring speech transcriptions using mismatched crowdsourcing. In *AAAI*, pages 1263–1269.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee. 2018. [Platforms for non-speakers annotating names in any language](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.
- Xuezhe Ma and Eduard Hovy. 2016. [End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.
- Stephen Mayhew. 2018. [TALen: Tool for Annotation of Low-resource ENtities](#). In *ACL Demonstrations*.

- Stephen Mayhew, Snigdha Chaturvedi, Chen-Tse Tsai, and Dan Roth. 2019. [Named Entity Recognition with Partially Annotated Training Data](#). In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*.
- Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. [Cheap Translation for Cross-Lingual Named Entity Recognition](#). In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. [Multi-source transfer of delexicalized dependency parsers](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215. Association for Computational Linguistics.
- 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.
- Nima Pourdamghani, Nada Aldarrab, Marjan Ghazvininejad, Kevin Knight, and Jonathan May. 2019. [Translating translationese: A two-step approach to unsupervised machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3057–3062, Florence, Italy. Association for Computational Linguistics.
- Mohammad Sadegh Rasooli and Michael Collins. 2017. [Cross-lingual syntactic transfer with limited resources](#). *Transactions of the Association for Computational Linguistics*, 5:279–293.
- Lev Ratinov and Dan Roth. 2009. [Design Challenges and Misconceptions in Named Entity Recognition](#). In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*.
- Nils Reimers and Iryna Gurevych. 2017. [Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.
- Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.
- Stephanie Strassel and Jennifer Tracey. 2016. [LORELEI language packs: Data, tools, and resources for technology development in low resource languages](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Jennifer Tracey, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott, and Neil Kuster. 2019. [Corpus building for low resource languages in the DARPA LORELEI program](#). In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 48–55, Dublin, Ireland. European Association for Machine Translation.
- Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. [Cross-Lingual Named Entity Recognition via Wikification](#). In *Proc. of the Conference on Computational Natural Language Learning (CoNLL)*.
- Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.
- Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*. Association for Computational Linguistics.
- Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. [Generalized data augmentation for low-resource translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.
- Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. [Neural cross-lingual named entity recognition with minimal resources](#). In *Proceedings of the 2018 Conference on*

Empirical Methods in Natural Language Processing,
pages 369–379, Brussels, Belgium. Association for
Computational Linguistics.