Enriching Interlinear Text using Automatically Constructed Annotators

Ryan Georgi
University of Washington
Seattle, WA 98195, USA
rgeorgi@uw.edu

Fei Xia
University of Washington
Seattle, WA 98195, USA
fxia@uw.edu

William D. Lewis
Microsoft Research
Redmond, WA 98052, USA
wilewis@microsoft.com

Abstract

In this paper, we will demonstrate a system that shows great promise for creating Part-of-Speech taggers for languages with little to no curated resources available, and which needs no expert involvement. Interlinear Glossed Text (IGT) is a resource which is available for over 1,000 languages as part of the Online Database of INterlinear text (ODIN) (Lewis and Xia, 2010). Using nothing more than IGT from this database and a classification-based projection approach tailored for IGT, we will show that it is feasible to train reasonably performing annotators of interlinear text using projected annotations for potentially hundreds of world's languages. Doing so can facilitate automatic enrichment of interlinear resources to aid the field of linguistics.

1 Introduction

In this paper we discuss the process by which a highly multilingual linguistic resource (greater than 1,200 languages) can be built and then subsequently automatically enriched. Although we touch upon tools for building and maintaining such a resource, our focus in this paper is not so much on the process by which we curate the data, but the process by which automatically enrich the data with additional layers of linguistic analysis. Crucially, we show that the linguistic knowledge encapsulated in all of the data, irrespective of the language, can improve the accuracy of NLP tools that are developed for any specific language. This is particularly true for languages that are otherwise highly under-resourced, and where the development of automated NLP tools, such as taggers, are either not possible or very expensive to develop using traditional methods.

We will focus on the development of Part-of-Speech (POS) taggers. POS tagging is generally thought of as a solved task for many languages, with per-token accuracies reaching 97% (Brants, 2000; Toutanova et al., 2003). While these high accuracies can certainly be achieved for languages with substantial annotated resources, many low-resource languages have little to no annotated data available, making such traditional supervised approaches impossible. Given the cost in developing such resources, many languages with insufficient economic or strategic interest may never see dedicated tools. If annotated resources are not available, what methods can be used?

Several approaches have been proposed to solve the problems posed by the shortage of labeled training data. The first are purely unsupervised techniques. POS induction techniques, such as class-based n-grams (Brown et al., 1992) or feature-based HMM (Berg-Kirkpatrick et al., 2010) induce parts-of-speech without the need for labeled data by finding the ways in which words appear to pattern similarly in clusters. However, as Christodoulopoulos et al. (2010) noted, the way to map the induced clusters to meaningful tags is not straightforward.

Other work has looked at solving the issue of a lack of data by using two or more closely related languages where one of the languages is resource-rich. Hana et al. (2004) used Czech resources to tag Russian. This, however, requires the languages to be closely related, and not all resource-poor languages have closely-related resource-rich languages.

Another path of inquiry has been to use one unrelated resource-rich language and alignment tech-

Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 58–67,
Beijing, China, July 30, 2015. ©2015 Association for Computational Linguistics and The Asian Federation of Natural Language Processing

```
LANG   nnisaau  daxalna  makaatibahunna
GLOSS  the-women(3.PL.F.)-NOM  entered-3.PL.F  office(PL.)-ACC-their(F.)
TRANS  "The women  have entered  their offices."
```

Figure 1: An example of Interlinear Glossed Text (IGT) in Arabic from (Nasu, 2001), with an English translation.

niques to "project" information from the resource-rich language to the resource-poor one. Yarowsky and Ngai (2001); Das and Petrov (2011) both investigated training POS taggers by projecting labels from one language to another, while Hwa et al. (2004) looked at projecting dependency parsers.

In this paper, we focus on using a resource known as Interlinear Glossed Text (IGT) as a possible source of linguistic knowledge for the POS tagging task on resource-poor languages, and apply it to the enrichment of a linguistic resource composed of IGT data. An example of IGT is shown in Fig. 1. IGT is a format used by linguists for giving examples of linguistic phenomena, and since linguists study a large number of languages, IGT instances can be found for hundreds of languages. We will explain the precise structure of the data in depth later, but IGT as a resource is appealing not only for its broad coverage, but also the linguistic knowledge it contains. Although it does not typically contain POS tags explicitly, these examples often contain enough data to make inferences which can be used to enrich the data, whether with POS tags or with other syntactic information (Lewis and Xia, 2010).

We present a system which takes advantage of the structure of IGT instances in order to perform automatic part-of-speech tagging of the target language, regardless of the language. While the tagging performance is not necessarily competitive with state-of-the-art supervised systems, it shows great promise for languages with which such supervised systems are not currently possible, and can increase the value of the entire resource to the linguistic and computational linguistic communities.

POS taggers are intrinsically valuable to computational linguists, since they are building blocks for a number of other NLP tools. Theoretical and descriptive linguists might question their value to them; however, they only represent a class of possible annotators. The projection methodology, especially the fact that projection accuracy is boosted by relying on an entire corpus, can be applied to other forms of annotation, such as tags or analyses that may be of benefit for subsequent analyses. Although such taggers will not be as accurate as human annotators, they could reduce workload by doing first pass analyses automatically.

2 The IGT Data Type

As shown in Fig. 1, IGT instances typically contain one line in the target language, then a word-for-word gloss in the language of the paper from which the example is drawn (typically English) and finally a translation. The gloss line is of particular interest for our purposes because of tokens such as the `(3.PL.F)-NOM` often found in IGT, as shown in Fig. 1. This token is intended to signify that the Arabic token `nnisaau` is third person, plural, feminine, and in the nominative case. Each portion of the token, `3`, `PL`, `F`, and `NOM` are grammatical annotations, or **grams** for short[1]. While these grams by themselves do not guarantee that the token is necessarily a noun, they are a strong indicator. We will show how this information may be used in Section 4.

Also of note is that while the gloss line aligns one-to-one with the language line, with three words mapping to three gloss tokens, the translation line has six words. Aligning these tokens is made easy by the words in the gloss line matching words in the translation line. This allows for projection to be performed more precisely than might otherwise be possible using statistical alignment methods.

Previous work on projecting syntactic information between languages differs from our approach in two significant ways. First, projection in previous work has relied on bitexts, which do not benefit from the additional information the gloss line of IGT provides. Therefore, these past methods have relied upon statistical alignment between languages. Obtaining alignments of sufficient quality would likely not be possible for resource-poor languages, since statistical alignment methods require large amounts of parallel text. Using IGT, however, alignment can be obtained more precisely with smaller amounts of data.

[1] While the "gram" moniker is typically used to refer to grammatical function tags specifically, we will use it in this paper to refer generally to all segments of the gloss line that are not whitespace delineated.

Second, while many-to-one are a source of problems for past approaches, IGT offers a possible solution for disambiguating tag-word alignments by examining the grams directly. For instance, the gloss token `boys.ran.3.PRES` may align to both NOUN and VERB tags, but the `3` and `PRES` grams provide evidence that the token is most likely a verb showing agreement. In this paper, we will explore how both of these solutions may help us over traditional approaches to projection.

2.1 Previous Uses of IGT

A number of studies have shown the linguistic knowledge contained in IGT data to be useful. The Online Database of INterlinear text (ODIN) (Lewis and Xia, 2010) contains over 190,000 IGT instances for over 1,000 languages. While still short of the approximately 6,900 languages that exist in the world (Lewis, 2009), this covers an enormous range of languages for which few other resources exist. Using ODIN as a resource, Lewis and Xia (2008) demonstrated via projection using IGT alignments that basic word order of a language could be determined with 99% accuracy if the language contained at least 40 instances of IGT. Georgi et al. (2014) used IGT instances to produce sets of dependency trees which were then corrected and used to learn automatic correction rules.

2.2 INTENT: a Package for Creating Enriched IGT

In the previous sections, we described the forms of useful information that IGT contains. The next step is to programmatically harness that information in order to construct automatic annotators. This is what our system, the INterlinear Text ENrichment Toolkit (INTENT), was designed for. INTENT takes IGT instances as input and produce automatically enriched IGT instances as output. A more in-depth discussion of the system can be found in Xia et al. (2015).

Word alignment is the first crucial phase of INTENT's enrichment strategy. Due to IGT's structure providing a one-to-one gloss and language word alignment, and a gloss line containing many English-language words that co-occur in the translation line, the gloss line can be used as a pivot to align the English language with the target language. INTENT does this either by matching the words from the gloss and translation lines, on a string and morphological heuristics, or by using GIZA++ as a statistical alignment approach.

INTENT's second primary purpose in this paper is to provide part-of-speech tags, which are produced in one of two ways. Either INTENT uses one of the word alignment methods to project the English POS tags onto the gloss, and subsequently the target language word, or it takes advantage of the extra grammatical markers such as *-Nom* (nominative case) or *-Dec* (declarative marker) as features for a classifier to recognize the part-of-speech tag that a gloss word is most likely to be annotating in the target language, without ever needing to look at the target language directly. This means that INTENT can theoretically provide tags for any language for which interlinear text is available.

2.3 Use of INTENT for Linguists

In the Spring of 2015 at the University of Washington, our colleague Prof. Emily Bender used the INTENT system as part of a course, Computational Methods in Language Documentation[2]. The INTENT system was used to enrich IGT instances from the Language CoLLAGE project (Bender, 2014). The students then worked on methods by which typological phenomenon might be determined from the automatically enriched data, following Bender et al. (2013, 2014). This type of inquiry shows the large-scale enrichment of a wide variety of languages that INTENT is intended for, and how this can be used to answer interesting linguistic questions.

Other such uses might be enriching collected IGT instances automatically, to create an annotated corpus from IGT data while being able to greatly reduce the amount of human annotators needed for the task.

For these goals to be effective, INTENT must be able to generate sufficiently reliable POS tags on resource-poor languages. Whether or not that is the case is the question we seek to answer in this paper.

3 Projecting Annotation in IGT

Projection-based approaches work by finding an alignment between two lines, where one has annotation and one does not, and "projecting" the annotations from one to the other. Figure 2 shows

[2] Course website available at: `http://faculty.washington.edu/ebender/2015_575/`

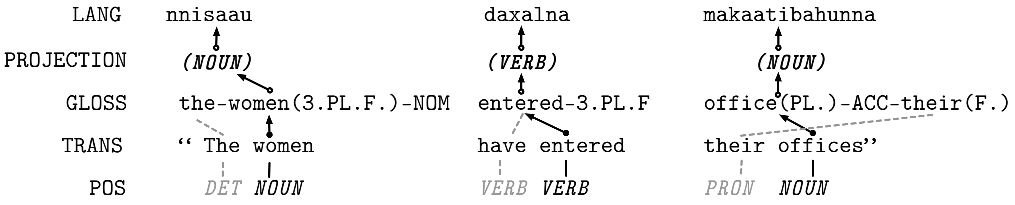

Figure 2: An illustration of how the gloss line from the IGT in Fig. 1 can be used for projection.

an illustration of how this projection occurs, using the sentence from Fig. 1.

Left unspecified is how the alignments between sentences are obtained. Previous papers generated alignments using statistical alignment. For instance, Yarowsky and Ngai (2001) showed 76% POS tag accuracy for projecting directly between English and French, using GIZA++ to automatically align the words and evaluating with a reduced tagset similar to the one used in this paper. Such an approach, however, required a set of parallel data consisting of roughly 2 million words per language, something which would not be available for resource-poor languages.

3.1 Using IGT to Bootstrap Alignment

While ODIN contains many IGT instances, it has nowhere approaching the 2 million sentences used in previous projection approaches, such as Yarowsky and Ngai (2001). Thankfully, IGT contains more information than simply the source and target language data—IGT also has gloss lines. The gloss line is a transliteration of the language line, containing many of the same words that are used in the translation, although in a different order. We can use the gloss line as a “pivot” to bootstrap our alignment, as shown in Fig. 2, and following Lewis and Xia (2008, 2010).

There are five steps to our process of projecting POS tags using the gloss line of IGT:

1. POS tag the translation line
2. Align the translation line with the gloss line
3. Disambiguate multiply-aligned gloss tokens
4. Attempt to resolve unaligned tokens in the gloss line
5. Project POS tags from gloss line to language line

1 – First, an English-language POS tagger is used to provide the POS tag sequence for the translation line. For our tests, we used the Stanford Tagger (Toutanova et al., 2003) trained on all sections of English Penn Treebank (Marcus et al., 1993) with the POS tags remapped from 45 down to the 12 tags in the universal tagset proposed by Petrov et al. (2011).

2 – Next, the words in the translation line and gloss line are aligned; this can be done by one of two ways: heuristically, or using statistical alignment. In the heuristic approach, words are aligned by exact string matches; then stemmed matches, and finally a series of gram mappings, such as `I` aligning to `1SG`. For this paper, we use the heuristic approach.

3 – After step 2, multiple translation words with differing POS tags may be aligned to the same gloss tokens. In Fig. 2, the translation tokens *The women* align to a single complex token `the-women(3.PL.F.)-NOM`. In an effort to use a language-independent way of resolving multiple tags on a single token, we specify an order of precedence by which a tag is selected, prioritizing content words over function words.[3]

4 – On gloss tokens that failed to receive an alignment, we attempt to retrieve a tag for the token based on a dictionary lookup of the individual subtokens. The dictionary is built using the part-of-speech tags from the English Penn Treebank (Marcus et al., 1993), and remapped to 12 universal POS tags following (Petrov et al., 2011). If one or more portions of the token are found in the dictionary, we use the most frequent tag for each of those words to label the token. Multiple tags are resolved the same as if the tags were projected.

5 – Finally, the tags projected to the gloss line are transferred to the language line assuming a one-to-one, monotonic alignment. Since the glosses of IGT are intended to be paired word-for-word with the target language, this is a reasonable assumption. Due to noise in the IGT text files[4],

[3] This order of precedence is: `VERB` > `NOUN` > `ADV` > `ADJ` > `PRON` > `DET` > `ADP` > `CONJ` > `PRT` > `NUM` > `PUNC` > `X`.

[4] Further discussion of the noise in these files is found in Xia et al. (2014).

however, there is not always a one-to-one alignment between the language line and gloss, and in these cases, we skip processing the IGT instance. In addition to alignment failures, noise is found in the form of a bias toward English, since all projections originate in English, as well as a bias toward unusual phenomena that the author of the paper from which the instance is extracted is focusing upon. These issues are discussed further in Lewis and Xia (2008), where they are referred to as the *English bias* and *IGT bias*, respectively.

3.2 Drawbacks of Projection

While Xia and Lewis (2007) show that the heuristic alignment approach can achieve 98% precision, and the recall is between 74% and 85% with fairly clean data. However, in our data, we found that upwards of 60% of tokens were unaligned (see Section 7.1). Tokens that are unaligned are left without a tag, and thus never labeled correctly. In order to address this issue, we next take a look at how the gloss line itself can be used as a means to obtain POS tags.

4 Building a Gloss-Line Classifier

There are three main areas in which the projection method discussed above shows weaknesses: gloss tokens with multiple POS tags aligned to them, gloss tokens that fail to be aligned or found in a dictionary, and the assumption that foreign-language words will share the same POS as the English words in the translation line[5]. For instance, a gloss token `run.NOM` might be labeled as **VERB**, due to aligning with the intransitive verb form of the word in the English line. However, the gram `NOM` is a strong indicator that the word is nominalized and should be tagged as `NOUN`.

It is not the case that the tokens in question lack information on which to base a decision, but rather that such information is perhaps not well-suited for a deterministic approach. Therefore, we propose building a gloss-line classifier that uses the individual subtokens of a gloss-line word as features to make a decision for the label of the token as a whole. Figure 3 illustrates how these subtoken level features might be used in this scenario. In addition to helping resolve the possible ambiguity of multiply-aligned tokens, our approach also avoids the indirection of finding the correct alignment for a gloss token and working on the gloss token directly.

By using the very precise, albeit low recall, heuristic alignment method, we can automatically generate training instances for the gloss-line tagger. Using these automatically-annotated gloss tokens, we then train a classifier using the MALLET package (McCallum, 2002) and its Maximum Entropy implementation. We experimented with different features, but the following set resulted in the highest performance on our development set for classifying a token i:

- Grams contained in token i
- Grams contained in token $i-1, i+1$
- Best dictionary tags for grams

Finally, this gloss line classifier can be used on IGT instances for the target language. After the gloss-line is labeled, the tags are transferred via one-to-one alignment with the language line.

4.1 Context-Sensitive Features

One of the things worth noting about this approach is that, while the IGT instances cover multiple different languages, we can opt to use the gloss lines from all the languages in our annotated data to train the classifier (as we do in Section 6.2). Although the gloss lines may annotate different languages, the tokens in the gloss line are all English words and grammatical markers. This results in a pseudo-language of sorts where the meaning of the tokens is largely consistent between languages.

Although it is convenient to think of the gloss-line as this pseudo-language for the purposes of POS tagging it, we also keep in mind that the word order of this pseudo-language is dependent upon

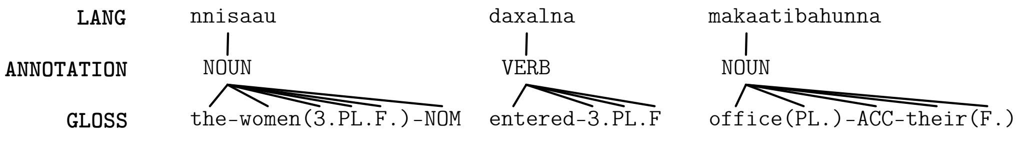

Figure 3: An IGT instance showing the classification-based approach, using the gram-level elements of the gloss line as features to choose a label for each token.

[5] This is an occurrence of the *English bias* noted by Lewis and Xia (2008)

the language it is annotating, and thus context-sensitive features might not generalize well.

5 The Data

For our experiments, we used IGT instances from the Documentation of Endangered Languages (DOBES) project by Bickel et al. (2011) on the Chintang language of Nepal[6], an endangered language with intricate morphology. This corpus included not only thousands of instances, but also gold-standard POS tags for the language line. This high-quality enriched resource is one that allows us to evaluate our method on a truly low-resource language.

5.1 Splitting the Data

From the data above, we split the corpus 80-10-10 for training, development, and testing. Since this work is still in the early stages, only the results on the development set are given here. The breakdown of this data can be seen in Table 1.

5.2 Chintang Tagset

Although tags are manually provided, they are not the same tags as the universal tagset that INTENT uses, so we must map one or the other to evaluate correctly. Table 2 shows our mapping from tags used in the Chintang (CTN) tagset to those in the universal tagset.

Ideally, this mapping should be many-to-1; that is, each Chintang tag maps to a single tag in the universal tagset. However, some Chintang tags can map to multiple tags in the universal tagset.

For instance, for words with tag "*gm*" in Chintang, some are *grammatical markers* that do not have counterparts in English (e.g., a TOPIC marker) and therefore is mapped to *PRT* (for *particle*); others are listed as the English words *and*, *or*, or *but*, and should likely be labeled as conjunctions and mapped to *CONJ*. Other "*gm*" tagged words have variations of *DEM* (for *demonstrative*), and are likely pronoun-like, requiring the *PRON* tag. Table 3 shows the top 12 gloss tokens labeled "*gm*" in the data.

We experiment with two mappings: the first mapping, **Basic Mapping**, is many-to-1, and all *gm* words are mapped to *PRT*; in the second mapping, **Extended Mapping**, the *gm* tokens are split between *CONJ*, *PRT*, and *PRON*, with a simple, 14-token wordlist, consisting of the dozen tokens in Table 3, plus *or* and *DEM* and their associated tags. For the projection-only POS tagger, this extended mapping occurs as post-processing remapping step on the projected tags. For classification, the gloss words are added to the existing dictionary, creating an expanded dictionary with CTN-specific (gloss-token, tag) pairs. This dictionary is then used to provide the best-guess tag feature to the classifier at training and test time.

6 Experiments

For this work, we wanted to test three overall scenarios: using projection alone (§6.1), using the classifier trained on ODIN data (§6.2), and then using the classifier trained on Chintang data (§6.3). All three scenarios will be evaluated on the dev set of the Chintang corpus.

6.1 Projection Only

For the first scenario, since the projection method is deterministic and does not require training, only instances from the Chintang dev set are used. The

	Training Set	**Dev Set**
Instances	7,120	876
Tokens in lang/gloss lines	31,116	3,884
Tokens in trans lines	39,396	4,872

Table 1: Corpus statistics for Chintang IGT data (Bickel et al., 2011).

CTN Tag	**Universal Tag**	**CTN Tag**	**Universal Tag**
adj	ADJ	interj	PRT
adv	ADV	gm	PRT
sound	ADV	vt	VERB
n	NOUN	vi	VERB
predadj	NOUN	v	VERB
num	NUM	v2	VERB
pro	PRON	NoPOS	X

Table 2: Tagset mapping from CTN tags to Universal tagset tags.

Gloss Word	**# Tokens**	**Tag**	**# Tokens**
FOC	1049	CIT	360
TOP	1027	REP	243
SEQ	855	and	237
ADD	621	SURP	236
EMPH	504	SPEC.TOP	223
BUT	365	COND	207

Table 3: Top 12 gloss tokens labeled "*gm*," sorted by decreasing frequency.

[6] http://dobes.mpi.nl/projects/chintang/

projection method described in Section 3 is used, tagging the English translation line, finding alignments on the gloss line, and then using 1:1 alignment from gloss and language to assign tags to the language line. We then evaluate by comparing the projected tags with the manually assigned tags, while also keeping track of the unaligned tokens.

6.2 Classifier Trained with ODIN Data

For the first of the two classification-based settings, we follow the approach described in Section 4, by using our full set of ODIN instances to automatically label the gloss line via projection. We then again use the 1:1 alignments between gloss and language to project to the language and evaluate with those tags.

6.3 Classifier Trained with Unlabeled CTN Data

Finally, for the second classification based approach, we train the classifier with the training portion of the Chintang coprus, ignoring the gold-standard POS labels, to see what effect using instances specific to Chintang might have. In particular, since all the instances used to train the classifier were coming from the same language, we used this experimental setting to test whether adding context features to the classifier would help, in the case that there was enough single-language data.

7 Results

The results are presented in the order given in Section 6, with the projection-only approach first (§7.1), followed by the different classification approaches (§7.2).

7.1 Projection

The results of using projection alone on the Chintang dev set can be seen in Table 4, which shows the POS tagging accuracy in the first column, as well as the unaligned tokens (tokens assigned *UNALIGNED*) in the data.

With only the most basic mapping, we see that the projection-only approaches achieves a mere 12.6% accuracy, with an 83.8% of all tokens in the gloss line unaligned. Figure 4 shows a Chintang instance that illustrates part of the reason for this high amount of unaligned tokens. While many of the instances in the ODIN database frequently contain words that match (if only in their stemmed forms) between translation and gloss line, many of the instances in the Chintang corpus glossed the words only in terms of grammatical markers, such as "SEQ" or "DEM-NMLZ-LOC" as shown in this example.

Method	**Accuracy**	**% Unaligned**
Basic Mapping	12.6	83.8
Extended Mapping	39.6	57.1

Table 4: Results of projection-only approach.

```
hun-ko-i       tis-u-m             pache
DEM-NMLZ-LOC   put.into-3P-1/2nsA  SEQ
(pro)          (vt)                (gm)
after putting dal or arum
```

Figure 4: An instance from the Chintang dev set, showing the lack of alignment between gloss line and translation line, as well as the gold standard POS tags.

The second row of Table 4 shows the result of adding Extended Mapping, which would correctly identify the gloss containing "DEM" as a *PRON* and the "SEQ" gloss as the mapped *PRT*.

7.2 Classification

Table 5 shows the results of the classification-based experiments outlined above, trained on either the ODIN instances or the Chintang (CTN) instances on their own, or combining the two. Shown for comparison also is the result of using the remapped gold standard tags from the Chintang training data to train the classifier, and evalu-

Training Data	Expanded Dictionary	Context Features	Accuracy
ODIN			43.1
	✓		53.0
CTN			75.0
	✓		74.9
		✓	74.8
	✓	✓	74.9
CTN+ODIN			61.6
	✓		70.7
	✓	✓	72.6
Supervised (with Labeled CTN)			89.6
	✓		90.6
	✓	✓	90.1

Table 5: Classification results showing different sets of training data and classifier features.
t

ating on the dev set.

7.2.1 ODIN-Only Training Data

As mentioned in Section 7.1, the IGT instances in the Chintang data look different from many of the instances in the ODIN data, and the accuracy results of 43.1% and 53.0% would seem to confirm the dissimilarities between the data sets. Though the expanded dictionary with CTN-specific gloss tokens seems to help somewhat, the ODIN data suffers because there are simply too many tags assigned by the classifier that do not occur in the CTN data, such as DET (for *determiner*) or ADP (for *adposition*). Even though the results are low, given that these are results for a system which has never been provided with a single instance of Chintang data, they are somewhat promising.

7.2.2 CTN Training Data

The classifier trained on CTN IGT instances fares much better, achieving 75% accuracy. It should be noted that when CTN instances were used to train the classifier, the automatically labeled training data is produced by the projection algorithm which uses the Extended Mapping described in Section 5.2, and thus the expanded dictionary is of little use above the training data that the classifier has already seen.

Finally, combining CTN training instances with ODIN IGT instances achieves only 61.6% without additional features, but when the expanded dictionary is added, as well as the CTN-specific contextual features, we see a result of 72.6%, getting closer to the result seen by the CTN data on its own.

While none of these methods come close to the 90% accuracies seen by the supervised system, our automated system shows promise given that it uses a far more impoverished set of information to train it. To compare these systems in a more real-world setting, we also looked at how the systems performed if the amount of data used to train each was scaled back from the approximately 32,000 tokens in the full training set.

7.2.3 Varying Amounts of Data

The graph in Fig. 5 shows the result of varying the amount of training data used for the different classification approaches. While all the classification-based approaches ultimately converge around 75% accuracy, we can see that when only 500 or fewer tokens are available, a setting which is much more realistic for low-resource languages, these approaches actually do not do poorly by comparison. We also see that the dictionaries help the most when the amount of training data is small.

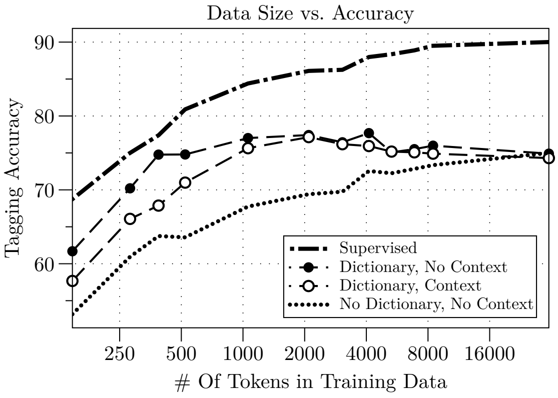

Figure 5: Graph showing how performance varies with respect to the number of training tokens available to the system.

8 Conclusion and Future Work

In this paper, we have demonstrated a proof-of-concept for a system that can potentially produce POS taggers for up to a thousand languages, many of which have little to no annotated linguistic resources available. While the performance is lower than state-of-the-art supervised systems for resource-rich languages, our approach demonstrates a method that can be applied to resource-poor languages, as shown by the Chintang results.

For subsequent work, our goal is to apply this technique to additional resource-poor languages, with different typological characteristics from Chintang, as well as take an additional step toward training monolingual parsers from the tagged language lines.

Acknowledgments

This work is supported by the National Science Foundation Grant BCS-0748919. We would also like to thank Balthasar Bickel and his team for allowing us to use the Chintang data set in our experiments, and our three anonymous reviewers for the helpful feedback.

References

Bender, E. M. (2014). Language CoLLAGE: Grammatical Description with the LinGO Grammar Matrix. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2447–2451, Reykjavik, Iceland. European Language Resources Association (ELRA). ACL Anthology Identifier: L14-1508.

Bender, E. M., Crowgey, J., Goodman, M. W., and Xia, F. (2014). Learning Grammar Specifications from IGT: A Case Study of Chintang . In *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Baltimore, MD.

Bender, E. M., Goodman, M. W., Crowgey, J., and Xia, F. (2013). *Towards Creating Precision Grammars from Interlinear Glossed Text: Inferring Large-Scale Typological Propertie*. In *Proceedings of the ACL workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities*.

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless unsupervised learning with features. Association for Computational Linguistics.

Bickel, B., Stoll, S., Gaenszle, M., Rai, N., Lieven, E., Banjade, G., Bhatta, T., Paudyal, N., Pettigrew, J., Rai, I., and Rai, M. (2011). *Audiovisual corpus of the Chintang language, including a longitudinal corpus of language acquisition by six children, paradigm sets, grammar sketches, ethnographic descriptions, and photographs.* DOBES Archive.

Brants, T. (2000). TnT — A Statistical Part-of-Speech Tagger. In *the sixth conference*, pages 224–231, Morristown, NJ, USA. Association for Computational Linguistics.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Christodoulopoulos, C., Goldwater, S., and Steedman, M. (2010). Two decades of unsupervised POS induction: how far have we come? pages 575–584.

Das, D. and Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 600–609, Portland, OR, USA.

Georgi, R., Xia, F., and Lewis, W. D. (2014). Capturing divergence in dependency trees to improve syntactic projection. *Language Resources and Evaluation*, 48(4):709–739.

Hana, J., Feldman, A., and Brew, C. (2004). A Resource-Light Approach to Russian Morphology: Tagging Russian using Czech Resources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 222–229. Rodopi Bv Editions.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2004). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 1(1):1–15.

Lewis, M. P., editor (2009). *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, sixteenth edition.

Lewis, W. D. and Xia, F. (2008). Automatically identifying computationally relevant typological features. In *Proceedings of the Third International Joint Conference on Natural Language Processing*.

Lewis, W. D. and Xia, F. (2010). Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Nasu, N. (2001). Towards a theory of non-cyclic A-movement. In *Essex Graduate Student Papers in Language and Linguistics*, volume 3, pages 133–160.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. pages 173–180.

Xia, F., Goodman, M. W., Georgi, R., Slayden, G., and Lewis, W. D. (2015). Enriching, Editing, and Representing Interlinear Glossed Text. In *16th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–16.

Xia, F., Lewis, W., Goodman, M. W., Crowgey, J., and Bender, E. M. (2014). Enriching odin. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Xia, F. and Lewis, W. D. (2007). Multilingual Structural Projection across Interlinear Text. pages 452–459.

Yarowsky, D. and Ngai, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second meeting of the North American Association for Computational Linguistics*, Stroudsburg, PA. Johns Hopkins University.