Sigmorphon 2020 Task 0: Data Augmentation for Morphological Inflection in Extremely Low-Resource Languages

Alexander Kahanek
University of North Texas
alexanderkahanek@gmail.com

Amelia Brown
University of Dallas
abrown@udallas.edu

Abstract

One of the great challenges of modern natural language processing (NLP) is the handling of low-resource languages. In this paper, we explore ways of augmenting data from low-resource languages to avoid overfitting and compensate for any frequent errors. We focus on one specific facet of NLP, namely morphological inflection, the process by which the dictionary form of a word is transformed into its various forms inflected forms. Beginning from the baseline system and data used in the SIGMORPHON 2020 Shared Task 0, we employed both multilingual systems and targeted pseudo-bilingual data in order to improve our neural model's ability to inflect words from low-resource languages. Our final results improve on the baseline for one of our target languages and are only slightly less effective on the other.

1 Introduction

There are many possible strategies for improving the performance of NLP models on extremely low-resource languages. These range from massively multilingual models trained on anywhere from tens to hundreds of languages, to multilingual word embeddings intended to extract meaning rather than morphology, to data hallucination, to handwritten rules that inject grammatical knowledge directly into a system.

Which of these strategies is most effective for a particular task depends partially on the task and partially on the time constraints. This experiment was carried out over ten weeks, which limited the time which could be spent exploring widely differing research strategies. In addition, all of the above strategies have been explored by previous researchers, such as **?**, whose work focused on building robust cross-lingual embeddings, Cotterell et al. (2017), who worked on a shared task extremely similar to this one, and the researchers of this year's SIGMORPHON conference, Vylomova et al. (2020). Our research focuses on Shared Task 0 from this year's SIGMORPHON (Special Interest Group on Computational Morphology and Phonology) conference, which deals with training models for morphological inflection. Morphological inflection is the process by which words are transformed from singular to plural forms, or from present to past tense. All languages have rules governing morphological inflection, but they vary widely from language to language, and they often include exceptions for specific words, making them challenging for a machine learning model to learn.

For this project, we focus on improving performance on specific languages which proved challenging to the systems submitted to the conference. We use Middle Low German and Norwegian Nynorsk as our research, or target, languages, because both of them were challenging to the majority of models submitted for the SIGMORPHON 2020 Shared Task 0. The average accuracy which these models achieved on Middle Low German and Norwegian Nynorsk was 11 percent and 14 percent, respectively. Part of the reason for this low performance was that both languages were "surprise languages," meaning that the models were not trained on data for either language, but even among the surprise languages, 14 and 11 percent accuracy were unusually low scores. We use English, German, and Icelandic as our source languages, because average accuracy on all three was over 80 percent and they are all high-resource languages. Our research methodology is to explore various types of data augmentation to improve model performance. We use the neural baseline provided by the task organizers, which contains both a transformer and a hard monotonic attention architecture, for modeling.

All of the five languages upon which our study focuses belong to the Germanic language family. Middle Low German is, as its name suggests, a very close relative of modern German, which was spoken in the Middle Ages before German became a somewhat unified language. Icelandic and Norwegian Nynorsk are both members of the North Germanic language family, and so share both grammar

and orthography to some degree. We train a large number of models on data sets produced by combining the source and target language data in a variety of ways. All of our models are multilingual to some extent, as they were trained on data from anywhere from two to five languages.

We explore multiple ways of augmenting target language data, but all of them in some way aim to make up for the lack of target language data by including words from the source languages in the training data. In some cases the augmentation data from the source language is selected randomly, in others it is targeted to teach the model to handle specific types of words which previous models had frequently gotten incorrect.

The most successful method for data augmentation is targeted augmentation by lemma. The least successful is a zero-shot transfer of a multilingual from the source languages to the target language. Our research shows that targeted data augmentation is a potentially useful strategy for future NLP systems.

2 Ethics

Like any developments in technology, improvements in machine learning (ML) technology are often double-edged swords. The same learning technology which enables the capture of dangerous criminals may also be used to suppress political dissenters, as is the concern of Andersen (2020). The facial recognition technology which makes identity fraud more challenging may also dangerously encroach upon the privacy of individuals. Therefore, this paper would be incomplete without a discussion of the potential ethical repercussions of our research.

Improvements in NLP for low-resource languages do not create any of the clear privacy-related issues that we mentioned above, but they are not without ethical content. The majority of NLP models have been written for so-called "high-resource" languages, which have massive and easily accessible collections of online documents and labelled data. This amount of data allows traditional deep learning methods to be highly effective. Lower-resource languages, however, especially those which have little to no labelled data, are more of a challenge to work with, and additionally there is less commercial demand for systems which use them, and so most of the models which have been built for them are comparatively recent.

Though the moral repercussions of these new models are not immediately apparent, a brief review of extant literature reveals how low-resource NLP can provide great benefits to the world. There are some languages now which are on the verge of linguistic extinction, and many field linguists are working tirelessly to preserve what can be preserved of the grammar and culture of their few remaining native speakers (Bird, 2009). An NLP model tailored to learning the grammatical structure of a language from small amounts of data could be of great assistance in this documentation task. Shcherbakov et al. (2016) applied data hallucination techniques to produce potential new words in a given language, which could then be given to native speakers to determine whether or not they are real words. Natural disasters often take place, and usually do the greatest damage, in localities where low-resource languages are spoken (Lewis et al., 2011). A translation model capable of quickly learning languages from unlabelled data could potentially save thousands of lives by enabling first responders to single out the most affected areas and target their efforts there.

As with any advance in NLP, there are also potential privacy concerns, but the extension of NLP models to encompass more languages does not carry with it any unique concerns beyond the usual potential pitfalls of invasions of privacy. A more pressing concern is that of data ownership. Language communities are sometimes concerned, and not without reason, that researchers are less interested in preserving their language and culture than with advancing their own academic careers. Others simply view their languages as private matters not to be shared with outsiders. The attempts of outside researchers to study and preserve a culture and language should not be conducted without regard to the wishes of the community to whom that culture and language primarily belong, and it behooves researchers who are looking into an extremely low-resource language to keep this in mind as they search for or attempt to record more data. For more detailed discussion of ethical issues surrounding language documentation and revitalization, we recommend Good (2018).

3 Morphological Inflection

Morphological inflection is the process by which words change form in order to fulfill different grammatical functions. For example, in English, the so-called "dictionary form," or "lemma," of a noun is usually its nominative singular form, i.e. the word *chair*. When used to refer to more than one chair, the lemma *chair* changes into the plural form *chairs,* according to the rules of English morphological inflection. All languages have multiple sets of rules for inflection. *Chair* follows the standard English inflection rules, for instance, but *mouse* does not. The plural of *mouse* is not *mouses,* but *mice:* the change is in the stem of the word rather than in the ending.

Verbs undergo similar changes, and have similarly diverse rules. A standard, or regular verb, in English, might follow the pattern of *love*. The in-

finitive form is *to love,* and the lemma simply *love*. When the verb changes to the third person, it becomes *loves,* and when it changes to the past tense, it becomes *loved*. This past tense form also serves as the participle, or verbal adjective, derived from the verb. Another type of regular verb follows the pattern of *to ride*. Rather than changing only the ending of the verb, *ride* follows a set of rules that change its stem. As a result, while *ride* behaves like *love* in the present tense (i.e. *I ride* becomes *he rides*), in the past tense it becomes *rode* rather than *rided*. Additionally, rather than reusing the past tense form, it has a separate form for the participle, *ridden*.

All languages have these rules of morphological inflection. Some have much more complex sets of rules than English, and include separate forms for a large number of potential nounal and verbal functions. Some verb forms include information about the physical location the object relative to the speaker, as is recorded in Sylak-Glassman (2016). Some noun forms include in the word information like whether or not the object under discussion accompanies another object or not, or whether it is to be avoided, once more documented in Sylak-Glassman (2016). Human beings learn the rules and exceptions of their native language as children. Adults who attempt to learn a new language have far more difficulty. Building a machine learning system which is capable of learning the rules and their exceptions, then discerning which rules to apply in each case and where the exceptions occur, is a problem that has not yet been solved.

The goal of Vylomova et al. (2020) in the SIGMORPHON 2020 Shared Task 0 was to produce systems capable of learning morphological inflection in a wide variety of languages, some of which were available during training, some of which were not. Our research builds off of the baseline provided for the task, but, while we take inspiration for certain research strategies both from the submissions to the shared task and from other papers cited herein, we follow a different direction for our research.

As mentioned in the the introduction, we train our models on selectively augmented data in order to remedy issues caused either by overfitting or by specific types of word being more challenging for the baseline model. Two bilingual models are trained by combining the training data from the two source languages most closely related to the target language, two multilingual models on data from both the related source languages and the target language, and two more broadly multilingual models on a combination of all the source language data, one with the target language data and one without. A number of models are also trained on bilingual data created by adding sufficient data from one source language to the training data of the target language to make up a certain percentage of the file. In some cases these models take only words from the source language for which the word's lemma was also present in the inflected form. In others the words are chosen randomly.

4 Data Collection

We used the language data provided by the SIGMORPHON Task 0. The data provided was split into two groups, the Development Languages and Surprise Languages. The Development Languages were provided at the start of the shared task, and the Surprise Languages were distributed during the final week. This was to challenge the multilingual models, but not to have the models fail. Each language was split into training (70%), development (10%), and testing (20%) sets. During our research we neglected the testing data sets for all of the languages, to help reduce bias and over-training in the models.

Our two target languages, Norwegian Nynorsk and Middle Low German, were distributed as Surprise Languages for the task, although this was not relevant for our own research. From Table 1 we can observe that the two target languages have significantly less data. We selected them specifically mostly because the models submitted to the task did very poorly on them, on average, but also because they are both in the same language family (Germanic), and because there are multiple high-resource languages in the same family. Icelandic and Norwegian Nynorsk are from the same subfamily, North Germanic, and Middle Low German is only approximately five hundred years of linguistic development away from modern German. This made it more likely that augmenting target language data with source language data would produce positive results.

As we can observe, by Table 1, that Norwegian Nynorsk and Middle Low German have drastically lower numbers of available words than our source languages. This leads to a multitude of issues, especially when it comes to utilizing the development data to judge the accuracy of the neural models, as there are only 127 words included in the Middle Low German Development Data set.

Language	Nwords	avg Ntags
nno	14431	2.9
gml	1272	4.4
deu	142008	3.8
eng	115522	2.4
isl	76915	4.1

Table 1: Total language statistics from the SIGMORPHON data sets.

Sigmorphon Task Data: https://github.com/sigmorphon2020/task0-data

5 The Baseline Model

For the Sigmorphon Task 0, the baseline Neural Transducer was distributed to teams for them to utilize and work from. This Neural Transducer includes two training methods, that are used throughout the entirety of all the work done in this paper. The two methods the Sigmorphon task outlines is the transformer model, and the mono-hmm model Vylomova et al. (2020).

In short, the transformer model is a character-level based neural network Vaswani et al. (2017). While the mono-hmm model is a hard monotonic attention system Wu and Cotterell (2019). Both of these models are using the lemma and tags as input, and the inflected form as output.

The task also outlines only one method of training data augmentation, which is referred to as data hallucination Anastasopoulos and Neubig (2019). This is effectively the act of randomly swapping characters in the lemmas, and setting a set amount of words, in this case 10,000 words, to train from. This is supposed to help with low-resource languages, as it gives additional pseudo language data, however when implementing their methods, we noticed mixed results for our two target languages see Table 2.

To get a basis of how the languages would preform we trained four models for each language, two transformer types, and two mono-hmm types. These were each split between the un-altered training data provided by the task, as well as the the hallucinated training data respectively see Table 2. As such, we consider these models to be the baseline for our research.

Language	model	Acc.
nno	transformer	88.7
	mono-hmm	81.3
	hall-transformer	86.3
	hall-mono-hmm	82.6
gnl	transformer	61.4
	mono-hmm	62.2
	hall-transformer	65.4
	hall-mono-hmm	54.3

Table 2: The accuracy of each baseline model for the target languages, of the development data.

5.1 Baseline Results

The results of these models, see table 3, produced drastically higher accuracy than the SIGMORPHON Task for our two target languages, and an accuracy greater than 95% for the source languages. Although the jump in accuracy was to be expected, as the task was focused on creating multilingual models that can perform on a variety of languages, where as these models are trained and tested on only one language. In addition to this, the two target languages chosen, Norwegian Nynorsk and Middle Low German, were both surprise languages. Which during the shared task, were distributed during the last week. This can also help explain why our target languages performed so poorly in the task results Vylomova et al. (2020).

Language	Our Mean Acc.	Task Mean Acc.
nno	84.7	14.0
gml	60.8	11.0
deu	97.8	88.0
eng	96.5	86.0
isl	96.6	84.0

Table 3: The Average Accuracy of the Baseline models for the chosen languages, of the development data.

When breaking these results down into the individual models see Table 2, we notice that the models trained on the hallucinated data had mixed results between the target languages. As such, we will use the average of the four baseline models accuracy as our goal to improve from.

When analyzing these models for a generalized language feature, if the lemma remains in the inflected word unchanged, see section 6.2 for more, we notice that Middle Low German had difficulty with this subset, as well as a low volume of words existing in this subset, see Table 4. However, for Norwegian Nynorsk, the models typically perform better when subsetted, see Table 4. When implementing a training data augmentation method that utilizes this subset of words we observed an improvement in accuracy for both languages, compared to the average baseline accuracy, see section 6.2.

Language	(%) words w/ lemma	model type	Acc. (%) difference
nno	81.7	transformer	+1.2
		mono-hmm	+2.3
		hall-transformer	+0.8
		hall-mono-hmm	+1.9
gml	11.8	transformer	-8.1
		mono-hmm	-2.2
		hall-transformer	-5.4
		hall-mono-hmm	-1.0

Table 4: Difference for each baseline model, when subsetted into words with an unchanged lemma.

6 Training Augmentation

When approaching the problem of improving the accuracy of models built to predict the inflected word for low-resource languages, there are quite a few different strategies to implement. For example, in the Sigmorphon Task 0 they used the method of hallucination to augment the training data. While our results from the baseline model presented proved this method to result in lower accuracy for Norwegian Nynorsk, it did result in an improvement in accuracy for Middle Low German.

Our approach to tackling this problem was to compare different methods of data augmentation, to improve the accuracy of both models. First, we wanted a guide for how well a model that was trained on larger Germanic languages, without seeing any words from our chosen low-resource languages. This is where the Full Data Augmentation, see section 6.1, approach comes in. This method shows how training on large languages of the same family achieved less than 10% accuracy. However with the simple addition of the target languages data when training these models, their accuracy increased significantly, see section 6.1. This served as a general guideline for multilingual models, without any additional steps.

The next method was to implement a language feature specific data augmentation. This was the Lemma training Data Augmentation, see section 6.2, approach that we took. We noticed in the baseline results, see section 5.1 that the models had a difficult time approaching words where the full lemma was unchanged in the inflected word. For example, when inflecting the word chair to chairs, the word 'chair' is fully contained in the word 'chairs'. We targeted this issue specifically for both Norwegian Nynorsk and Middle Low German by augmenting the training data directly. The method was designed to take a target languages full data set and add words from other larger languages, where the condition of the lemma being unchanged in the inflected word was true. The amount of additional language that was added to the training data was another variable we measured with this approach. When implementing this model and comparing the results to the baseline we noticed accuracy improvements in both languages, for nearly every model attempted.

As a control for the lemma augmentation method, we utilized a similar approach. The only difference being the words added were not subsetted, and were chosen randomly. These results proved to be better than the baseline, however not better than the lemma implementations, see section 6.3.

6.1 Full Data Augmentation

Full data augmentation is a process by which a multilingual model is trained on data produced by combining the complete training data sets of multiple languages. Two methods of full data augmentation are included in our research. The first is a simple zero-shot transfer of a model trained on related high-resource languages to the target low-resource languages. The model for Middle Low German trains on English and German, and the model for Norwegian Nynorsk trains on German and Icelandic. An additional model trains on all the source languages and is tested on both target languages. The accuracies of the resulting models, which are very low, are summarised in Table 5.

Sources	Targets	Accuracy
eng, deu	gml	9.055
deu, isl	nno	5.509
eng, deu, isl	gml, nno	5.541

Table 5: The numbers are the average accuracy across the mono-hmm and transformer models, except for all-langs, which was only run as mono-hmm.

This method of data augmentation is clearly not feasible for field application, though analysis of the errors which it produces reveals valuable information about the relationships between source and target languages. For instance, a model trained on English and German and tested on Middle Low German has a disproportionately high number of errors in prefixes and stems, but a very low instance of errors in suffixes. In most languages, the pattern is the opposite, because most Indo-European languages rely heavily on suffix changes in inflection. This result would seem to indicate that Middle Low German either relies more heavily on prefixes and stem changes than the other two languages, or that its inflectional patterns with regard to prefixes and stem changes are fairly different from those of English and German, whereas its patterns with regard to suffixes are not. A graph of the incidence of prefix, stem, and suffix errors classed by model architecture and training data is visible in Figure 1.

The second method of full data augmentation is the addition of target language data to the combined training data from the high-resource source languages, in order to gain some of the benefits of fine-tuning on target data. This improves results across the board, though not to a level competitive with the targeted lemma augmentation models. The error distributions of these models are seen in Figure 2.

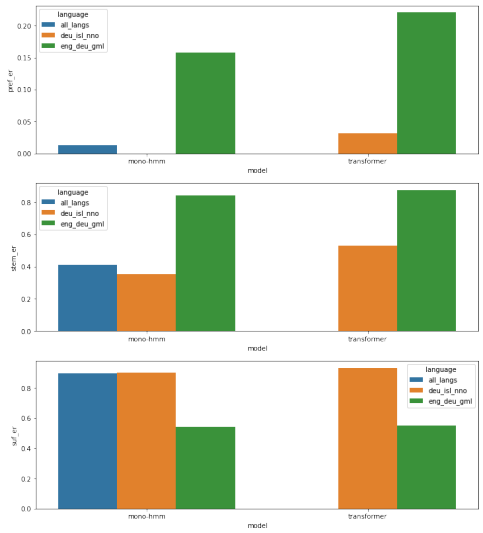

Figure 1: Errors from the multilingual source language models classified as prefix, stem, and suffix errors, divided by model and language. The all-langs model was run only as mono-hmm and not transformer.

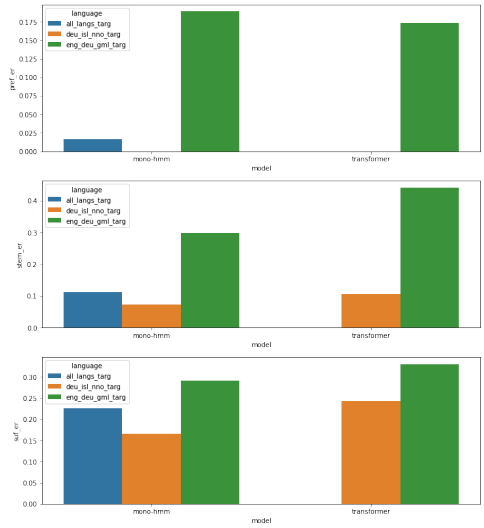

Figure 2: Errors from the multilingual source+target language models classified as prefix, stem, and suffix errors, divided by model and language. Once more, the all-langs model ran only as mono-hmm, not transformer.

6.2 Lemma in Inflected Form

As mentioned in section 6 this method was designed to take all of the data from a target language, Norwegian Nynorsk or Middle Low German. Then add percentages of additional data from a source language of the same family, in our implementation we used German, English, Icelandic. We also attempted to combine Norwegian Nynorsk and Middle Low German with each other and themselves, although this resulted in low amounts of data being added, as these languages already have low amounts of data.

This method is defined by the addition of words from two language data sets, the target language data (L_0), and an additional language (L_1). However, L_1 is a subset of the full language where the lemma appears in the inflected word unchanged. The amount of words used is defined as N_L. Thus, the number of words from L_1 added to L_0 is defined as follows:

$N_{L_1} = L_0 * \frac{percent}{100}$

As such, the number of words added scales proportionally to the target language. The words chosen, from our subset L_1, for the augmented training data set are picked at random, to reduce the bias. This allows us to effectively measure the effect additional language data provides to the final models, among other attributes.

An example of this would be combining Middle Low German with an additional 50% data from English. The amount of training data for Middle Low German was 890 words, as such we added 445 words from the English data set, where the lemma appeared unchanged in the inflected word. This new augmented training data would then be represented as 'gml+50eng'.

6.2.1 Lemma Implementation Results

Overall we trained 62 models for Middle Low German and 46 models for Norwegian Nynorsk, split evenly between transformer and mono-hmm type models, from a mix of the target and source languages, using percentages that ranged from 25% to 200%, when possible. These percentages were not always the same as the data limited itself, due to not enough unique words being available.

When aggregating the results from all the different models and percentages, we observed that the mean accuracy outperformed the baseline for both Norwegian Nynorsk and Middle Low German. Showing that this method could have strong potential for other low-resource languages as well.

Language	Augmented Mean Acc.	Baseline Mean Acc.
nno	86.1	84.7
gml	65.5	60.8

Table 6: Comparison of Lemma Implementation and Baseline models, using the development data.

As shown Middle Low German seemed to get the biggest rise in accuracy, +4.7%. While Norwegian Nynorsk had an increase of +1.4%. However, looking at the full breakdown of models in table 10, 11, we notice that while Middle Low German produces the best accuracy overall, Norwegian Nynorsk did not beat the best baseline model. None-the-less,

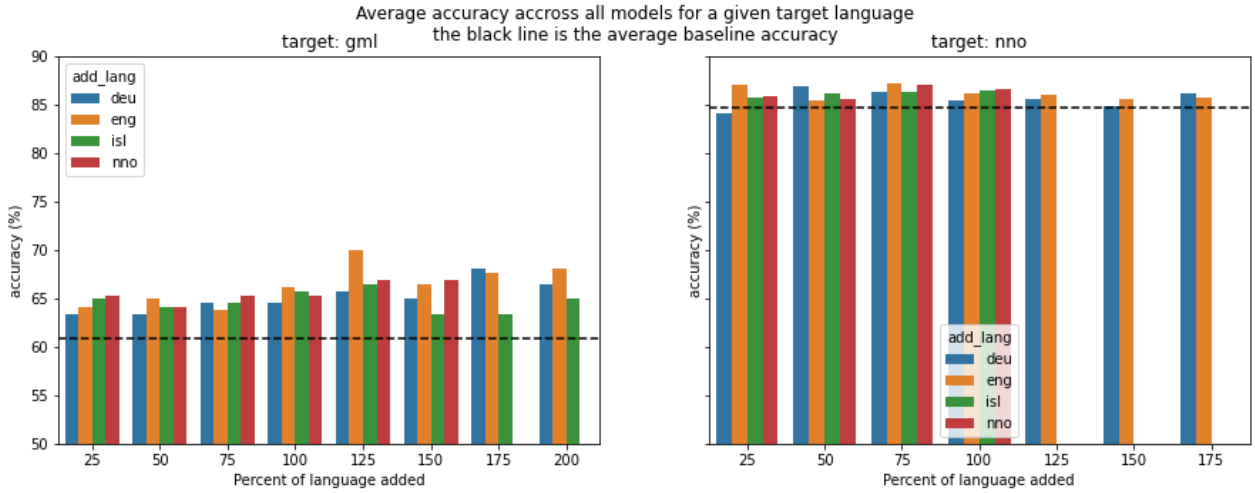

Figure 3: Accuracies are binned to multiples of 25%, this is using the Lemma Implementation and development data.

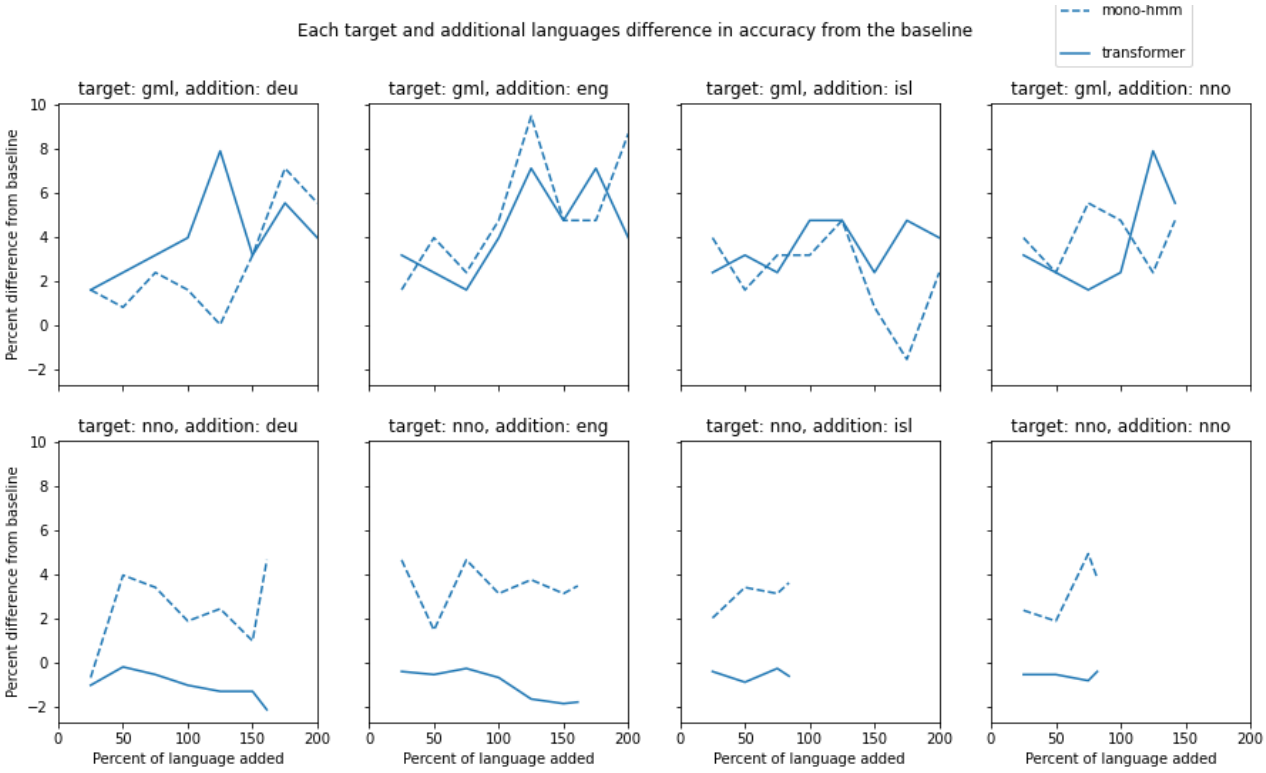

Figure 4: The difference is using the average target baseline, it is using the Lemma Implementation and development data.

on average the models outperformed the baseline drastically.

Referencing table 7, we notice that the best language pairs for Norwegian Nynorsk is itself, with the runner up being Icelandic. For Middle Low German we see that being paired with English produced the best results.

Next lets dive into the differences each language has when looking at only the percent of language added, these groupings were paired together as some language pairs could not get exact multiples of 25%, for example, only 82% of Norwegian Nynorsk could be added together. The full table of the exact broken down results is Table 10, 11.

The table 8 suggests that the most optimal percentage of words to add is 51 - 75% for Norwegian

Source Language	Additional Language	Augmented Mean Acc.
nno	deu	85.6
	eng	86.20
	isl	86.24
	nno	86.3
gnl	deu	65.2
	eng	66.4
	isl	64.7
	nno	65.7

Table 7: Based on the Lemma Implementation using the development data.

Source Language	Percent Added	Augmented Mean Acc.
nno	0 - 25%	85.7
	26 - 50%	86.1
	51 - 75%	87.8
	76 - 100%	86.2
	101 - 125%	85.8
	126 - 150%	85.2
	151 - 175%	86.0
	176 - 200%	NA
gnl	0 - 25	64.5
	26 - 50%	64.2
	51 - 75%	64.6
	76 - 100%	65.5
	101 - 125%	67.3
	126 - 150%	65.5
	151 - 175%	66.4
	176 - 200%	66.5

Table 8: Mean accuracy for percentage of language added to target, based on the Lemma Implementation and using the development data.

Nynorsk and 101 - 125% for Middle Low German. Although, in general we notice small variations in the accuracy based on the percentage added, and these variations vary more drastically when looking at the individual language pairings based on the percentage added. This should only serve as a rough guide, and not absolute fact.

To find the most optimal pairing, we should look at a full breakdown of the models, see Table 10, 11. As well as the table, there is an included bar graph of the binned percentages 3. From this we can see generally which languages perform best, and at what percentage do they perform best. Furthermore, to get the most detailed look at how these models performed, we should look at the difference of accuracy, from the baseline, for each language pairing and model type, for each amount of language added. This breakdown is displayed in Figure 4. This figure shows a stark difference between model types. For Norwegian Nynorsk we consistently see the mono-hmm model type under perform, and in most cases get an accuracy lower than the baselines average. However, for Middle Low German, we see that the model type tends to provide mixed results.

In general, we can observe that the simple addition of another same-family language can provide boosts in accuracy to the target language, however the best percentage of the target language to add can vary depending on the additional language, and the model type. At the end, the model that performed the best for this Lemma Data Augmentation implementation were the mono-hmm type model 'gml+125eng' with an accuracy of 71.7%, and the transformer type model 'nno+50deu' with an accuracy of 88.5%.

6.3 Random Implementation

The Random Implementation was similar in regards to the Lemma Implementations, see section 6.2, except for the words being added to the target language (L_0). The added words from (L_1) is not a subset, i.e. they can be any word in that language, and they are chosen at random. The percentage scales from the target language (L_0), just as it does in the Lemma Implementations.

This experiment is used a control for the Lemma Implementations, to see just how well it matches up. In total 64 Middle Low German models were trained, split evenly between the transformer and mono-hmm model types, ranging from 25% - 200% of the target language added. While 56 models were trained for Norwegian Nynorsk with the same setup as Middle Low German. All of the models will not be displayed, as the individual match ups did not outperform the Lemma Implementations.

Language	Augmented Mean Acc.	Lemma Mean Acc.
nno	86.0	86.1
gml	64.9	65.5

Table 9: Comparison of Random Implementation and Lemma models, using the development data.

However, there are a few things to note. Table 9 shows that the Random Implementation is close to the lemma Implementations, however it under-performs in every aspect when broken down. Although it under-performs, there are still useful insights. For example, we observed that the best language pairings are the same as the Lemma Implementations, see section 6.2.1, for both target languages. Which is Norwegian Nynorsk and itself, with the runner up being Icelandic, while for Middle Low German the best paired language was English. As these statistics are the same, one could

assume that training these pairs of languages together would produce the best results.

Another interesting conclusion is the percentage of language to add, based on a scale of the target language, see section 6.2. For Norwegian Nynorsk we again confirm that 51 - 75% had the highest average accuracy, however, for Middle Low German we observed that 176 - 200% has the best accuracy, with the close runner up being tied for 101 - 125% and 151 - 175%, very similar to what was observed with the Lemma Implementations, see section 6.2.1.

7 Discussion

The results of our research conclude that Low-Resource languages can be utilized in Morphological Inflection tasks, with some additional steps. We observed that models trained purely on a low-resource language, see section 5, can produce significant results on those target languages. However, with the help of additional languages a model can improve the target language accuracy.

The Baseline transformer model for Norwegian Nynorsk had the best accuracy (88.7%) than any other Norwegian Nynorsk model produced. However, this model would perform poorly on other languages besides Norwegian Nynorsk. With the addition of other languages, via Lemma Implementations section 6.2 and Random Implementations section 6.3, we can observe that improvements can be made through the use of simple targeted language feature engineering for the training data. Through these training data augmentations we observed the most optimal pairs of languages tend to be Norwegian Nynorsk and itself, and Middle Low German and English. We Also observed that Norwegian Nynorsk had the best results with a 51 - 75% additional language and Middle Low German was roughly 101 - 125%, see section 6.2.1, 6.3.

Another experiment we performed was a zero-shot transfer for Full Data Augmentation, see section 6.1. This method showed that a model with no language from a low-resource target language will perform poorly. However, through the addition of another high-resource language, from the same family, we see big accuracy improvements. Although not as accurate as the Lemma and Random Implementations, this could prove useful if the goal is to maximize a multilingual model.

To test if these models are over-fitting, we utilized the testing data at the end of our research. All of the models only drop in accuracy by about 0 - 4%, suggesting that these models might be slightly over-fitting. This data was used to perform a set theory analysis and is referenced in Figure 5, 6. The first figure 5 shows the number of correct predictions for each word index. This can help us identify clustering in the word predictions, such as the beginning and end of Middle Low German. Making this simpler, it becomes apparent there are certain words that do well with every model ran. Looking at figure 6 we notice there tends to be a higher percentage of words that have variability in Middle Low German (70 - 80%), as opposed to Norwegian Nynorsk which has roughly 30 - 48% variability in its words. We also notice a portion of words that were incorrectly predicted by every model. This could indicate there might be a few words that are impossible for any model to correctly predicted, due to limited data, or this might be due to chance.

We hope this work will encourage other researchers to explore the difficulties of low-resource languages. There is still a lot of ground to cover in this area, for example, can we improve the accuracy of a low-resource language, utilizing training data augmentation methods, for massively multilingual models? We have showed that this technique can be utilized for small targeted models, however with the nature of low-resource languages, a single language targeted model may not serve as much benefit as a multilingual model.

8 Conclusion

The results of our research show that targeted language training data augmentation can boost the accuracy of low-resource languages, specifically in Middle Low German and Norwegian Nynorsk. We showed that the addition of data from other languages in the same family to a model's training data can help guide a model into finding the inflectional patterns of a lower-resource language. The average accuracy from the SIGMORPHON Task 0 for our target languages were: Middle Low German (11%) to our best accuracy of 71.7% with the Lemma mono-hmm model 'gml+125eng', see section 6.2.1, as well as the baseline transformer Norwegian Nynorsk Model with an accuracy of 88.7%, see section 5.1.

We also showed that targeted general language features can prove more successful in model improvements, as opposed to a random approach. We hope this inspires other researchers to explore other low-resource languages, methods, and approaches to tackling the problem of Morphological Inflection in low-resource languages.

References

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection.

Ross Andersen. 2020. The panopticon is already here. *The Atlantic*.

Steven Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35:469–474.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Jeff Good. 2018. Ethics in language documentation and revitalization. In Kenneth L. Rehg and Lyle Campbell, editors, *The Oxford Handbook of Endangered Languages*, pages 419–440. Oxford University Press, Oxford, GBR.

William Lewis, Robert Munro, and Stephan Vogel. 2011. Crisis MT: Developing a cookbook for MT in crisis situations. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 501–511, Edinburgh, Scotland. Association for Computational Linguistics.

Andrei Shcherbakov, Ekaterina Vylomova, and Nick Thieberger. 2016. Phonotactic modeling of extremely low resource languages. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 84–93, Melbourne, Australia.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). Technical report, The UniMorph Project, Baltimore, Maryland.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction.

Source Language	Percent Added	Additional Language	Model Type	Acc. (%)
nno	25%	deu	transformer	87.7
			mono-hmm	80.6
		eng	transformer	88.3
			mono-hmm	85.9
		isl	transformer	88.3
			mono-hmm	83.3
		nno	transformer	88.1
			mono-hmm	83.6
	50%	deu	transformer	88.5
			mono-hmm	85.2
		eng	transformer	88.1
			mono-hmm	82.7
		isl	transformer	87.8
			mono-hmm	84.7
		nno	transformer	88.1
			mono-hmm	83.2
	75%	deu	transformer	88.1
			mono-hmm	84.7
		eng	transformer	88.4
			mono-hmm	85.9
		isl	transformer	88.4
			mono-hmm	84.4
		nno	transformer	87.9
			mono-hmm	86.2
	82%	nno	transformer	88.3
			mono-hmm	85.1
	84%	isl	transformer	88.1
			mono-hmm	84.9
	100%	deu	transformer	87.7
			mono-hmm	83.2
		eng	transformer	88.0
			mono-hmm	84.4
	125%	deu	transformer	87.4
			mono-hmm	83.7
		eng	transformer	87.0
			mono-hmm	85.0
	150%	deu	transformer	87.4
			mono-hmm	82.3
		eng	transformer	86.8
			mono-hmm	84.4
	161%	deu	transformer	86.6
			mono-hmm	85.9
		eng	transformer	86.9
			mono-hmm	84.8

Table 10: Full NNO table of every model trained and tested using the Lemma Implementations, section 6.2 and the development data.

Source Language	Percent Added	Additional Language	Model Type	Acc. (%)
gml	25%	deu	transformer	63.0
			mono-hmm	63.8
		eng	transformer	64.6
			mono-hmm	63.8
		isl	transformer	63.8
			mono-hmm	66.1
		nno	transformer	64.6
			mono-hmm	66.1
	50%	deu	transformer	63.8
			mono-hmm	63.0
		eng	transformer	63.8
			mono-hmm	66.1
		isl	transformer	64.6
			mono-hmm	63.8
		nno	transformer	63.8
			mono-hmm	64.6
	75%	deu	transformer	64.6
			mono-hmm	64.6
		eng	transformer	63.0
			mono-hmm	64.6
		isl	transformer	63.8
			mono-hmm	65.4
		nno	transformer	63.0
			mono-hmm	67.7
	100%	deu	transformer	65.4
			mono-hmm	63.8
		eng	transformer	65.4
			mono-hmm	66.9
		isl	transformer	66.1
			mono-hmm	65.4
		nno	transformer	63.8
			mono-hmm	66.9
	125%	deu	transformer	69.3
			mono-hmm	62.2
		eng	transformer	68.5
			mono-hmm	71.7
		isl	transformer	66.1
			mono-hmm	66.9
		nno	transformer	69.3
			mono-hmm	64.6
	142%	nno	transformer	66.9
			mono-hmm	66.9
	150%	deu	transformer	64.6
			mono-hmm	65.4
		eng	transformer	66.1
			mono-hmm	66.9
		isl	transformer	63.8
			mono-hmm	63.0
	175%	deu	transformer	66.9
			mono-hmm	69.3
		eng	transformer	68.5
			mono-hmm	66.9
		isl	transformer	66.1
			mono-hmm	60.6
	200%	deu	transformer	65.4
			mono-hmm	67.7
		eng	transformer	65.4
			mono-hmm	70.9
		isl	transformer	65.4
			mono-hmm	64.6

Table 11: Full GML table of every model trained and tested using the Lemma Implementations, section 6.2 and the development data.

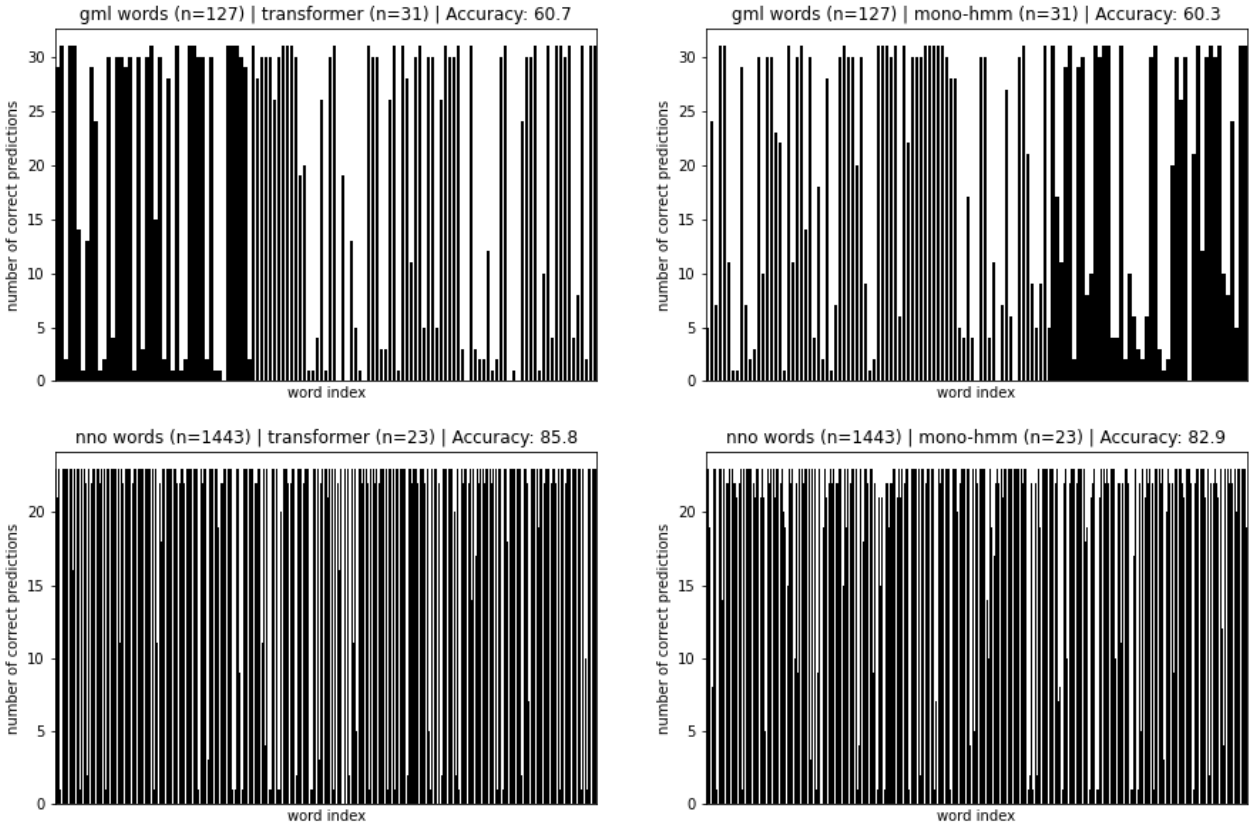

Figure 5: The raw prediticions of Lemma Implementation models, utilizing the testing data.This shows the words that are predicted correctly predicted by N models.

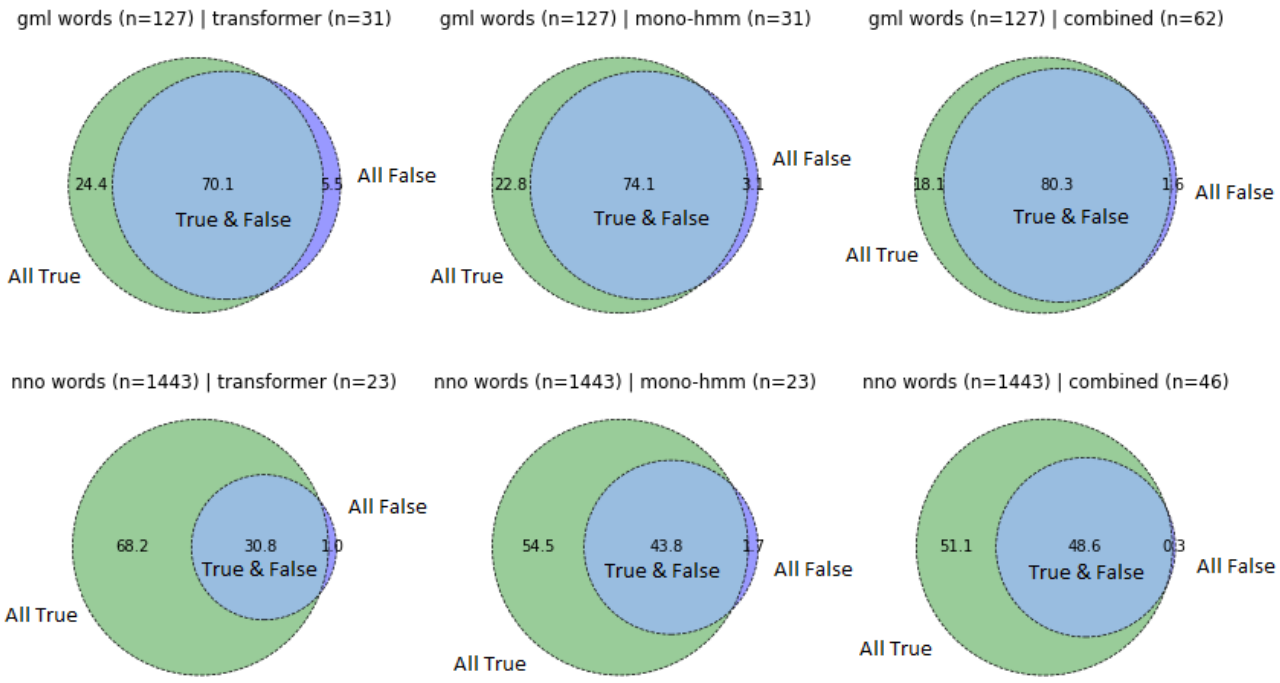

Figure 6: The sets of Lemma Implementation models, utilizing the testing data. This shows the percentage of words that are predicted All True and All False by every model.