Applying Backtranslation to Enhance Morphological Inflection Systems in the Danish Language

Ignacio Talavera Cepeda
Euskal Herriko Unibertsitatea
italavera002@ikasle.ehu.eus

Abstract

Backtranslation is the technique that, in the context of NLP, refers to the process of processing a text from its translated version back to its original language. This technique can be used to create synthetic pairs of words that can be used on parallel-language datasets. Taking inspiration from Machine Translation systems, we propose a backtranslation technique in the context of morphology inflection. By creating systems that, given a lemma and an inflected word, are able to obtain the morphological tags in the UniMorph schema referring to the morphological processes taking place, we can leverage datasets used for morphological inflection systems, in which the objective is to obtain the inflected word given the lemma and the tags. In the project, we show that morphological tagging systems are easy to train in low-data scenarios, and can be used to obtain significant improvements in morphological inflection systems. We also create morphological tagging and inflection systems in the Danish language, testing scenarios of high and low data availability.

1 Introduction

Backtranslation is a technique for improving Neural Machine Translation systems (NMT) to obtain state-of-the-art performance with only parallel training data. It consists of translating a sentence from one language to another and translating it back to the original language. This resulting backtranslated sentence can be used to identify discrepancies and improve the underlying models, as new, synthetic data is generated and can be used for fine-tuning. Backtranslation is especially beneficial in scenarios where parallel data is scarce for training, as it allows researchers to generate additional training examples (Sennrich et al., 2016).

However, backtranslation techniques can be also very helpful in computational morphology, especially for scenarios where morphological data is scarce. Instead of using it over parallel data in a pair of languages, it can be used to obtain the inflected morphological form of the lemma of a word and the morphological tags of that inflected word.

In this paper, the use of backtranslation techniques for improving morphological inflection systems is explored, in the context of Nordic Languages, focusing on Danish. We explore architectures that combine several neural networks to create pipelines in which the same input data is used to train an inflection system and a morphological tagger. This morphological tagger is then used to obtain labelled data for the inflection system to retrain and try to enhance its performance. You can see a visualization of this workflow in Figure 1.

The main contributions of this paper are the following:

- we show that we can improve morphological inflection systems by training morphological taggers that can label new training examples.
- we compare this semi-supervised learning scenario against a fully-supervised learning scenario, where data scarcity is not a problem.
- we successfully create a pipeline that not only trains an inflectional model but uses the same data to train a morphological tagger that can be used to leverage the first model, allowing for a loop of further fine-tuning of the model to push performance.
- we create both morphological inflection systems and morphological tagging systems that can be used for real use cases in the Danish language.

2 Related Work

The Transformer architecture (Vaswani et al., 2017), a family of powerful language models that

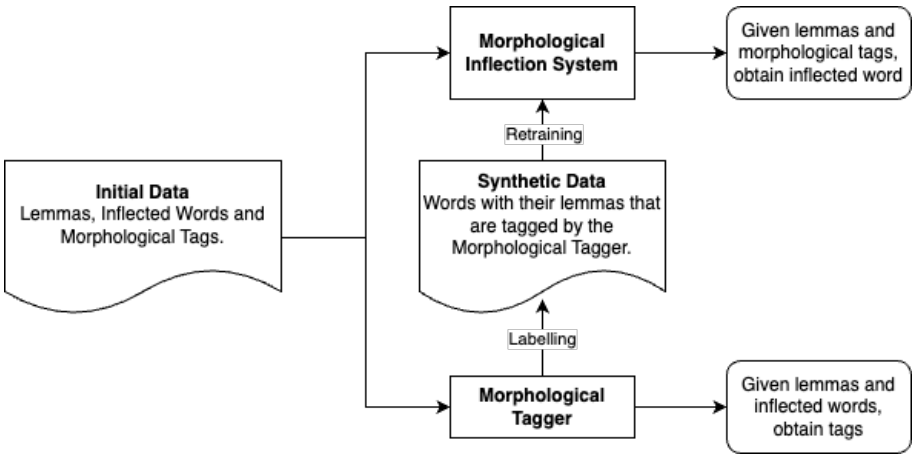


Figure 1: Schema of the Backtranslation system implemented in this paper. From our initial data, we train two different models, the Morphological Inflection System, and the Morphological Tagger. This last model can label data for retraining the Inflection System.

can be scaled up by billions of parameters and can be pre-trained with high amounts of text (Devlin et al., 2018; Liu et al., 2019; Sanh et al., 2019) to be later fine-tuned for more specific tasks, has led to state-of-the-art results in many NLP tasks, making the transfer learning approach, where this fine-tuning onto certain downstream task with a much smaller amount of annotated, specific data has become the standard in the industry. One of these NLP tasks is morphological inflection.

Morphological inflection is the NLP task where, given a sequence x corresponding to a word lemma and the morpho-syntactic features of a morphological inflection process, the NLP models have the goal of finding a mapping to obtain a sequence y corresponding to the morphological inflected version of the lemma, given the morpho-syntactic features (Makarov et al., 2017).

This project gets the inspiration for its main technique from classic backtranslation architectures. Backtranslation techniques take advantage of monolingual data that is *a priori* unsuited for Neural Machine Translation systems, that feed on parallel data for training. Using backtranslation this monolingual data can be transformed into pairs of bilingual, parallel data, that can be later treated as additional training data. This constitutes a technique of semi-supervised learning, where synthetic data is being used alongside human-labelled data (Sennrich et al., 2016).

Semi-supervised learning is a type of machine learning that combines elements of supervised and unsupervised learning. In semi-supervised learning, the model is trained on a mix of labelled and unlabelled data. The idea behind this technique is to leverage large amounts of unlabelled data to improve the performance of the model, and this is especially useful in scenarios where there is a limitation in the amount of training data available. The main goals of this approach are to reduce the dependency on labelled data, the ability to leverage large amounts of unlabelled data and the possibility to improve the accuracy of state-of-the-art systems. However, mixing labelled and unlabelled data is required to ensure the quality of this new data, and it is computationally more expensive than supervised learning, as usually more pipelines are used to process the unlabelled data and introduce it to the training datasets (Learning, 2006).

The UniMorph schema (Sylak-Glassman et al., 2015) is used in NLP to represent morphological information about words, providing a standardised way of representing the internal structure of words across different languages. It consists of three main components: **morphemes**, which are the smallest units of meaning in a word, **inflectional features**, which describe how these morphemes change their form to indicate grammatical function such as tense, number, or gender, and **derivational features**, which describe how a morpheme contributes to the overall meaning of a word. Each inflected word in any given language is represented by its lemma and a set of different UniMorph features.

3 Data Exploration

This section explores the data corpus used for training the different NLP pipelines, both the

morphological taggers and the morphological inflection systems. The main systems have been trained for the Danish language with data from CoNLL–SIGMORPHON 2017 Shared Task[1] (Cotterell et al., 2017).

Danish is a language that belongs to the East Scandinavian branch and is spoken by approximately six million people, primarily in Denmark. It is characterised by a complex phonological system (especially for its distinctive vowel inventory, including monophthongs and diphthongs, with vowel length distinctions playing a crucial role in lexical and grammatical contrasts) and a rich inflectional morphology. The language follows a subject-verb-object (SVO) order and relatively free word order, due to its strong reliance on inflection for grammatical relations. Danish underwent significant linguistic reforms during the 19th century, including spelling standardisation and simplification of grammatical forms, so data collected before is not aligned with the actual use of the language (Basbøll, 2005; Heltoft and Hansen, 2007; Fischer-Jørgensen, 1994).

The data made available for CoNLL–SIGMORPHON 2017 is highly multilingual, spanning 52 unique languages. For the majority of the languages, including the Scandinavian languages, the data comes from the English edition of Wiktionary[2], a large multi-lingual open-source repository containing morphological paradigms for many lemmas.

The format of the data consists of triplets of the form *(lemma, inflected form, morphological tags)*. We can see some examples of the dataset in Table 1. This data is coded following the UniMorph schema (Sylak-Glassman et al., 2015), and was divided into two tasks: the first task is focused on obtaining the inflected form from the lemma with sparse training data; the second task tackling the paradigm cell filling problem (Ackerman et al., 2009), which requires predicting many inflections of the same lemma. The training data for both tasks followed the same schema so that it could be merged into one single dataset.

This dataset contains 11203 records. Once merged, we perform a data exploration phase, in which we focus on extracting information about the frequency and distribution of lemmas and tags. The main conclusions are shown in Figure 2. There are 16 different morphological tags, highly unbalanced in terms of frequency distribution. There is also common repetition of lemmas, as the box plot shows. On average, lemmas appear 3 times, with some outliers lemmas being present 8 times or more.

Even though there is a high class imbalance which must be taken into consideration (in the usage of micro-averaged F1 score when possible, for example), initially it seems like there is enough data to create a supervised model able to do morphological inflection.

4 Methodology

In this section, we detail our process to create the different NLP pipelines that can perform the morphological inflection and the morphological tagging tasks used to test backtranslation as a method to leverage morphological inflection.

There are two different types of models presented in this project: *morphological inflection systems* and *morphological taggers*.

Morphological inflection systems Morphological inflection systems are models trained with the sequence modelling toolkit *Fairseq* (Ott et al., 2019). The input for these models is a combination of the columns *Lemma* and *Morphological Tags* from the dataset, combined internally by *Fairseq*, and the output is the resulting inflected word from the lemma and the morphological tags given, as seen in column *Inflected Form* of Table 1. All systems belonging to this category in this project are named with Greek upper-case letters, starting from A, in ascending order. The metric used to evaluate these systems is accuracy, which is a standard in CoNLL–SIGMORPHON (Cotterell et al., 2016, 2018), alongside Levenshtein's distance, a measure of the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into another (Levenshtein, 1965).

Morphological tagging systems Morphological tagging systems are models trained using Hugging Face Transformers (Wolf et al., 2020), a state-of-the-art platform for fine-tuning Transformer models. The foundation model chosen for all these systems in the project is *distilbert-base-uncased*[3] (Sanh et al., 2019). The input for this model is a combination of the columns *Lemma* and *Inflected*

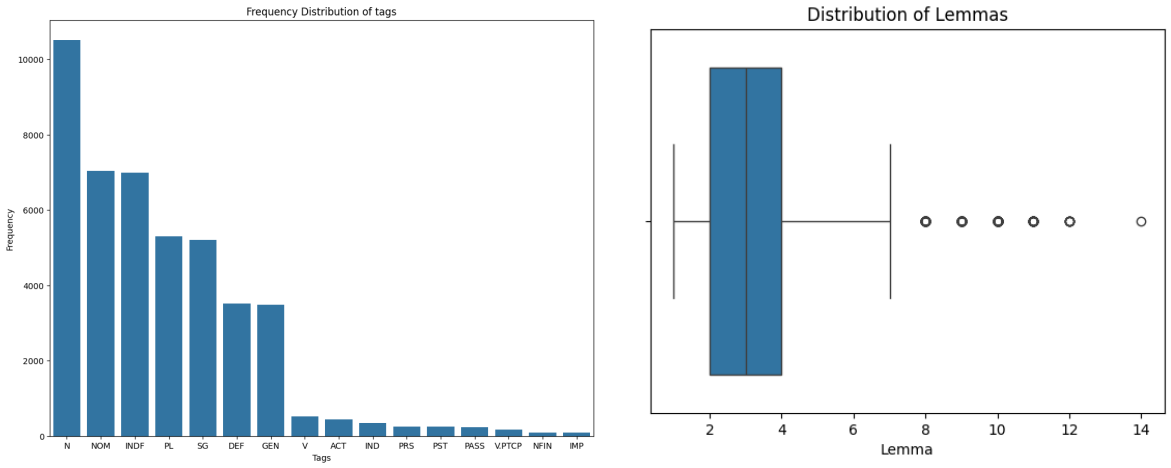
[1] https://github.com/sigmorphon/conll2017

[2] https://en.wiktionary.org/

[3] https://huggingface.co/distilbert-base-uncased

Lemma	Inflected Form	Morphological Tags
kule	kulernes	N;DEF;NOM;PL
bangebuks	bangebuks	N;INDF;NOM;SG
affaldsstof	affaldsstofs	N;INDF;GEN;SG
fosfolipid	fosfolipid	N;INDF;NOM;SG
jubilæum	jubilæums	N;INDF;GEN;SG

Table 1: Example of records from the Danish datasets of the corpus provided by CoNLL–SIGMORPHON 2017. Each record is a triplet consisting of the lemma, the inflected form of the lemma and the morphological tags responsible for the transformation from the lemma to the inflected form.



(a) Frequency distribution of morphological tags.

(b) Box plot showing the distribution of lemmas.

Figure 2: Results from the data analysis over the Danish dataset, showing a column chart with the frequency distribution of morphological tags and a box plot showing the distribution of lemmas.

Form from the dataset (see Table 1), which is combined into a single string, separated by the [SEP] token, and passed to the tokenized, and outputs predictions similar to *Morphological Tags*, in one-hot-encoding format. All systems belonging to this category in this project are named with Greek lowercase letters, starting from ω, in descending order. As a multilabel classification problem, these systems will be evaluated using accuracy and micro-averaged F1 Score[4].

4.1 Pipelines with full training data

As we see in Section 3, we have a data corpus of 11203 rows, which a priori is a sufficient amount to perform supervised learning techniques and obtain a pipeline capable of performing morphological inflection, which is the ultimate goal of this project. To test if this amount of data is enough, and to have models to which we can compare the results obtained with further pipelines that apply morphological tagging to create new records that can be used for training other morphological inflection systems, we propose the first two models:

- A model: Morphological inflection system trained using the whole data corpus at our disposal.
- ω model: Morphological tagging system trained using the whole dataset at our disposal.

Both models are trained with the whole data corpus provided by CoNLL–SIGMORPHON 2017 Shared Task[5] (Cotterell et al., 2017), which corresponds to 11203 fully annotated as we see in Section 1, with a 70% training split, a 15% development split and a 15% test split. A model was trained following the Transformers-based encoder-decoder architecture architecture, with a learning rate of 0.001, a dropout value of 0.3 and RELU as the activation function; *Fairseq* default parameters,

[4]F1 score is a metric that calculates the harmonic mean of precision and recall across all classes in a multi-class classification problem, treating the entire dataset as a single aggregate.

[5]https://github.com/sigmorphon/conll2017

during 90 epochs. The ω model is the result of fine-tuning *distilbert-base-uncased* with a learning rate of 2e-5, a weight decay value of 0.01 and a batch size of 16 data points, during 7 epochs. In both cases, the models were trained until the validation or test metrics stopped growing.

4.2 Pipelines with limited training data

After defining the pipelines trained with the full data corpus, we need systems trained with limited data, which we can compare to A and ω. These systems are expected to be less accurate, given the limited data scenario. Still, they should be effective enough on their own to be able to improve when we use the morphological tagger obtained in this section to tag more data that can be used to train an enhanced version of the morphological inflection system. We train this data using a subset of 1000 instances of the available corpus in Danish. The two models proposed of this type are:

- B model: Morphological inflection system trained using limited data.
- ψ model: Morphological tagging system trained using limited data.

To be able to compare the limited data scenario with their previous models, the same training hyperparameters have been maintained for the morphological inflection system and the morphological tagging system, with the only exception of the training epochs, which varied in each training. For B, 222 epochs were needed to reach convergence, and for ψ, 13. These values are also obtained thanks to the validation metrics.

4.3 Backtranslation-based system

If our thesis is correct, we can improve B model by using ψ to, given tuples of lemmas and inflected words, obtain their morphological tags. Once we obtain triplets of lemmas, inflected words and morphological tags, we can add it to the data available in the limited training data scenario, and retrain a new version of B that should perform closer to the initial model A. We call the system implementing this approach *backtranslation-based system*, and the model performing it is Γ. The data upon which we perform backtranslation is another subset of 1000 records from the initial data corpus, not overlapping the subset used for training B and ψ. To be able to make a proper comparison with this final morphological inflection system, again, the conditions of training have been maintained, with the only exception of using more training epochs, as the training needed more to reach convergence. 220 epochs were needed for convergence.

5 Results

The models proposed in Section 4 were trained as described, obtaining 5 models that were tested using their test splits. Test results of the models are shown in Tables 2 and 3.

Model	Accuracy	Levenshtein's distance
A	0.83	0.44
B	0.62	1.38
Γ	0.73	0.81

Table 2: Results of morphological inflection systems trained for the Danish language.

Model	Accuracy	F1 Score
ω	0.92	0.98
ψ	0.9	0.95

Table 3: Results of morphological tagging systems trained for the Danish language.

Table 2 shows the morphological inflection systems, A, B and Γ, compared using the test accuracy and Levenshtein's distance. By comparing the accuracy, we can see that the model trained with all the available data shows a higher accuracy, correctly classifying 83% of the test instances. B, trained with only 1000 instances, obtains a test accuracy of 0.62, and Γ, trained with these 1000 instances and another 1000 instances, manually tagged by the model, shows an improvement over B, with 0.73 test accuracy. The same behaviour can be observed in Levenshtein's distances, which show the average number of insertions, deletions and modifications of characters that need to be done to get from the predicted word to the ground truth. The reinflected model performs better than B, not obtaining a value as high as A.

On the other hand, Table 3 shows the two morphological taggers. ω can be seen as the benchmark, as it was trained using the whole data corpus available with the only purpose of being compared with ψ, which was used to label the data included in the training dataset of Γ. Here, we can see almost no difference, showing close values in terms of accuracy and micro-averaged F1 score. This means that

limited data is sufficient to finetune a DistilBERT model into this multilabel classification task.

A summary of all the models obtained, along with their type, the amount of training data and epochs they were trained, and their test metrics, can be seen in Table 4.

Given these results, we can state that backtranslation is a technique that improves the inflection capabilities of Transformer-based models in scenarios of low availability of data. The backtranslation-based model, while being trained with only 2000 data instances (18% of the total dataset, 1000 being from the original data corpus, and the other 1000 being obtained by the morphological tagger ψ from lemmas and inflected words), can reach 87.95% of the total accuracy obtained by the model trained with the whole data corpus. We can also state that the morphological tagging models, obtained by fine-tuning DistilBERT instances, can obtain similar results trained with both the whole dataset and just a subset of 1000 instances. With only 1000 instances, we have been able to train a morphological tagger capable of labelling data later used to obtain the improvement seen in the morphological inflection systems.

6 Conclusions

After carrying out these experiments and obtaining the results, we conclude that the thesis stated at the beginning of the project is correct, and backtranslation is a method that effectively improves the performance of a morphological inflection NLP system by labelling pairs of lemmas and inflected word, making them ready to be added as more training data to perform a retraining or a fine-tuning of the model.

This backtranslation-based process, inspired by the backtranslation technique used for improving Machine Translation systems that feed on parallel data, consists of a pipeline of just two elements, as shown in Figure 1: a *morphological tagger*, that can label lemma-inflected word pairs, and a *morphological inflection system*, that, given a lemma and a set of morphological tags, can obtain the inflected word.

Thanks to the high amount of available data in the Danish language, we were able to create benchmark models that were used to evaluate the obtained models, both the morphological inflection system before and after backtranslation and the morphological tagger under the low-data scenario. These benchmark models showed the improvement that backtranslation-based systems made on the morphological inflection system, while also showing that a scenario with high data availability is preferable. In the case of the morphological tagger, it was shown that the model trained with 10% of the data had a performance close to the model trained with the whole data corpus at our disposal.

7 Future Work

Given the similarities between some languages, and the effectiveness of multilingual pipelines, especially in cases where data is scarce, like the Basque language (Urbizu et al., 2022), experimenting with multilingual pipelines could enhance morphological inflection. Danish is very similar to Norwegian and Swedish, to the point that speakers of one language can maintain conversations with speakers of the other languages, so trying a multilingual pipeline that combines data from the three languages could make a system capable of doing morphological inflection in any Nordic language (Haugen and Borin, 2018).

Another way of expanding the work done in this project could be by focusing on the hyperparameter optimisation process or trying a more powerful foundation model. The goal of this project was to show that there was an improvement by using this technique, and thus most of the hyperparameters were as static as possible, properly researching different Large Language Models could result in models that surpass the test metrics of both A and Γ models.

References

Farrell Ackerman, James P Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. *Analogy in grammar: Form and acquisition*, 54:82.

Hans Basbøll. 2005. *The phonology of Danish*. OUP Oxford.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Model	Type	Training Data	Training Epochs	Test Accuracy	Test LD	Test F1
A	Inflection System	11203 instances	90	0.83	0.44	
B	Inflection System	1000 instances	222	0.62	1.38	
Γ	Inflection System, trained with backtranslation	1000 labelled instances + 1000 synthetic instances	220	0.73	0.81	
ω	Tagger	11203 instances	7	0.92		0.98
Ψ	Tagger	1000 instances	13	0.9		0.95

Table 4: Summary of all the models developed for the project, with their type, trainng data and test metrics. LD refers to Levenshtein's distance. Inflection system are not evaluated using F1 score, and taggers are not evaluated using Levenshtein's distance.

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.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The sigmorphon 2016 shared task—morphological reinflection. In *Special Interest Group on Computational Morphology and Phonology Workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Eli Fischer-Jørgensen. 1994. Fonetik og fonologi. *NyS, Nydanske Sprogstudier*, (19):37–56.

Einar Haugen and Lars Borin. 2018. Danish, norwegian and swedish. In *The world's major languages*, pages 127–150. Routledge.

Lars Heltoft and Erik Hansen. 2007. Grammatik over det danske sprog. *Ny forskning i grammatik*, (1).

Semi-Supervised Learning. 2006. Semi-supervised learning. *CSZ2006. html*.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Peter Makarov, Tatiana Ruzsics, and Simon Clematide. 2017. Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 49–57, Vancouver. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *Systems and Frameworks for Computational Morphology: Fourth International Workshop, SFCM 2015, Stuttgart, Germany, September 17-18, 2015. Proceedings 4*, pages 72–93. Springer.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.