Using Graph-Based Methods to Augment Online Dictionaries of Endangered Languages

Khalid Alnajjar[1,2,3], Mika Hämäläinen[1,2,3], Niko Partanen[1] and Jack Rueter[1]
[1]University of Helsinki, Finland
[2]École Normale Supérieure & CNRS, France
[3]Rootroo Ltd, Finland
firstname.lastname@helsinki.fi

Abstract

Many endangered Uralic languages have multilingual machine readable dictionaries saved in an XML format. However, the dictionaries cover translations very inconsistently between language pairs, for instance, the Livonian dictionary has some translations to Finnish, Latvian and Estonian, and the Komi-Zyrian dictionary has some translations to Finnish, English and Russian. We utilize graph-based approaches to augment such dictionaries by predicting new translations to existing and new languages based on different dictionaries for endangered languages and Wiktionaries. Our study focuses on the lexical resources for Komi-Zyrian (kpv), Erzya (myv) and Livonian (liv). We evaluate our approach by human judges fluent in the three endangered languages in question. Based on the evaluation, the method predicted good or acceptable translations 77% of the time. Furthermore, we train a neural prediction model to predict the quality of the automatically predicted translations with an 81% accuracy. The resulting extensions to the dictionaries are made available on the online dictionary platform used by the speakers of these languages.

1 Introduction

For many endangered languages there are several existing dictionaries and other bilingual lexical resources for different language pairs. For example, for many Uralic languages there are German dictionaries, as that has traditionally had a strong role as a scientific language of the field. Also the dictionaries in local majority languages such as Finnish, Estonian, Latvian and Russian are very common. Although the fact that a great many of them exist only as printed copies limits their use in the digital era.

Nevertheless, dictionaries play an important role in language documentation and revitalization efforts. For endangered Uralic languages, Akusanat online dictionary (Hämäläinen and Rueter, 2019) collects multilingual dictionary resources in multiple endangered languages such as the ones in focus of our paper: Komi-Zyrian, Livonian and Erzya. Making it possible for native speakers and language learners to access such a resource has a very big societal impact within the language communities.

Furthermore, online resources such as Wiktionary have gathered very large amounts of lexical data for majority languages. This data does not necessarily represent a fully curated and finalized product in which all entries would be of an equal quality. Only more recently has there been interest in building such resources in the languages that are nowadays more widely used, such as English. As creating these resources is an enormous undertaking, we investigate in this study the possibility of predicting translations from endangered languages to resource-rich languages automatically from existing translations in these high-resource language Wiktionaries.

We would like to point out that the languages we are working with in this paper are endangered, not just low-resourced (see Hämäläinen 2021). According to UNESCO Atlas of World languages (Moseley, 2010), Komi-Zyrian (kpv) has 217,316 native speakers and Erzya (myv) 400,000 native speakers. Livonian (liv), however, does not have any surviving native speakers[1], but has a small community of second language speakers.

Apart from Livonian, these languages have received some digital language documentation interest. Erzya (Rueter and Tyers, 2018) and Komi-Zyrian (Partanen et al., 2018) have small Universal Dependencies tree banks and morphological transducers (Rueter et al., 2020).

[1] https://www.thetimes.co.uk/article/death-of-a-language-last-ever-speaker-of-livonian-passes-away-aged-103-8k0rlplv8xj

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 139 - 148
May 26-27, 2022 ©2022 Association for Computational Linguistics

In the method we investigate in this study, the translations for a word in different languages are represented as graphs. This allows for an effective use of a large number of lexical resources that are not complete, but support one another.

Our main contributions in this work are:

1. We describe a method for inferring translations by combining different graph-based link prediction methods in endangered language data.

2. We evaluate their performance and applicability by conducting a manual evaluation, followed by detailed analyses and discussions.

3. We implement an artificial neural network model to determine the quality of predictions by the algorithmic methods automatically.

4. The prediction results of our method are published in an online dictionary after being verified by lexicographers to have a direct impact on the endangered language communities in question.

Our approach makes it possible for lexicographers to bootstrap new languages into existing multilingual dictionaries. This saves time as instead of building a lexicon from the ground up, their task becomes more of a post-editing, where new translations need only to be verified rather than written from scratch. In the context of larger languages, post-editing has become mainstream in lexicographic work (see Jakubicek et al. 2018), however in the context of endangered languages post-editing has thus far received less lexicographic interest.

2 Related Work

There is a plethora of NLP work out there relating to endangered languages ranging from rule-based approaches (Tyers, 2010; Zueva et al., 2020; Rueter and Hämäläinen, 2020) to latest neural models (Ens et al., 2019; Alnajjar, 2021; Wiechetek et al., 2021). In this section, however, we focus more on work on extending dictionaries.

There has been several attempts in the past in predicting new translations in bilingual and multilingual dictionaries. In this section, we describe the most relevant ones to our work. There has been related approaches to extending semantic knowledge bases (Raganato et al., 2016; Pasini and Navigli, 2017; Gesese et al., 2020), but we leave their detailed description out of this section as the problem the approaches try to solve is fundamentally different in terms of the availability and magnitude of the data.

Lam and Kalita (2013) have proposed a method for reversing bidirectional dictionaries (e.g., reversing Hindi-English to English-Hindi). Their approach requires WordNet[2] (Fellbaum, 1998) for at least one of the languages, and uses the similarities between the words and their synonyms, hyponyms and hypernyms in WordNet to estimate the quality of the reverse translations. They have tested the method by reversing resource-poor and endangered language dictionaries (e.g. Karbi, Hindi and Assamese) to have English as the source language instead of the destination language. It is worth noting that this approach is not capable of producing dictionaries or translations in new languages.

Lam et al. (2015) proposed a method for creating new dictionaries for resource-poor languages. In their work, a dictionary of a low-resource language to a resource-rich language with a high-quality WordNet is needed. To translate a word from the source language to a new language (e.g. Arabic), their method uses links between the English WordNet and existing multiple intermediate WordNets of other languages such as Finnish and Japanese to highlight the relevant words in the WordNets. Thereafter, each of these words are translated to the desired destination language using existing machine translation systems such as Google Translate. The higher the agreement between multiple machine translation systems, the higher the score given to the translation.

A constraint-based approach for inducting new bilingual dictionaries for low-resource languages that are share the language family has been proposed by Wushouer et al. (2015). In their approach, a graph is constructed from two bilingual dictionaries (i.e. A-B and B-C, where B is the intermediate language), and new potential translation links are examined by treating the problem as conjunctive normal form (CNF) and using WPMaxSAT solver to identify the new translations. This work has been extended further in (Nasution et al., 2016) to generalize the method to work for a larger group of languages and identify the best constraint set according to the language pairs.

A graph-based method for combining multiple

[2] https://wordnet.princeton.edu/

Wiktionaries and inferring new translations using graph-based probabilistic inference measured by random walks was proposed by Soderland et al. (2009). The goal of their work is to construct a huge dictionary covering the well-resourced languages (e.g., English, French, Spanish, . . . etc) and suggest new dictionary translations; nonetheless, their work does not address endangered or resource-poor languages. Another graph-based method was embraced by Alnajjar et al. (2021).

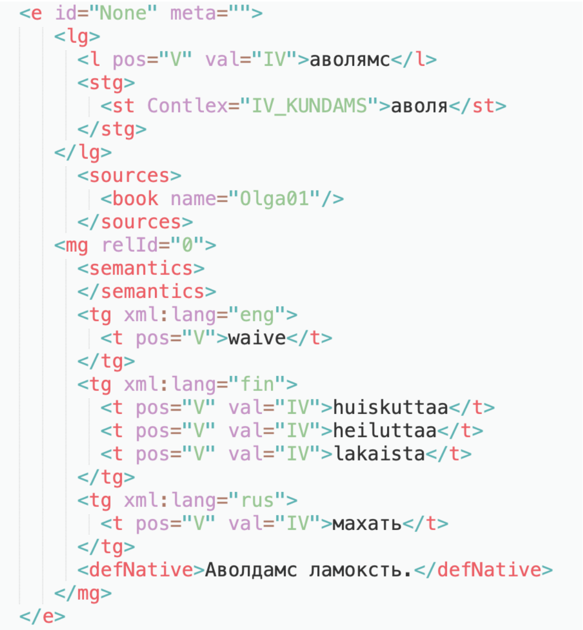
Donandt et al. (2017) have trained a Support Vector Machine (SVM) model to predict whether a new translation is valid or not. Given multiple bilingual dictionaries, a directed graph is constructed where nodes are unique words with their language and part-of-speech tag. Depth-first search is applied to find cycles in the graph. Translations found in cycles with a translation in the dictionary from the target word back to the source are considered to be positive examples, whereas translations found in paths but not cycles are treated as negative instances. Additional features are passed to the model as well, such as the frequency of source word in a dictionary, number of available paths between the source and target words, and, in the case of sharing the language family, the average Levenshtein distance between all the words in the path. This method was not investigated nor evaluated for endangered languages.

3 Data

Two types of resources are used in our approach, 1) XML dictionaries of endangered languages (such as Komi-Zyrian, Livonian and Erzya, with kpv, liv and myv as ISO 639-3 codes respectively) and 2) Wiktionaries[3] of resource-rich languages (such as English and French). While we could utilize the Finnish WordNet (Lindén and Carlson, 2010) as an additional resource in this task as done in some of the previous work, however, in practice it would introduce more noise due to the relatively low quality of the Finnish WordNet[4].

3.1 XML Dictionaries

The XML dictionaries have been created in connection with the development work at morphological analysers, and they contain both materials from already published dictionaries and also individually added entries. In this work, we use dictionaries of three endangered languages Komi-Zyrian, Livonian and Erzya. The Komi and Erzya dictionaries are built as part of the Giella Project (Moshagen et al., 2014)[5] and they are available through UralicNLP (Hämäläinen, 2019), while the Livonian dictionary has been outlined in Rueter (2014).



```
<e id="None" meta="">
    <lg>
      <l pos="V" val="IV">аволямс</l>
      <stg>
        <st Contlex="IV_KUNDAMS">аволя</st>
      </stg>
    </lg>
      <sources>
        <book name="Olga01"/>
      </sources>
    <mg relId="0">
      <semantics>
      </semantics>
      <tg xml:lang="eng">
        <t pos="V">waive</t>
      </tg>
      <tg xml:lang="fin">
        <t pos="V" val="IV">huiskuttaa</t>
        <t pos="V" val="IV">heiluttaa</t>
        <t pos="V" val="IV">lakaista</t>
      </tg>
      <tg xml:lang="rus">
        <t pos="V" val="IV">махать</t>
      </tg>
      <defNative>Аволдамс ламоксть.</defNative>
    </mg>
</e>
```

Figure 1: An example of the XML structure in the Erzya dictionary.

As seen in Figure 1, an XML dictionary contains lexemes, their parts-of-speech, and translations grouped by the meaning group. Out of the three, the Livonian dictionary is the most consistent dictionary with multi-translations to Finnish (19,210), Latvian (18,064) and Estonian (18,684). Komi-Zyrian mostly has Russian (32,744) and Finnish (11,745) translations, and some English (6,702). Erzya has Finnish (12,631), Russian (7,572) and English (3,739).

While in theory these multilingual dictionaries have their translations divided into meaning groups that group semantically similar translations together, in practice these meaning groups are of a poor quality (see Hämäläinen et al. 2018) and thus omitted in our approach. The problem can already be seen in Figure 1 with the Erzya word аволямс where Finnish words *huiskuttaa* (to wave) and *heiluttaa* (to wave) are in the same meaning group as *lakaista* (to sweep).

[3] https://www.wiktionary.org/

[4] For instance, the word for a *dog* (koira) is linked as a synonym for a *pig* (sika), and unacceptably the word for a *woman* (nainen) is linked as a synonym for *whore* (huora) among others.

[5] https://giellalt.uit.no

3.2 Wiktionary

Wiktionaries are rich multilingual online dictionaries consisting of an enormous number of words, translations, examples. There are Wiktionaries for many resource-rich languages and they are publicly available.

We have crawled and parsed the Finnish (fin), Estonian (est), French (fra), Latvian (lav) and Russian (rus) Wiktionaries to extract all words and translations provided in them. Despite the humongous linguistic data supplied, the data in each Wiktionary is structured differently and is not well aligned with other dictionaries (e.g. a given translation does not necessary exist in the reverse direction). These dictionaries do not have many translations in our endangered languages of interest, but they serve as an important resource for our link prediction approach.

4 Inferring New Translation Candidates

Representing translations in a graph, where words are represented as nodes and translations between words as edges, is intuitive and has been successfully used for the task in the past, as described in the related work. In fact, some of the modern approaches to lexicography have also rejected the traditional tree structure of a dictionary in favor of a graph representation (Mechura, 2016). Similarly, we represent both types of dictionaries, XMLs and Wiktionaries, in a graph-based network using NetworkX library (Hagberg et al., 2008). Unlike some of the previous work such as (Donandt et al., 2017), the graph is not directional, given that nearly all lexical translations work bidirectionally.

Let $G = (V, E)$ denote the graph, where V is all the vertices/nodes in the graph and E is all undirected edges/links between two nodes. We initialize the graph with all translations from the five Wiktionaries in such a way that their entries become interconnected based on words and their translations.

To predict new translations from the source language S to the target language T, we load the XML dictionary of the desired endangered language to the graph while omitting any existing translations to the target language. This is done to ensure that all translations to the target language are projected by the method.

Once the graph is constructed, we iterate over all nodes from the source language $V_S = \{s|s \in V \cap S\}$ and their neighbouring nodes $N(s) = \{n|ns \in E\}$. For all the neighbouring nodes linked to the source language n, we examine whether they belong to the target language, i.e. $n \in T$. When such a constraint is satisfied, a new translation between the source lexeme s and n is considered as a candidate translation and assessed using link predictions methods. All candidates scoring zero on any of the link predictions methods described below are pruned out.

We employ four link prediction methods to discover new translations; these are 1) Jaccard coefficient (Jaccard, 1912), 2) Adamic-Adar index (Adamic and Adar, 2003), 3) resource allocation index (Zhou et al., 2009), and preferential attachment score (Liben-Nowell and Kleinberg, 2007). In short, Jaccard coefficient computes a score based on the common neighbours between the source and target nodes with respect to the total number of their neighbours. The Adamic-Adar index is defined as:

$$\sum_{w \in N(s) \cap N(n)} \frac{1}{log|N(w)|}$$

The resource allocation index is defined similarly but without taking the log of the denominator. Lastly, the preferential attachment score measures the magnitude of the neighbours of each node, which is defined as $|N(s)||N(n)|$.

An example of a sub-graph containing the Livonian lexeme (Japān) along with links to existing translations in the XML dictionary (which are Japani in Finnish, Jaapan in Estonian and Japāna in Latvian) is shown in Figure 2. All the remaining nodes in the graph and their black connections to the other nodes are from Wiktionaries. By running the link prediction methods described above to infer translations from Livonian to English, two new links are suggested and they point to the lexemes Japan and Nippon, shown in red dashed lines. The methods were able to recommend the link to the Japan with high confidence as there is a strong support based on their neighbouring nodes (i.e. liv_Japāna, fin_Japani and est_Jaapan), whereas the link to Nippon had a low confidence as only one node supports it (i.e. est_Jaapan).

5 Manual Evaluation

In our evaluation, we run the link prediction method for the following four language pairs, 1) Erzya and English 2) Livonian and English, 3) Komi-Zyrian and English and 4) Komi-Zyrian and

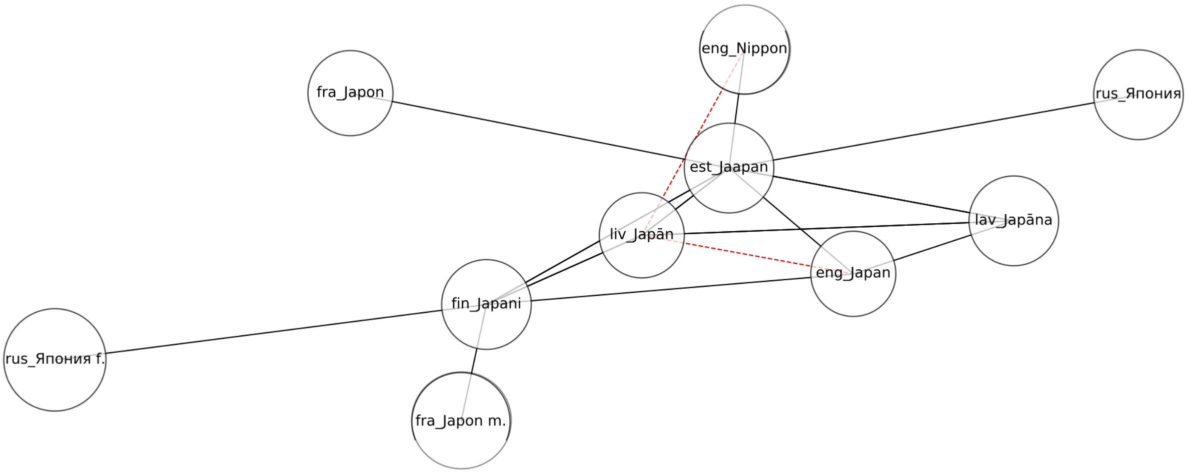


Figure 2: A sub-graph illustrating an example for inferring new translations from Livonian (Japān) to English (Japan and Nippon) by the methods (highlighted in red).

French. Which resulted in 17,042, 22,911, 9,611 and 7,765 translation suggestions for the four language pairs, respectively.

To evaluate the method, we have reached to fluent speakers in the source and target languages and requested them to manually annotate 200 randomly selected predictions. None of these predictions existed before in the XML dictionaries between each language pair. For each translation, they were instructed to indicate whether it is 1) good, 2) acceptable, 3) incomplete or 4) bad. Good translations are dictionary-ready entries and can be automatically populated as they are. Acceptable instances are correct predictions but may contain ambiguity due to, for example, synonymy or polysemy. Incomplete translations are close to the desired translation but require manual modifications, while bad translations are completely off predictions and should be removed.

In total, we obtained 800 annotated predictions. Table 1 shows the summary of annotations per language pairs. The annotations point out that the majority (44.62%) of inferred translations are good and can be used as they are. 16.62% and 15.5% of the predictions were seen as acceptable and incomplete, in the given order. Overall, this demonstrates the effectiveness of the method in predicting translations for endangered languages, with 76.75% good or potential translations, and only 23.25% bad translations.

We can see some examples of the predictions and human annotations in Table 2. In the table, we can see examples of all four annotation categories for Komi-Zyrian to English translations. The annotator also wrote notes for non-good translations.

Pair	Good	Acceptable	Incomplete	Bad	Total
myv-eng	76	34	36	54	200
liv-eng	88	23	39	50	200
kpv-eng	102	35	29	34	200
kpv-fra	91	41	20	48	200
Total	357	133	124	186	800

Table 1: A summary of the manual annotation of predicted translations from endangered languages to resource-rich languages.

Next, we calculate the Pearson correlation coefficient to determine if there is a linear correlation between each of the four link prediction methods and the manual annotations. We assigned the annotation a value of from 3 (for good) to 0 (for bad). Our results indicate that there is a positive weak correlation between the annotation values and the predicted scores for three methods Jaccard coefficient, Adamic-Adar index, and resource allocation index. For preferential attachment, no correlation existed. All of the four correlations are with very strong statistical significance, i.e. p-value < 0.001. These correlation scores indicate the importance of considering the total and common neighbouring translations of the source and target words, something that is not taken into consideration in the preferential attachment method.

5.1 An automated evaluation attempt

Komi-Zyrian and Erzya dictionaries contain some English translations. As these translations were ignored during the automatic prediction phase, we

Komi-Zyrian	English	Annotation	Note
норматив	norm	good	
во пом	year	incomplete	end of the year
чуксасьны	crow	acceptable	verb
сöгластöм	indeclinable	bad	uncompromising

Table 2: Examples of Komi-Zyrian to English predictions and annotations.

can use them as a simplistic automatic evaluation metric to test if the method infers them correctly. To do so, we only consider English translations which exist in the initial graph (i.e., constructed from Wiktionaries) because some of these translations are placeholders (i.e., 'YY') or contain additional meta-data (e.g., the context or specification), not to mention that Wiktionaries are not complete resources and some words will be missing. This filtering resulted in 4,096 and 3,386 Komi-English and Erzya-English translation pairs to be assessed by the link prediction methods. For Komi-Zyrian to English, 2,419 (59%) of translations were predicted correctly; however, we were able to verify only 423 (13% of) Erzya to English translations by the existing XML dictionary.

These numbers indicate that at least this many translations were correct based on this automated evaluation method, however, this method cannot assess how many of the predicted translations that were not in the dictionaries, were correct as well. In our experience, dictionaries (even larger Wiktionaries) have an inconsistent coverage of synonyms in the translations. Which means that if our method predicts a synonym of an existing translation that is not in the dictionary, this simplistic automated evaluation cannot capture that. With a quick look into the data, we were able to see several of these cases.

Because no dictionary is perfect, and even less so in the context of endangered languages, it is difficult to conduct the kind of automated evaluation that would be functional in assessing the degree to which our predictions are correct. For this reason, we believe that the manual evaluation by people knowledgeable in the languages in question is the best way of evaluating the performance of the method. This also creates a very useful gold standard dataset that can be used in further evaluation of different approaches.

Layer (type)	Output Shape	Param #
Linear-1	[-1, 64, 64]	320
ReLU-2	[-1, 64, 64]	0
BatchNorm1d-3	[-1, 64, 64]	128
Linear-4	[-1, 64, 64]	4,160
ReLU-5	[-1, 64, 64]	0
BatchNorm1d-6	[-1, 64, 64]	128
Linear-7	[-1, 64, 64]	4,160
ReLU-8	[-1, 64, 64]	0
BatchNorm1d-9	[-1, 64, 64]	128
Dropout-10	[-1, 64, 64]	0
Linear-11	[-1, 64, 1]	65

Table 3: A summary of the architecture of the neural network.

6 Automatic Detection of Good Predictions

To further aid lexicographers in creating dictionaries, especially for endangered languages, we build an artificial neural network model for detecting whether a predicted translation by the methods is a good one. An automated way of filtering out the bad translations cuts the time needed for going through the predictions manually.

We have experimented with different neural architectures and techniques. For the scope of this work, we describe the outperforming model which is a multilayer feedforward neural network (for a summary of the architecture, see Table 3). The input to the network is the prediction scores computed by the link prediction methods and the output is a binary score, 1 denoting a good prediction and 0 a bad one. We follow the rule-of-thumb of introducing hidden layers based on 70-90% of the size of the input (Boger and Guterman, 1997), which yields three hidden layers and each layer consists of 64 neurons. Rectified linear unit (ReLU) is used as an activation function after each layer. Subsequently, batch normalization (Ioffe and Szegedy, 2015) and dropout (Srivastava et al., 2014) (with a probability of 10%) are applied to accelerate train-

	Precision	Recall	F1-score	N
Baseline				
Good	77%	51%	61%	124
Bad	22%	47%	30%	36
Accuracy	50%			160
Neural Model				
Good	81%	98%	89%	124
Bad	73%	22%	34%	36
Accuracy	81%			160

Table 4: The accuracy, precision, recall and F1-score of a random baseline and our neural model for detecting good translation candidates.

ing, and reduce internal covariate shift and overfitting. In total, the network had 9,089 trainable parameters.

In our model, we utilize Adam optimizer (Kingma and Ba, 2014) and a sigmoid layer combined with binary cross entropy as the loss function due to its suitability for the binary classification task. To obtain the classification from the model, a sigmoid function followed by rounding the result is applied post inference.

For the problem we are tackling, there are no available training datasets, neither for endangered languages nor resource-rich languages. To overcome this, we exploit our manual annotations and split them into 80-20 splits for training and testing. To convert the annotations into binary classes, we treat all good, acceptable and incomplete translations as positive instances and bad ones as negative.

After 1,000 epochs of training with a learning rate of 0.001, the model reached an accuracy of 81%. Table 4 reports a summary of the performance metric of the model in comparison to a random classifier as a baseline.

7 Discussion

When looking at the bad candidate translations, the reasons why they were predicted by our method can be divided roughly into two categories: polysemy and wrong translations in the original XML dictionaries. A polysemy of a word in one language can cause a wrong translation to appear in another language that does not exhibit the same polysemy. For example the Komi-Zyrian word гол had been translated into *paint* instead of the correct translation *goal*. This is due to polysemy in Finnish as the Finnish word *maali* means both *goal* and *paint*. Had there been more translations in between languages for these words that do not have the Finnish polysemy, the graph based model would have been less likely to predict this translation.

We have attempted to test the method by focusing solely on Wiktionary data, where we would omit all existing translations from a particular source language to another (e.g., Finnish to French or English). Nonetheless, many of the predicted translations were good but were missing from the Wiktionary of the source language, making it infeasible to assess the effectiveness of the method. Despite that, this is a strong indication that the proposed method with our model could be employed to enrich existing Wiktionaries further.

An idea we had for training a neural model for predicting whether the new predictions are good or bad was to generate synthetic training data automatically. In practice, collecting examples of good translations from Wiktionaries is easy, but producing automatically examples of bad translations is more difficult. Predicting random links between words would result in all of the link prediction models outputting such a low score that it would hardly be representative of the real case of bad translations that are mainly due to polysemy or wrong initial translations.

We tried out producing a dataset of bad translations with the idea that if an English word, for instance *can* is translated into *voida* (be able to) and *purkki* (can as a container) in Finnish and *võima* (be able to) and *purk* (can as a container) in Estonian, then predicting *voida* as a translation of *purk* and *purkki* as a translation of *võima* would make our synthetic data have very representative examples of bad translations. However, in practice, we ran into a coverage issue in Wiktionaries. For example, the English Wiktionary did not have any entry that would have had at least two translations into Finnish and Estonian. This made our good idea in theory impossible in practice.

While quality and coverage of the existing data pose challenges, our work has provided some insight for the lexicographers working with these resources about the limitations of the current state of the lexical resources. This has been well received as a form of a sanity check among the lexicographers in question given that the lexicographic resources have been built by different people depending on their funding situation. This means that a lot of the work done in the dictionaries has been there before the current people working with the resources have started extending them.

In our graphs, we have omitted the part-of-speech because it is not present for all lexemes, whether in the XML dictionaries or Wiktionaries. Taking them into consideration would have resulted in inferring low-quality translations in smaller magnitudes. Therefore, we believe that incorporating part-of-speech tags is a crucial step, once new translations are inferred. As this would assist in detecting some ambiguous cases where a miss-match between the parts-of-speech is sufficient to prune them out. The part-of-speech tags could be automatically predicted by taking advantage of neural- and graph-based methods (Angle et al., 2018; Das and Petrov, 2011; Thayaparan et al., 2018). However, in some cases, ensuring the same part-of-speech tag, might lead to correct translations being filtered out. For instance, the Finnish word *alla* may be an adverb or a postposition, whereas its English translation *under* is a preposition.

In terms of the features used in our neural model, we use the prediction scores returned by the link prediction methods. This causes the neural model to act as an expert voter observing the various scores and to make the executive decision of whether the prediction is valid or not. Additional features could be passed to the model, such as the strings of both source and target words, and meta-information about their nodes (e.g., the number of their distinct and common neighbours). Based on Donandt et al. (2017) work, using the Levenshtein distance between source and target words resulted in poor classifications. Such features contribute differently to the performance of the model depending on the languages and would limit the model to closely related languages with a high number of cognates. This motivated our choice of judging the quality of predictions based on the link prediction scores, which causes our model to be generic and appropriate for many different language pairs as we assume no phylogenetic relation between the languages in question. This also makes it possible for our approach to work across writing systems as we are dealing with languages written in Latin and Cyrillic alphabets.

8 Conclusions and Future Work

We have released the source code of our method and its predictions on Github[6]. Our method could, in the future, be integrated with the existing dictionary editing infrastructures for Uralic languages such as Giella (Moshagen et al., 2014) and Ve'rdd (Alnajjar et al., 2020). This would make link prediction an active part of the process of building lexical resources, making it a more dynamic human-in-the-loop task.

We have presented our work on extending the existing lexical resources for several endangered languages. For the time being, human annotators are needed to go through the predicted translations, although we have perceived promising results with our neural approach.

Regardless of the accuracy of the current method for identifying good predictions or what any future method might reach, we believe that a lexicographer needs to go through the predictions at any rate. Compiling dictionaries for an endangered language is an important step in the language documentation and, if done right, can greatly benefit the native speakers of the language in learning foreign languages, and also anyone interested in learning the endangered language in question. This being said, any fully automatically produced lexicon will have errors that ultimately lead to misunderstandings and can be harmful for the language community.

We envision that our work opens the door for constructing aligned multilingual word-embeddings between endangered languages and high-resource languages. This would narrow the gap between severely scarce-resource languages and the latest neural machine translation techniques, making it possible to build a functional neural translation system from languages at the risk of dying to a vast number of big languages which in return would greatly benefit the communities of endangered language.

The results produced by our method will be manually filtered by lexicographers and included in the Akusanat online dictionary[7]. The goal of our paper has been that of extending existing lexicographic resources so that the language communities can directly benefit from our research. Without releasing our results and having them manually verified, we would be embracing an unethical research tradition that relies on cultural and linguistic appropriation for a purely academic benefit.

Acknowledgments

Many thanks to our Livonian evaluators Uldis Balodis and Valts Ernštreits. This work was funded in part by the French government under manage-

[6] https://github.com/mokha/translation-link-prediction/

[7] https://www.akusanat.com/

ment of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Khalid Alnajjar. 2021. When word embeddings become endangered. *Multilingual Facilitation*, pages 275–288.

Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. Ve'rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 1–6.

Khalid Alnajjar, Jack Rueter, Niko Partanen, and Mika Hämäläinen. 2021. Enhancing the erzya-moksha dictionary automatically with link prediction. *Folia Uralica Debreceniensia*, 28:7–18.

Sachi Angle, Pruthwik Mishra, and Dipti Mishra Sharma. 2018. Automated error correction and validation for POS tagging of Hindi. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Z. Boger and H. Guterman. 1997. Knowledge extraction from artificial neural network models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 4, pages 3030–3035 vol.4.

Dipanjan Das and Slav Petrov. 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*, pages 600–609, Portland, Oregon, USA. Association for Computational Linguistics.

Kathrin Donandt, Christian Chiarcos, and Maxim Ionov. 2017. Using machine learning for translation inference across dictionaries. In *LDK Workshops*, pages 103–112.

Jeff Ens, Mika Hämäläinen, Jack Rueter, and Philippe Pasquier. 2019. Morphosyntactic disambiguation in an endangered language setting. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 345–349, Turku, Finland. Linköping University Electronic Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Genet Asefa Gesese, Mehwish Alam, and Harald Sack. 2020. Semantic entity enrichment by leveraging multilingual descriptions for link prediction. In *arXiv preprint arXiv:2004.10640*.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

Mika Hämäläinen. 2019. Uralicnlp: An nlp library for uralic languages. *Journal of open source software*, 4(37).

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! *Multilingual Facilitation*.

Mika Hämäläinen and Jack Rueter. 2019. An open online dictionary for endangered uralic languages. In *Electronic lexicography in the 21st century Proceedings of the eLex 2019 conference*. Lexical Computing CZ sro.

Mika Hämäläinen, Liisa Lotta Tarvainen, and Jack Rueter. 2018. Combining concepts and their translations from structured dictionaries of uralic minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 862–867. European Language Resources Association (ELRA).

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Milos Jakubicek, Michal Měchura, Vojtech Kovar, and Pavel Rychly. 2018. Practical post-editing lexicography with lexonomy and sketch engine. In *The XVIII EURALEX International Congress*, page 65.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *arXiv preprint arXiv:1412.6980*.

Khang Lam, Feras Al Tarouti, and Jugal Kalita. 2015. Automatically creating a large number of new bilingual dictionaries. In *AAAI Conference on Artificial Intelligence*.

Khang Nhut Lam and Jugal Kalita. 2013. Creating reverse bilingual dictionaries. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 524–528, Atlanta, Georgia. Association for Computational Linguistics.

David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

Krister Lindén and Lauri Carlson. 2010. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17:119–140.

Michal Mechura. 2016. Data structures in lexicography: from trees to graphs. In *RASLAN 2016 Recent Advances in Slavonic Natural Language Processing*, page 97.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. In *The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era"*, pages 71–77.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2016. Constraint-based bilingual lexicon induction for closely related languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3291–3298, Portorož, Slovenia. European Language Resources Association (ELRA).

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian universal dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Tommaso Pasini and Roberto Navigli. 2017. Train-O-Matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88, Copenhagen, Denmark. Association for Computational Linguistics.

Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2016. Automatic construction and evaluation of a large semantically enriched wikipedia. In *IJCAI*, pages 2894–2900.

Jack Rueter. 2014. The livonian-estonian-latvian dictionary as a threshold to the era of language technological applications. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 5(1):251–259.

Jack Rueter and Mika Hämäläinen. 2020. Fst morphology for the endangered skolt sami language. In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, pages 250–257, France. European Language Resources Association (ELRA).

Jack Rueter, Mika Hämäläinen, and Niko Partanen. 2020. Open-source morphology for endangered mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. The Association for Computational Linguistics.

Jack Michael Rueter and Francis M Tyers. 2018. Towards an open-source universal-dependency treebank for erzya. In *International Workshop for Computational Linguistics of Uralic Languages*.

Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 262–270, Suntec, Singapore. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Mokanarangan Thayaparan, Surangika Ranathunga, and Uthayasanker Thayasivam. 2018. Graph based semi-supervised learning approach for Tamil POS tagging. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Francis Tyers. 2010. Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.

Linda Wiechetek, Flammie Pirinen, Mika Hämäläinen, and Chiara Argese. 2021. Rules ruling neural networks - neural vs. rule-based grammar checking for a low resource language. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1526–1535, Held Online. INCOMA Ltd.

Mairidan Wushouer, Donghui Lin, Toru Ishida, and Katsutoshi Hirayama. 2015. A constraint approach to pivot-based bilingual dictionary induction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 15(1).

Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. 2009. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630.

Anna Zueva, Anastasia Kuznetsova, and Francis Tyers. 2020. A finite-state morphological analyser for Evenki. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2581–2589, Marseille, France. European Language Resources Association.