CLUSTERING OF ANALOGIES FOR INTER-LANGUAGE SIMILARITIES

Cindy Pereira
IDMC
Université de Lorraine
Nancy
cindy.pereira1@etu.univ-lorraine.fr

Justine Diliberto
IDMC
Université de Lorraine
Nancy
justine.diliberto4@etu.univ-lorraine.fr

Anna Nikiforovskaja
IDMC
Université de Lorraine
Nancy
anna.nikiforovskaja1@etu.univ-lorraine.fr

ABSTRACT

This report presents the results of a project about finding similarities between languages. For this, analogies between lemmas and inflected forms are generated, and clustering is computed on these results. In addition, morphological rules are manually studied, to predict or explain the results of the computations. A website is also made available to display the rules and percentage of similarities by pairs of languages.

***K*eywords** Inflections · Analogies · Languages · Clustering

1 Introduction

Organizing languages into family trees has been the work of archaeologists, linguists and geneticists [1]. While language hierarchies have been built by identifying common ancestors, this work presents the idea of finding similarities in languages by comparing word inflections through analogies. Languages from different families are thus studied, in the hope of finding new links between them. This research consists in computing language resemblance through clustering of analogies, but also in extracting morphological rules manually, as these could predict or explain similarities found between languages. The results can be accessed on a website to compare pairs of languages [1].

This research is based on the work of Alsaidi et al. [2], who introduced a deep learning approach to detect morphological analogies. Their approach is to study how a CNN model which was trained to predict analogies for one language would transfer to the other. In this paper we continue studying this type of transfer in depth, and provide a discussion on why the model transfers well or not well from one language to another.

This paper introduces next the experimental setup, that is, the baseline used and the corpus. Then, the early stages of the project are described, before focusing on the results. Finally, elements of discussion are given.

2 Experimental setup

2.1 Baseline

Alsaidi et al. [2] introduced a CNN-based model which is used as a baseline for the experiments presented in this paper. This model is a character-based model, which is trained to classify if four input words are a morphological analogy or

[1] https://inter-language-analogies.herokuapp.com/

not. The quadruple $a : b :: c : d$ is considered to be a morphological analogy in case a refers to b the same way as c refers to d. This generalisation describes a set of morphological inflections from a to b, and from c to d.

2.1.1 Analogies

For example, $stay : stayed :: play : played$ is a morphological analogy, as both parts describe an evolution from the infinitive form of the verb to the past participle form.

Analogies have also several properties which are utilised by Alsaidi et al. [2] to generate both positive and negative analogies to train the CNN models. For example, if $(a : b :: c : d)$ is a morphological analogy, then $(a : c :: b : d)$ is also a morphological analogy. Actions like this help to increase the number of positive analogies to train on.

2.1.2 Model training

Originally the training of the CNN model by Alsaidi et al. [2] was not balanced well. This aspect of the model was improved by generating the same amount of different positive and negative analogies for each quadruple. The reasons behind the model's behavior are explained in this paper. To do so, a way to include the accuracy of classifying both positive and negative analogies into a single score is computed. These accuracy scores are combined by taking their harmonic mean.

Precisely, $F = \frac{2 \cdot a_{\text{pos}} \cdot a_{\text{neg}}}{a_{\text{pos}} + a_{\text{neg}}}$, where:

F – the final score of the model on the language,

a_{pos} – accuracy of predicting positive analogies (a ratio of correct answers on positive analogies),

a_{neg} – accuracy on negative analogies.

Alsaidi et al. [2] also study both partial and full transfers of the model. The difference between them is that with the full transfer from on language to another, the whole model for the first language is used. On the other hand, the second language still has its embedding model and then only the body of the model is taken from the first language, with the partial transfer. This paper will mainly study the partial transfer, as the full one is alphabet-dependent, as it was shown in the above mentioned paper.

2.2 Corpus

The data used is taken from the SIGMORPHON 2020 corpus, presented in [3], for the shared Task 0 entitled "Typologically Diverse Morphological Inflection" [2].

SIGMORPHON, or Special Interest Group on Computational Morphology and Phonology, is a subgroup of the Association for Computational Linguistics (ACL). Workshops and shared tasks are organized yearly by SIGMORPHON about computational morphology and phonology.

The corpus provides morphological inflections for 90 languages (45 development and 45 surprise languages). Each language data is divided into training, development and test sets. More precisely, the corpus consists in text files containing lemmas, their inflection forms, and the corresponding inflection rule types (in the form of Universal Morphological Feature Schema[3]) separated by tabs. Below is an example of a few data entries from the English development part:

```
fabulize fabulized V.PTCP;PST

digiscope digiscopes V;SG;3;PRS

tempre tempre V;NFIN

keypunch keypunching V.PTCP;PRS
```

[2] `https://github.com/sigmorphon2020/task0-data`

[3] `https://raw.githubusercontent.com/unimorph/unimorph.github.io/master/doc/unimorph-schema.pdf`

3 Early stages

3.1 Preliminary research and experiments

3.1.1 Language families

The very first steps of this research were to get information on how languages are classified into families, and which families exist. An approach for gathering languages into families is to study how some words are related, from a purely orthographic point of view [4]. Another article focuses on a quantitative method to classify languages, that is, to use a mathematics-like methodology and to get inspiration from genetics to build trees [5]. Some other methodologies use actual genetics to measure relations between languages, as language contact is often associated with immigration and people from a common geographic area will share some genes [6]. Schemes of language hierarchies were also examined, to visualize the proximity of the languages in the corpus, and their place in language families from a broader scale [7].

Given the amount of languages in the corpus, the task of manually studying all possible inflections of a language was deemed too tedious. Also, a huge number of languages would not make it possible to understand them thoroughly and compare them effectively. It was thus decided to keep a small number of languages, that will be presented in the next part.

3.1.2 First transfers

The preliminary experiments with transferring the model from one language to another showed the importance of the data size for each language. To illustrate that, the results of one of the preliminary experiments are presented below. This experiment was to compare languages from three different groups of languages, which are Germanic group (represented by English and Old English), Amuzgoan group (represented by Amuzgo, from Central America), and Greater Central Philippine group of languages (represented by Tagalog, Hiligaynon and Cebuano). The final score of partial transfers between these languages is shown in fig. 1.

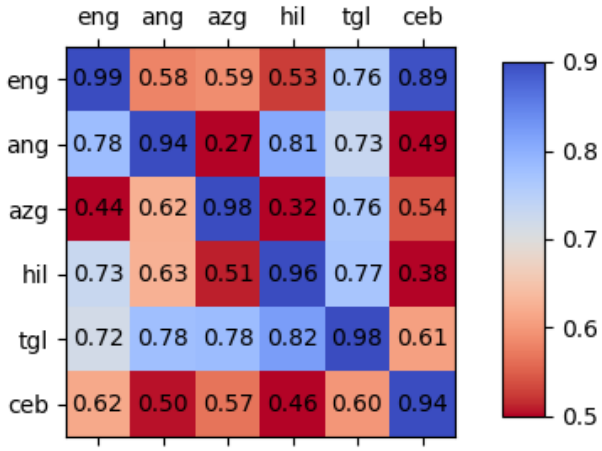

Figure 1: Partial transfer performance from one language to another in a group of English (eng), Old English (ang), Amuzgo (azg), Hiligaynon (hil), Tagalog (tag), Cebuano (ceb)

One may notice several unusual things about the results. One of these things is that the transfer works well from English to Cebuano, which are extremely distant languages. Old English has a relatively high correspondence with Tagalog and Hiligaynon, which is also surprising. English does not transfer well with Old English, however the opposite transfer score is quite high, which can be explained by the fact that English comes from Old English, not the opposite. When considering the Philippine language group, transfers from Cebuano to the two others languages were very low, whereas the results were satisfying between Hiligaynon and Tagalog. Interestingly, Amuzgo obtained a quite good transfer result with Tagalog. The data sizes of these languages are sparse, which can provide some explanation on these observations.

Table 1: Number of inflections per language

Language	#Inflections
Mezquital Otomi	22,962
Swedish	54,888
Karelian	80,216
English	80,865
Finish	99,403
German	99,405

3.2 Language analyses

3.2.1 Chosen languages

For this study, six languages were analysed: English, German, Swedish, Finnish, Karelian, and Mezquital Otomi.

They were chosen according to several criteria. First, a reasonable amount of inflections is available in the corpus (having too scarce data would not favor the model). The number of inflections for each language can be found in the table 1. Second, they belong to one of the three language families that were kept (Germanic, Uralic and Oto-manguean). By doing so, the distance between the languages can be predicted, as the two first language families are European, and the last is South American. Third, it is possible for us to find at least a person who can speak each language (except for Mezquital Otomi).

3.2.2 Rules extraction

To analyse if inflections between languages are compatible, it was necessary to extract inflection rules from the corpus. No automatic tool was proven to be effective for this specific task, given the format and content of the data. As a result, rules were preprocessed using Python and extracted manually.

The preprocessing step included an automatic extraction of rules using NLG package [8] to get an overview of the way the lemmas are inflected. However, the results were not conclusive and the grids were not used after all. Then, the number of inflections for each language has been computed automatically. The distribution was as follows: 5 different types of inflections in English, 18 in Mezquital Otomi, 34 in Swedish, 35 in German, 91 in Finnish, and 161 in Karelian. Finally, these inflections were split and classified in terms of their morphological tags. This helped to find the similarities and understand the underlying rule.

The manual study consisted in comparing each inflection from the same category to its equivalent in the same language. The inflections are created by appending words with suffixes for a majority of the languages of this study, so they are easily noticeable. In most cases, there was a regularity in the inflected forms of a same rule, however some very irregular variants were also found. As a result, even if the constitution of the rules was as accurate as possible, some nuances might have been missed. Nevertheless, the rules are very thorough and make it possible to compare inflections across languages.

3.3 Hypotheses

3.3.1 Very close language pairs

The first language pair that was found to be really close is Karelian and Finnish. As was expected even before the extraction of rules, these two languages belong to the same family and share many characteristics. Karelian has 161 different inflection types, whereas Finnish has 91. Among these inflections, 51 types are shared by both languages and are very similar. Moreover, 29 out of these 51 inflections are exactly the same. Even the differences are not so far from one another, for example the indicative present in the first person singular is marked by "*n*" in Karelian and "*an*" in Finnish. Another example is the genitive plural, ending in "*in*" for Karelian and "*en*" in Finnish.

3.3.2 Relatively close language pairs

English and Swedish can be considered as a relatively close pair of languages, as they share some inflections. First, the active infinitive form in Swedish is the same as the infinitive in English, that is, the lemma unchanged. The past participle forms in English are ending in "*ed*", "*en*" or are irregular, while they are ending in "*d*" or "*en*" in Swedish. Also, the indicative present of the 3rd person singular in English is marked with "*s*" and could be related to the present

passive indicative form in Swedish with the same ending (as this inflection does not exist in English). Finally, there is a similarity in the past indicatives, with "*ed*" for English and "*de*" in Swedish.

German has some similarities when compared to Finnish and Karelian. First, German inflections that are common to Finnish and Karelian can be noted: the 2nd person singular indicative in German is "*est*" and could be linked to "*t*" in 2nd person singular active indicative, also the nominative and accusative singular are the same in the three languages. There are some cases where Finnish and German are approximately similar, like for the 2nd person plural present indicative in "*t*" in German and "*tte*" in Finnish for the same rule. Swedish has the same inflection as German for the imperative, and the definite singular genitive ends with "*ens*" or "*ets*" whereas it ends with "*s*" in German.

3.3.3 Far language pairs

Given that English has very few rules in the dataset, finding similarities with any other language was very complex. One common inflection could possibly be found with German: the 1st and 3rd person plural past indicatives in German "*ten*", with the irregular past participle in English that may end in "*en*". No common rule was found with Finnish or Karelian.

Swedish had two exact similarities with Finnish or Karelian, which was quite few. Both the infinitive active and the nominative indefinite inflection are similar to their lemma, across these languages.

Mezquital Otomi was the furthest from the other languages, which was not surprising either, given that it belongs to a language family that is greatly distant from the other considered language families. The corpus only contains verb inflections, with rules for irregular verbs, present and past perfect, present and past imperfective, and perfective. The inflections are very different from the ones seen in the other languages, as no similarity was found.

4 Results

4.1 Comparing languages

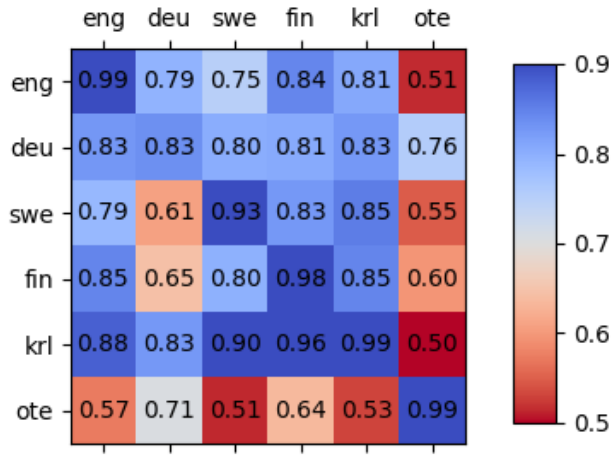

Figure 2: Partial transfer performance from one language to another in a group of English (eng), German (deu), Swedish (swe), Finnish (fin), Karelian (krl), and Mezquital Otomi (ote)

One can see the final partial transfer performance in fig. 2.

The first noticeable fact is that Finnish and Karelian transfer really well (0.85 and 0.96). This fits the hypothesis on the closeness on these languages based on the similarity between the inflection rules.

Karelian is the language with the highest number of different inflection rules. This could explain why a model trained on this language will perform well on all the other languages, except for Mezquital Otomi.

One can also see that Mezquital Otomi transfers poorly with every language, and no language transfers well with it. This was expected as there is no other language from the same family, and no similarity was found in terms of rules.

However, some interesting results were unexpected. For instance, German doesn't transfer really well with itself with a result of 83%, whereas other languages transfer perfectly with themselves. The German-to-German result is actually

very similar to all the transfers from German, which vary from 76% to 83%. For example, the results of the model trained on German and applied on English is of 83%.

Also, a model trained on English will transfer better with Finnish and Karelian than with German or Swedish. This is unexpected because English is in the Germanic group and not in the Uralic group (containing Finnish and Karelian). The same fact happens with models trained on Swedish.

4.2 Final product

4.2.1 Website content

A website[4] was built to display the results of language comparisons in an interactive way. The idea is to be able to select 2 languages in the list of the 6 studied languages, in the form of two scrolling menus. After the selection of languages, the morphological rules (that were extracted earlier) are shown side by side. For more readability, only the categories of rules are displayed first, that are *nouns*, *adjectives* or *verbs*, and then it is possible to click on them to show their content. Similar rules, i.e. the morphological inflections which are represented the same way in two languages, are highlighted to make the analysis easier. The ratio of similar rules is shown. Also, the percentage of the similarity of the chosen language pair, resulting from the analogies generated by the model, is exposed. In addition, some of the analogies are shown, where the model performed well for the chosen language.

4.2.2 Website implementation

The microframework Flask [9] was used to build the website. A simple scheme of the website architecture is shown on fig. 3. When a user fills in information on what language they want to compare, this information is passed via Flask to the backend application. All the information about the languages chosen is retrieved from three types of files: the prepared JSON files with extracted rules, JSON files with example analogies and a CSV file with model transfer results.

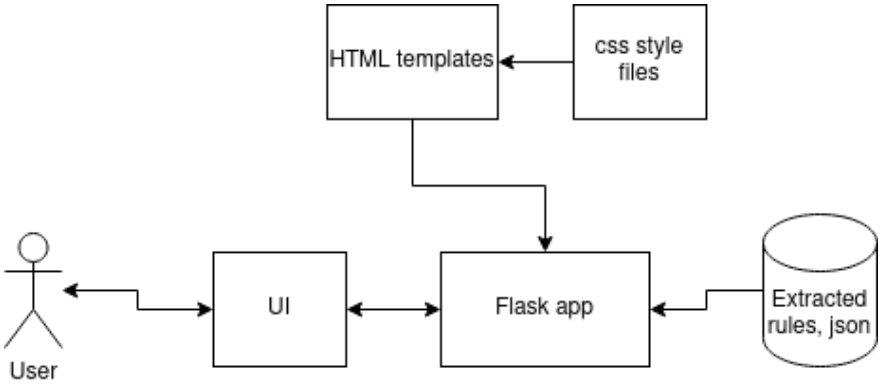

Figure 3: A scheme of website design

The rules for each language were all represented on the same format. That is, a JSON file, where the possible morphological tags are nested one into each other, and in the end the possible transitions are listed via a "|" sign. An example JSON file constructed for English rules can be viewed in fig. 4. Then, similar rules for the given languages were automatically highlighted by iterating through all the rules, checking that one language's tags are subset of another language's tags for the given rules. If the condition is true, and if texts of the rules are equivalent, they are considered to be a common rule and that information is highlighted on the website. This method obviously underestimates the amount of similar rules, as it only counts the exact equivalent ones.

5 Discussion

We have managed to extract morphological rules for 6 languages and to compare the results of the transfer of the model for classification of analogies.

Interesting findings have been made, such as the relative closeness of Karelian, Finnish and Swedish. Karelian is also quite close to English and German. In addition, Mezquital Otomi has been found to be really far from other languages.

[4] `https://inter-language-analogies.herokuapp.com/`

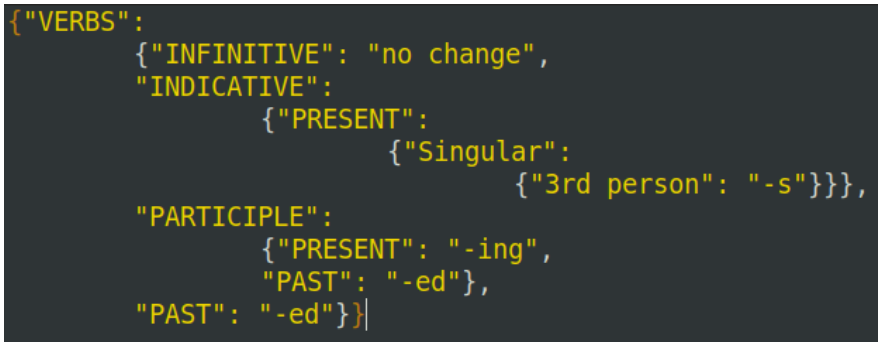

Figure 4: Rules extracted from English corpus in a JSON format.

We have also built a website to make it easier to compare the extracted rules and study the best performing analogies.

Some further research still needs to be made, as some transfer results were quite different from what was expected. Indeed, some of these results are surprising because there is no common morphological rules between the languages they belong to.

As was noticed in the analysis, Karelian is very close to Finnish and also scored good transfers to Swedish and English. Knowing this, a new study could try to link Finnish to the Germanic group (containing German and Swedish) through Karelian.

In addition, more languages could be added to extend this work.

Acknowledgment

We would like to express our gratitude to our professors, Miguel Couceiro and Esteban Marquer, for their numerous useful recommendations and remarks throughout this research.

Experiments presented in this paper were partially carried out using the Grid'5000 testbed[5], supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations.

We are also thankful to Bryan Pereira, who provided a server to host our website.

References

[1] K. Rexová, D. Frynta, and J. Zrzavỳ, "Cladistic analysis of languages: Indo-european classification based on lexicostatistical data," *Cladistics*, vol. 19, no. 2, pp. 120–127, 2003.

[2] S. Alsaidi, A. Decker, P. Lay, E. Marquer, P.-A. Murena, and M. Couceiro, "A neural approach for detecting morphological analogies," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021, pp. 1–10.

[3] E. Vylomova, J. White, E. Salesky, S. J. Mielke, S. Wu, E. M. Ponti, R. Hall Maudslay, R. Zmigrod, J. Valvoda, S. Toldova, F. Tyers, E. Klyachko, I. Yegorov, N. Krizhanovsky, P. Czarnowska, I. Nikkarinen, A. Krizhanovsky, T. Pimentel, L. Torroba Hennigen, C. Kirov, G. Nicolai, A. Williams, A. Anastasopoulos, H. Cruz, E. Chodroff, R. Cotterell, M. Silfverberg, and M. Hulden, "SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection," in *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Online: Association for Computational Linguistics, Jul. 2020, pp. 1–39. [Online]. Available: https://aclanthology.org/2020.sigmorphon-1.1

[4] H. H. Hock and B. D. Joseph, "Comparative method: Establishing language relationship," in *Language History, Language Change, and Language Relationship*. De Gruyter Mouton, 2009, pp. 427–454.

[5] A. McMahon and R. McMahon, "Finding families: Quantitative methods in language classification," *Transactions of the Philological Society*, vol. 101, no. 1, pp. 7–55, 2003.

[5] https://www.grid5000.fr

[6] G. Longobardi, S. Ghirotto, C. Guardiano, F. Tassi, A. Benazzo, A. Ceolin, and G. Barbujani, "Across language families: genome diversity mirrors linguistic variation within e urope," *American journal of physical anthropology*, vol. 157, no. 4, pp. 630–640, 2015.

[7] V. Blažek *et al.*, "On the internal classification of indo-european languages: survey," *Linguistica online*, 2005.

[8] R. Fam and Y. Lepage, "Tools for the production of analogical grids and a resource of n-gram analogical grids in 11 languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1171

[9] M. Grinberg, *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.