Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish)

Mateusz Lango[1,2], **Magda Ševčíková**[2] **and Zdeněk Žabokrtský**[2]
[1]Poznan University of Technology, Faculty of Computing, Institute of Computing Science
[2]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
mlango@cs.put.edu.pl,{sevcikova,zabokrtsky}@ufal.mff.cuni.cz

Abstract

The paper presents a semi-automatic method for the construction of derivational networks. The proposed approach applies a sequential pattern mining technique in order to construct useful morphological features in an unsupervised manner. The features take the form of regular expressions and later are used to feed a machine-learned ranking model. The network is constructed by applying resulting model to sort the lists of possible base words and selecting the most probable ones. This approach, besides relatively small training set and a lexicon, does not require any additional language resources such as a list of alternations groups, POS tags etc. The proposed approach is applied to the lexeme sets of two languages, namely Polish and Spanish, which results in the establishment of two novel word-formation networks. Finally, the network constructed for Polish is merged with the derivational connections extracted from the Polish WordNet and those resulting from the derivational rules developed by a linguist, resulting in the biggest word-formation network for that language. The presented approach is general enough to be adopted for other languages.

Keywords: derivation, derivational morphology, Polish, Spanish, lexical network, learning to rank, sequential pattern mining

1. Introduction

Derivational morphology has moved into focus of Natural Language Processing (NLP) only recently. For some languages, we observe a significant research effort in the construction of resources specialized in derivation, e.g. DerivBase (Zeller et al., 2013) for German, Démonette (Hathout and Namer, 2014) for French, DerivBase.Hr (Šnajder, 2014) for Croatian, DeriNet (Ševčíková and Žabokrtský, 2014; Žabokrtský et al., 2016) for Czech, or Word Formation Latin (Litta et al., 2016). However, for many other languages the data resources which provide information about derived words are scarce or even lacking. Unfortunately, the creation of such resources requires a considerable human effort and is highly time-consuming.

In this paper, we propose a method for semi-automatic construction of a derivational network which can be applied to under-resourced languages. The proposed approach requires only two resources: a set of lexemes and a relatively small training set which should contain examples of derived lexemes with their base words.

First, the method is looking for frequent patterns in a given lexeme set in order to automatically detect character-level regularities in the word construction of the language under consideration. The mentioned process is conducted in a completely unsupervised manner and no hand-crafted rules are used. Given those frequent patterns, each lexeme can be described by the presence (or absence) of a given pattern. Such lexeme descriptions serve as feature vectors for machine learning techniques and allow to train a ranking model.

Later, we start the network construction by making an attempt to find a base word for each lexeme. In order to avoid the consideration of all possible parents for each lexeme, only a candidate set of most morphologically similar words is considered. As a measure of morphological similarity, the Proxinette distance (Hathout, 2009; Hathout, 2014) is used. Next, the previously trained ranker is applied to order each candidate set. Finally, for each lexeme the highest ranked candidates are selected but only if the confidence of the ranker is high.

In order to evaluate the proposed approach, the method is applied on sets of Polish and Spanish lexemes, resulting in two new derivational networks for those languages. Even though these languages belong to different families, both have rich inflectional and derivational morphology, with the derivation as the most productive word-formation process. For Polish more than 50 thousand derivational pairs were automatically detected, and above 18 thousand for Spanish.[1] Moreover, the experimental evaluation demonstrates the high precision of constructed networks.

Finally, the Polish network was enriched with additional connections extracted from Polish WordNet (Maziarz et al., 2016) resulting in the Polish Word-Formation Network which contains 261 822 lexemes with more than 192 thousand connections. To the best of our knowledge, it is the biggest language resource of derivational morphology for Polish.

The rest of the paper is organized as follows. In Section 2. we briefly present related work. Section 3. contains a description of the proposed approach. In Section 4., we proceed with the discussion of experimental evaluation and with the analysis of the resulting resources. Finally, in Section 5. we draw conclusions and discuss lines of future research.

2. Related work

Morphology was intensively studied by the NLP community, with the research primarily concentrated on inflectional morphology. However, in recent years researchers

[1] Spanish lexeme set was considerably smaller.

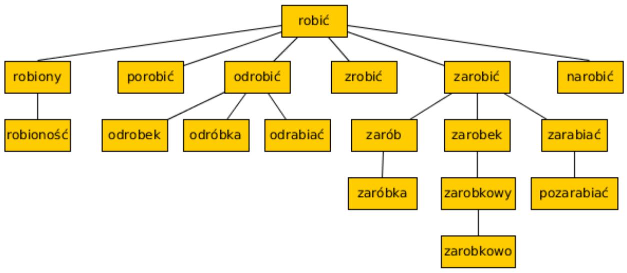

Figure 1: An example of a tree structure in word-formation network.

noticed the potential of derivational morphology to improve the performance in many important areas of NLP, which caused the development of novel language resources which focus on word formation.

One novel type of such resource is word-formation network which represents information about derivational morphology in a form of the graph. In such networks, the derivational relations are represented as directed edges between lexemes. In this work, as in the majority of related works, we consider word-formation networks whose connected components have a tree structure (see an example on Figure 1). This means that derivatives can have only one base word, hence, e.g. compounding is not considered.

Although such networks were created for some languages (see Sect. 1.), there is still a demand for the creation of such resources for many other languages. For instance, there is no word-formation network for two languages which are considered in the present work, Polish and Spanish. For those languages the number of works on automatic detection of derivatives is also quite limited.

Piasecki et al. (2012) propose an approach based on bootstrapping and supervised learning in order to construct derivational rules for Polish. Also, a hand-crafted list of alternation groups is used in order to make the derived words as similar as possible to the base word, allowing for a more effective application of derivational rules. Besides using this additional resource, the approach uses a relatively large training set of 15 718 examples (approx. 10 times bigger than the one used in this work) and deals only with derivation by suffixation or prefixation. Recently, an algorithm for the automatic pairing of perfective and imperfective Polish verb forms has been developed (Kaleta, 2017). The algorithm fully relies on the hand-crafted rules of Polish morphology, hence it is inapplicable for the construction of the network for other languages.

The literature related to Spanish is quite more extensive. Vilares et al. (2001) develop a system for automatic generation of morphological families[2] in order to improve their information retrieval systems. However, the derivational rules are incorporated into the system by a human expert rather than automatically learned. There were also some attempts to automatically discover affixes and suffixes used in Spanish derivations (Urrea, 2000), but no language resources were created. Spanish was also one of the languages studied by Baranes and Sagot (2014) in their language-independent approach. Although their approach is completely unsupervised, it requires POS tagging and, particularly for Spanish, it provides a considerably lower accuracy of extracted relations than for the other languages (73% for Spanish compared with 98% for English and German). In our opinion, it creates an opportunity for a use of supervised approaches which do not require large training sets nor additional handcrafted features or resources.

Sequential pattern mining Sequential pattern mining is one of the most important topics in the area of frequent pattern mining, and in the data mining in general (Han et al., 2007). The problem of sequential pattern mining is the extraction of all frequent subsequences with the support[3] greater than a specified threshold. Informally, a sequence a is a subsequence of the sequence b if one can remove items from the sequence b (without changing the order of them), to finally get the sequence a. Due to the importance of the task, a lot of approaches have been proposed. Among them, SPADE (Zaki, 2001) which bases on breadth-first search and Apriori pruning on the vertical data format. For formal definitions and a detailed review consult e.g. Mabroukeh and Ezeife (2010).

Learning to rank Learning to rank is a widely studied area of machine learning which was originally researched in the context of automatic ranking of web search results in the information retrieval community. However, it proved to be useful in many other areas such as statistical machine translation, see (Watanabe, 2012). The task of learning to rank is the construction of a model which is able to sort new objects according to their degrees of importance. The approaches for machine-learned ranking can be divided into three groups: the pointwise, the pairwise and the listwise approaches. Pointwise methods make use of classification or regression techniques in order to predict a score for each object given the query. An idea of predicting the order of each pair of objects is explored by pairwise methods. Finally, listwise approaches directly optimize metrics defined on a whole list of objects. For a review of those methods see e.g. Liu (2009).

3. Proposed approach

For the construction of the networks for Polish and Spanish, a machine-learned model is constructed first. This allows for building a large part of the word-formation network automatically, using a small training set only. Our focus is rather on the precision of discovered connections than on the coverage, hence a connection is constructed only if its probability exceeds a predefined threshold. For Polish then, a second step is done, namely the network established in the first step is merged with derivational connections extracted from Polish WordNet.

3.1. Machine learning approach: Polish and Spanish

Our machine learning approach can be split into several steps. First, a sequential pattern mining method is used to

[2] In the context of word-formation networks, a morphological family is a connected component of the network.

[3] The support of a subsequence is equal to the number of sequences which contain that subsequence in the database.

construct features describing lexemes in an unsupervised manner. Then, for each lexeme in the lexicon, we construct a list of possible base words using nearest-neighbors search. Finally, a machine-learned ranker is trained and consecutively applied to each candidate list. If the model ranks one position on the list much higher than any other, we create a connection between the lexeme and the selected candidate. In the following paragraphs, we will discuss each step in detail.

As the first stop of our approach, a sequential pattern mining is applied to the lexicon in order to find frequent subsequences. To perform this task we selected the SPADE algorithm (Zaki, 2001) since it is computationally efficient for usual lexicon sizes and its implementation was easily available. Any other algorithm which solves the sequential pattern mining problem could be adopted in a straightforward way. The algorithm treats each word as a sequence of characters and the lexicon is interpreted as a database of them. Hence, the resulting subsequences are in fact lists of characters which often occur in a particular order. The examples of such frequent subsequences in Polish lexicon are $\{$n, i, e$\}$, $\{$o, w, y$\}$ and $\{$n, o, ś, ć$\}$. Our hypothesis is that by finding subsequences covered by a lot of words, we will be able to discover useful morphological patterns in the lexeme set. We hope that those patterns will be useful in the feature construction for machine learning techniques.

To unify notation with the following paragraphs, we will represent each frequent subsequence as a regular expression. For example, the aforementioned subsequence $\{$n, i, e$\}$ will be further denoted as ˆ⋆n⋆i⋆e⋆$ where ˆ and $ mark the beginning and the end of the word, and ⋆ represents any string (including an empty one). At this point one faces two problems: first, the extracted patterns are too general and second, the number of frequent subsequences is large. Hence, we proceed with a procedure whose goal is to make the patterns more specific but also more meaningful. Later, a method for pruning the set of frequent subsequences is introduced.

In order to make patterns more specific, we apply a greedy approach which iteratively tries to delete one of the symbols of any string (⋆) from the pattern. If the deletion of that symbol results in a small decrease in the support (less than a threshold provided by the user), the newly created regular expression is accepted. The execution of this procedure results in more specific patterns with some of them having a linguistic interpretation. For example, ˆ⋆n⋆i⋆e⋆$ is replaced by ˆnie⋆$ which is a prefix used for the creation of negated forms in Polish, e.g. *dobry* (good) and *niedobry* (bad); ˆ⋆o⋆ś⋆ć⋆$ is converted to ˆ⋆ość$ which is a common suffix for Polish nouns, e.g. *męski* (manly) and *męskość* (manhood); *żwawo* (briskly) and *żwawość* (briskness). Since many other methods for automatic discovery of suffixes and prefixes has been proposed in the literature, it is important to note that our approach is able to detect more complex patterns than affixation only. For example patterns like ˆnie⋆ość$ or ˆ⋆cz⋆ność$ are also constructed.

Next, one must deal with a high number of patterns generated. We have observed that some of the patterns match almost the same lexemes. For example, ˆ⋆cz⋆noś⋆$ and ˆ⋆cz⋆ność$ have approximately the same support and cover the same lexemes, so keeping both of them seems to be redundant. In order to detect such pairs of redundant patterns, we perform a specific correlation analysis. First, we describe each lexeme in the lexicon by a binary feature vector. Each previously created regular expression is converted into one feature which takes 1 when the pattern matches the lexeme, and 0 otherwise. Having such representation of the lexicon, we are able to measure the association between patterns by calculating the phi coefficient (see e.g. Kotz et al., 2006). We identify a pair of patterns as a redundant one if its phi coefficient is greater than 95 %. Then, we shrink the number of patterns by selecting the most specific pattern from each indicated pair. This results in a considerably smaller set of patterns.

The aforementioned representation of lexemes as binary feature vectors enable us to use machine learning techniques for the construction of word-formation network. Since the network has a tree-like structure, we construct it by finding a base word for each lexeme. We approach this problem by sorting the list of possible base words for a lexeme from the most plausible ones to the least probable ones. Because sorting a whole lexicon for each lexeme is infeasible, we restrict the list of possible base words to the 100 most similar words according to Proxinette measure (Hathout, 2009; Hathout, 2014). Such constructed candidate list is sorted by a ranker, which is previously trained on a relatively small training set provided by a linguist. A connection in the network is established when the difference between the rank of the first and second element on the sorted list exceeds a threshold provided by the user.

3.2. Enriching the Polish network

3.2.1. Extraction of derivational connections from WordNet

The connections constructed in the Polish network by the machine learning method, the connections extracted from Polish WordNet (Maziarz et al., 2016) are added. Polish WordNet contains many relations which store information about derived words such as "feminity" which links masculine nouns with its feminine counterparts, "inhabitant" which connects geographical names with the name of their inhabitants, "aspectuality" which relates verbs of different aspects and many others. A list of the main relations related to derivation, together with descriptions can be found in (Piasecki et al., 2012). In order to merge the automatically constructed network with the connections from the WordNet, we iteratively analyze each lexeme for which a base word was not discovered by the machine learning approach. For each such lexeme, we try to find a base word using one of the 53 relations which were extracted from the WordNet. Additionally, the direction of some of the relations was reversed, in order to obtain a coherent network structure.

3.2.2. Adding connections by derivational rules

Although our machine learning approach is able to detect pairs of words which are derived in a fairly complex way, we discovered that many new connections can be added to the network by following some simple handcrafted rules.

A language expert created 20 derivational rules for Polish which are most productive. The derivational rules are given in the form of regular expressions and express suffix/prefix addition or substitution. In order to improve rule's performance, the base word indicated by the rules is accepted only if it already exists in a lemma set.

3.2.3. Elimination of cycles

One disadvantage of constructing a network in an iterative way is the lack of any prevention from creating a cycle. In our approach, a network is constructed by searching a parent for each lemma. This ensures that every node in a graph has at most one outgoing edge and causes that many of the graph's components will have the desired tree structure. Nevertheless, there is no guarantee that a cycle in a graph will not be created (e.g. $A \rightarrow B \rightarrow C \rightarrow A$). To handle such situation, we use a simple heuristic to eliminate graph's cycles. Our heuristic relies on the rather naive assumption that the derivatives are usually longer than the base word since they are often created by adding to it a suffix or a prefix. Hence, we eliminate cycles by iterating over them and removing the first connection between a shorter child and a longer parent. If all words in a cycle have the same length, we drop a random connection.

4. Experimental evaluation

4.1. Machine learning approach

As the base for the creation of the Polish Word-Formation Network, we have used the Grammatical Dictionary of Polish (Saloni et al., 2017), which is a comprehensive lexicon of the Polish language, covering more than 261 thousand lexemes. This dictionary is quite popular in the Polish NLP community as it was used in the creation of many resources and tools such as Morfeusz morphological analyzer (Woliński, 2006) or the Great Dictionary of Polish (Żmigrodzki, 2011).

Using this lexeme set, we created a training set which consists of 1500 pairs of base words with their derivatives. Polish native speakers, who created the training set, were asked to provide examples of as many different derivation schemas as possible. The construction of the training set took approximately 12 man-hours.

In the implementation of our machine learning approach, we have used the SPMF data mining library (Fournier-Viger et al., 2016) for the extraction of frequent subsequences. The implementation of the ranker based on Gradient Boosting Decision Trees was taken from XGBoost machine learning library (Chen and Guestrin, 2016). The ensemble of 100 decision trees was used with the maximum depth of a tree set to 40.

The application of the SPADE algorithm with the minimal support set to the 1% of the lexicon size resulted in roughly 27 thousand frequent subsequences. However, our filtering technique based on the phi coefficient limited the set of patterns to 13 441 regular expressions. Using that set of expressions, we created a feature vector together with two additional features: the length of the common prefix and the length of the common suffix. The training set for a ranker consist of automatically constructed groups containing a lexeme together with 100 candidates. In each group, the rank of the correct base word is set to 1, whereas the rank of the rest of candidates is set to 0.

We evaluated our approach using 5-fold cross-validation and we obtained the accuracy of 82.33% without applying any threshold on the confidence of the ranker. However, since we prefer precision to coverage, we have chosen a threshold which allowed us to obtain 98.8% of precision with the recall of 38.2%. By applying this thresholded model to the set of lexemes, we were able to create more than 53.5 thousand links in the network.

Since our dataset does not contain negative examples (all of the words are derivatives), we decided to perform a manual verification of the precision of discovered connections. We created a random sample of 200 connections and manually checked each of them. Only 6 of them were incorrect, so we estimate the precision of created connections to 97%. Then, we estimated the recall to 26.5% by sampling 200 lexemes and verifying if their parent exists in the network.

Encouraged by the high precision of constructed network, we decided to augment the training set with the connections from the network. In this way, we have around 55 thousands lexeme pairs in the training set without any additional effort of language experts, although the training set has become somewhat noisy. By applying our approach to this larger set, we have obtained almost 75 thousand connections. The manual evaluation on a random sample of 200 connections yielded 95% of precision and 34% of recall.

Since our machine learning approach is general enough to be adopted for other languages, we have also evaluated its performance on the set of 159 035 Spanish lexemes taken from Leffe (Molinero et al., 2009) lexicon. Using the training set of 1026 examples, we obtained 98.1% precision with the recall of 27.7%. The model was able to discover 18.5 thousand connections with the manually evaluated precision of 85% and recall of 44%. The precision for Spanish is considerably lower than for Polish due to a large amount of French proper names in the lexicon which were not taken into account by linguists while creating the training set. The precision without taking them into account is much closer to the one for Polish (94.9%).

Language	Method	Precision	Recall
Polish	Rules	90.0%	66.7%
	ML	95.0%	34.0%
Spanish	Rules	84.0%	59.0%
	ML	94.9%	44.0%

Table 1: A comparison of precision and recall for machine learning and rule-based approach

4.2. Comparison against the rule-based approach

We compared the performance of our machine learning approach with substitution rules provided by a human expert. We asked a linguist to provide us with 20 highly productive and reliable derivational rules for each language. The results are presented in Table 1.

For both languages, the recalls obtained by hand-crafted

Step	# of conn.	Precision	Recall
Machine Learning	53 487	97.0%	26.5%
Machine Learning (retraining)	74 985	95.0%	34.0%
Merge with WordNet	110 553	94.5%	47.0%
Derivational rules	192 289	95.0%	72.0%

Table 2: The number of connections, precision and recall of the Polish Word-Formation Network evaluated after each step of the construction.

rules are considerably higher than those for machine learning approach. However, our method is significantly better in terms of precision. Since the provided rules were the very productive ones, we would like to emphasize that further development of a higher number of hand-crafted rules will result in rather small increases of the recall. Furthermore, the impact of applying the rules in a particular order will significantly grow with their number. This will cause that substitution rules will have to be designed with special caution and be thoroughly tested. Conversely, one can expect the quality of machine learning approach to improve together with the addition of more training examples. Also, a particular rule typically handles only one type of derivation whereas machine learning approach it is able to connect lexemes derived in a plenty of different ways.

4.3. Construction of the Polish Word-Formation Network

In order to create a bigger language resource, we decided to merge our automatically detected connections with other available resources.

First, we extracted derivational connection from *Słowosieć*, the Polish WordNet. By applying these connections to our network's lexemes, we were able to construct about 52 thousand links. Finally, merging the automatically discovered connections together with links constructed on the basis of the information from the WordNet resulted in a network of more than 110.5 thousand connections.

The manual verification of precision and recall, performed as was previously described in Section 4.1., yielded 94.5% of precision and 47% of recall. The addition of connections from the Polish WordNet, introduced some new errors mainly due to the incorrect direction of the link or due to the omission of lexemes in the chain of derivation e.g. *poatomowy* is directly connected with *atom* whereas we would expect *atom* (atom) →*atomowy* (atomic) →*poatomowy* (property of the result of a nuclear explosion).

Finally, a set of 20 highly-productive hand-constructed derivational rules was applied to the lemma set. The resulting network has 192 289 connections and, to the best of our knowledge, it is the biggest word-formation network for Polish language. The manual verification of the resource yielded 95% of precision and 72% of recall. This demonstrate that the resource is reliable and has significant coverage of Polish derivations.

Both the Polish Word-Formation network and the data for Spanish are available at `ufal.mff.cuni.cz/derinet`.

4.4. Visualization of word-formation networks

To better evaluate the created resource, we decided to visualize some parts of the Polish Word-Formation Network using a visualization tool developed by Vidra and Žabokrtský (2017). In Figure 2 one can see the derivation tree for a proper noun *Szczecin* (a Polish city). The derived adjective (*szczeciński*), the name of its inhabitants (*szczecinianin*) and even a name of another smaller city (*Szczecinek*) are correctly connected in a network, despite the fact that e.g. the word-formation process of *szczeciński* is potentially difficult because of occurring alternation of *n* to *ń*. One can also observe that the tree has multiple layers with further derivatives e.g. *podszczeciński* (property of Szczecin's surrounding area). Furthermore, a derivation tree for *herbata* (tea) in Figure 3 also seems correct, containing derived adjectives (e.g. *herbaciany*), nouns (e.g. *herbaciarnia* tea shop), negated forms (e.g. *nieherbaciany*) and diminutives (*herbatka*). One can notice the lack of adverbs, however they are absent in our lemma set for this particular case.

The presented approach can also handle some untypical cases. For example, *ponauczać* is a verb which does not have an imperfective form (see Figure 4). Even though a blind application of morphological rules would result with *ponauczyć* (e.g. *nauczyć* and *nauczać* are correctly connected), this verb is correctly connected with another perfective form *nauczać*.

However, on visualizations also some errors in the network structure can be found. For example, in Figures 5 and 6 we present the derivation trees for *karta* (sheet) and *karty* (card game), respectively. First of all, we would rather expect these two trees to be mutually connected. The lack of the connection between *karta* and *karty* causes the absence of a whole derivational subtree in the derivations of *karta*. Second, the words *Djakarta/Dzakarta* (Jakarta) are definitely not derived from *karta* and *gokartowość* (noun from go-cart) is not derived from *kartowość*. Such errors related to the words adopted from other languages are quite frequent. Another example of such error can be found in Figure 7 where *pastoforium* is derived from Greek.

We have also visualized some parts of the Spanish Word-Formation Network. On Figure 8 the derivation tree for the verb *ilustrar* (to illustrate) is presented. Similarly to the visualizations of the Polish network, one can see that the lexeme is correctly linked to many derivatives. The derived adjectives like *ilustrativo* illustrative or *ilustrado* illustrated as well as derived nouns (e.g. *ilustración* illustration) are properly connected in the network. However, as a result of a low recall, many lexemes have too small derivation trees. A good example of this is the derivation tree for the verb *hacer* (to do) which is presented on Figure 9. Only the noun *hecedor* (maker) is connected with this verb and the absence of many other lexemes is rather evident e.g. *rehacer* (to redo) or *deshacer* (to undo) are lacking.

5. Conclusion and future research

In the present paper, a new semi-automatic approach for the construction of word-formation networks is presented. In particular, it applies sequential pattern mining to construct useful morphological features. To the best of our knowledge, no one has used these techniques in that context.

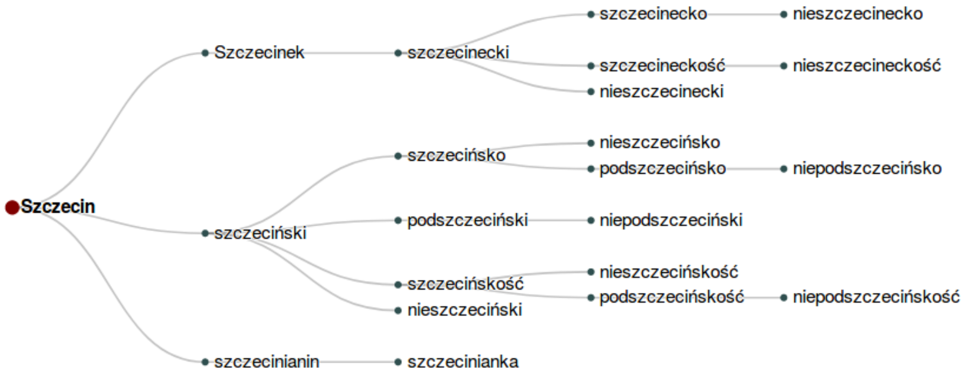

Figure 2: The noun *Szczecin* (the capital of West Pomeranian Voivodeship) and derivationally related lexemes (displayed as a tree structure) in the Polish Word-Formation Network.

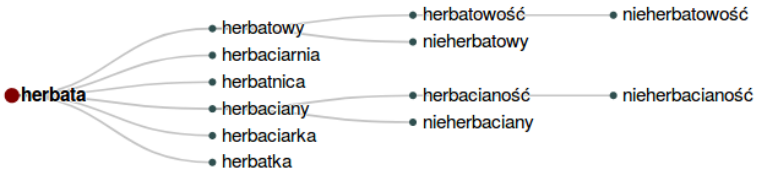

Figure 3: The noun *herbata* (tea) and derivationally related lexemes (displayed as a tree structure) in the Polish Word-Formation Network.

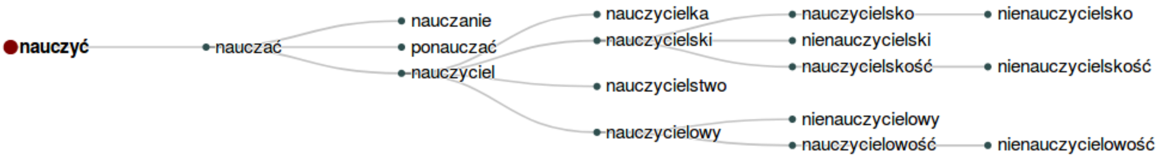

Figure 4: The verb *nauczyć* (to learn *perf*) and derivationally related lexemes (displayed as a tree structure) in the Polish Word-Formation Network.

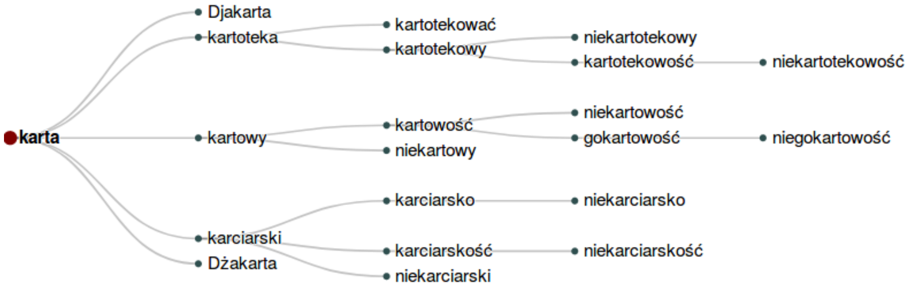

Figure 5: The noun *karta* (sheet) and derivationally related lexemes (displayed as a tree structure) in the Polish Word-Formation Network.

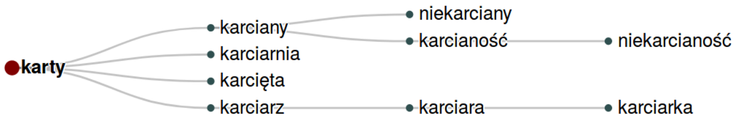

Figure 6: The noun *karty* (card game) and derivationally related lexemes (displayed as a tree structure) in the Polish Word-Formation Network.

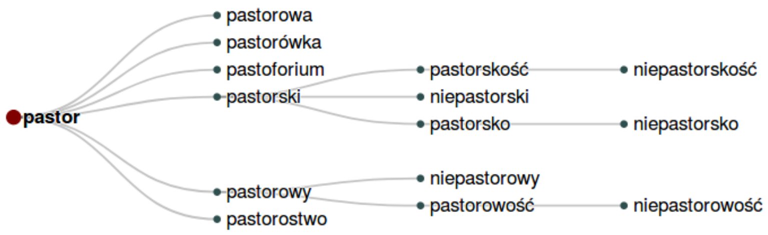

Figure 7: The noun *pastor* (pastor) and derivationally related lexemes (displayed as a tree structure) in the Polish Word-Formation Network.

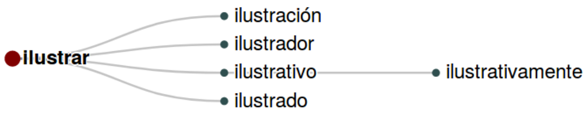

Figure 8: The verb *ilustrar* (to illustrate) and derivationally related lexemes (displayed as a tree structure) in the Spanish Word-Formation Network.

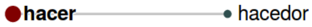

Figure 9: The verb *hacer* (to do) and derivationally related lexemes (displayed as a tree structure) in the Spanish Word-Formation Network.

Moreover, the approach was successfully evaluated on one Slavic and one Romance language, namely Polish and Spanish. For both languages, the newly proposed method discovered plenty of true connections between base words and their derivatives. The resources created by the method are characterized by a high precision, and their creation does not require large human effort.

Finally, this paper introduces the Polish Word-Formation Network which is the result of merging Polish WordNet with automatically discovered connections by our machine learning approach. The network is constructed over a large lexeme set and is characterized by high precision. Furthermore, it is also the biggest free language resource about Polish derivations.

However, the coverage of the created resources still needs to be improved. As a future research, the active learning framework (Settles, 2010) could be incorporated into our approach. The active learning methods allow for the iterative construction of machine learning models with the participation of human experts. In such approaches, training examples which should be annotated are chosen by the algorithm itself with the ultimate goal of constructing an accurate model using as few examples as possible. We hope that such extension of our work would lead to the significant improvement in coverage of constructed networks without significantly increasing the effort of human experts.

Currently, we explore the possibility of using a cross-lingual transfer to the fully automatic construction of word-formation networks. As we mentioned in the introduction, derivational resources have been already developed for a limited number of languages. This creates the possibility of employing those resources to the construction of word-formation networks for related languages. For instance, the Czech DeriNet (Vidra et al., 2017) may be used to extend the proposed Polish network. Each connection from the Czech network could be translated into Polish using one of the available dictionaries such as Treq (Škrabal and Martin, 2017). Then, some notion of morphological similarity should be adopted in order to eliminate incorrect connections. In the context of the present work, the constructed candidate sets may be used to check if the indicated base word is morphologically related. The preliminary experiments which we have run so far are quite promising. The described approach was able to find over 40 thousand connections but with rather poor precision. Roughly 40% of the discovered connections were correct, however, one-third of errors could be simply fixed by reversing the direction of relation. We believe that there are many possibilities to improve this result e.g. by taking the probability of translation into account.

Another issue of the current approach is that it is limited to derivation only. Hence, there is a need of developing novel methods for the automatic discovery of other word-formation processes, such as compounding. Additionally, some effort to design the representation of such enhanced resources is also needed.

Acknowledgements

This work was supported by the Grant No. GA16-18177S of the Grant Agency of the Czech Republic. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

Mateusz Lango acknowledges also the support from the Institute of Computing Science Statutory Funds and from the internship program of the Institute of Formal and Applied Linguistics.

6. Bibliographical References

Baranes, M. and Sagot, B. (2014). A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 2793–2799, Reykjavik, Iceland, May.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA. ACM.

Fournier-Viger, P., Lin, C., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., and Lam, H. T. (2016). The SPMF Open-Source Data Mining Library Version 2. In *Proc. 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III*.

Han, J., Cheng, H., Xin, D., and Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86.

Hathout, N. and Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5):125–168.

Hathout, N. (2009). Acquisition of Morphological Families and Derivational Series from a Machine Readable Dictionary. In Fabio Montermini, et al., editors, *Selected Proceedings of the 6th Décembrettes*, Cascadilla Proceedings Project, pages 166–180, Bordeaux, France.

Hathout, N. (2014). Phonotactics in morphological similarity metrics. *Language Sciences*, 46:71 – 83. Theoretical and empirical approaches to phonotactics and morphonotactics.

Kaleta, Z. (2017). Automatic Pairing of Perfective and Imperfective Verbs in Polish. In *Proceedings of the Eight Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Kotz, S., Read, C. B., Balakrishnan, N., and Vidakovic, B. (2006). *Encyclopedia of statistical sciences*. Wiley, 2nd edition.

Litta, E., Passarotti, M., and Culy, C. (2016). Formatio formosa est. Building a Word Formation Based Lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 185–189, Naples.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331.

Mabroukeh, N. R. and Ezeife, C. I. (2010). A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys (CSUR)*, 43(1):3.

Piasecki, M., Ramocki, R., and Maziarz, M. (2012). Recognition of Polish Derivational Relations Based on Supervised Learning Scheme. In *LREC*, pages 916–922.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.

Ševčíková, M. and Žabokrtský, Z. (2014). Word-Formation Network for Czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1087–1093, Reykjavik, Iceland, May.

Šnajder, J. (2014). DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 3371–3377, Reykjavik, Iceland, May.

Urrea, A. M. (2000). Automatic discovery of affixes by means of a corpus: A catalog of Spanish affixes. *Journal of quantitative linguistics*, 7(2):97–114.

Vidra, J. and Žabokrtský, Z. (2017). Online software components for accessing derivational networks. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology*, pages 129–139.

Vilares, J., Cabrero, D., and Alonso, M. A. (2001). Applying productive derivational morphology to term indexing of Spanish texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 336–348. Springer.

Watanabe, T. (2012). Optimized online rank learning for machine translation. In *Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 253–262. Association for Computational Linguistics.

Woliński, M. (2006). Morfeusz - a practical tool for the morphological analysis of Polish. In *Intelligent information processing and web mining*, pages 511–520. Springer.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1):31–60.

Zeller, B., Šnajder, J., and Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1201–1211.

Żmigrodzki, P. (2011). Polish Academy of Sciences Great Dictionary of Polish: history, presence, prospects. *Studies in Polish Linguistics*, 6(1):7–26.

Žabokrtský, Z., Ševčíková, M., Straka, M., Vidra, J., and Limburská, A. (2016). Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1307–1314, Portoroz, Slovenia.

7. Language Resource References

Maziarz, Marek and Piasecki, Maciej and Szpakowicz, Stanisław. (2016). *plWordNet*. Wrocław University of Technology, 3.0.

Molinero, M., Sagot, B., and Nicolas, L. (2009). A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe. In *RANLP 2009-Recent Advances in Natural Language Processing*.

Saloni, Zygmunt and Gruszczyński, Włodzimierz and Woliński, Marcin and Wlosz, Robert and Skowrńska, Danuta. (2017). *Slownik gramatyczny języka polskiego*. University of Warmia and Mazury.

Vidra, Jonáš and Žabokrtský, Zdeněk and Ševčíková, Magda and Kalužová, Adéla and Mediankin, Nikita and Straka, Milan. (2017). *DeriNet 1.5*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2422.

Škrabal, M. and Martin, V. (2017). Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii*, 99(2):245–260.