Paradigm classification in supervised learning of morphology

Malin Ahlberg
Språkbanken
University of Gothenburg
malin.ahlberg@gu.se

Markus Forsberg
Språkbanken
University of Gothenburg
markus.forsberg@gu.se

Mans Hulden
Department of Linguistics
University of Colorado Boulder
mans.hulden@colorado.edu

Abstract

Supervised morphological paradigm learning by identifying and aligning the longest common subsequence found in inflection tables has recently been proposed as a simple yet competitive way to induce morphological patterns. We combine this non-probabilistic strategy of inflection table generalization with a discriminative classifier to permit the reconstruction of complete inflection tables of unseen words. Our system learns morphological paradigms from labeled examples of inflection patterns (inflection tables) and then produces inflection tables from unseen lemmas or base forms. We evaluate the approach on datasets covering 11 different languages and show that this approach results in consistently higher accuracies vis-à-vis other methods on the same task, thus indicating that the general method is a viable approach to quickly creating high-accuracy morphological resources.

1 Introduction

Use of detailed and sophisticated morphological features has been found to be crucial for many downstream NLP tasks, including part-of-speech tagging and parsing (Tseng et al., 2005; Spoustová et al., 2007). However, creating an accurate wide-coverage morphological analyzer for a new language that can be used in tandem with other higher-level analyses is an arduous task.

Learning word inflection patterns by organizing related word-forms into morphological paradigms based on the longest common subsequence (LCS) found in an inflection table has recently been proposed as a method for supervised and semi-supervised induction of morphological processing tools from labeled data (Ahlberg et al., 2014). Also, the argument that the LCS shared by different inflected forms of a word—even if discontinuous within a word—corresponds strongly to a cross-linguistic notion of a 'stem' has later been advanced independently on grounds of descriptive economy and minimum description length (Lee and Goldsmith, 2014).

We used this idea in (Ahlberg et al., 2014) to create a relatively simple-to-implement system that learns paradigms from example inflection tables and is then able to reconstruct inflection tables for unseen words by comparing suffixes of new base forms to base forms seen during training. The system performs well on available datasets and results in human-readable and editable output. The longest common subsequence strategy itself shows little bias toward any specific morphological process such as prefixation, suffixation, or infixation. Using the model, we argued, a selection of ready-inflected tables could be quickly provided by a linguist, allowing rapid development of morphological resources for languages for which few such resources exist.

Potentially, however, the model's commitment to a simple suffix-based learner is a weakness. To assess this, we evaluate a similar LCS-based generalization system with a more refined discriminative classifier that takes advantage of substrings in the example data and performs careful feature selection. We show that much higher accuracies can be achieved by combining the LCS paradigm generalization strategy with such a feature-based classi-

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1024–1029,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

fier that assigns unknown words to the LCS-learned paradigm based on substring features taken from word edges. This holds in particular for languages where paradigmatic behavior is triggered by material in the beginning of a word (e.g. German verbs).

We present experiments on 18 datasets in 11 languages varying in morphological complexity. In all the experiments, the task is to reconstruct a complete inflection table from a base form, which usually corresponds to the lemma or dictionary form of a noun, verb, or adjective. The experiments are divided into two sets. In the first, we use an earlier dataset (Durrett and DeNero, 2013) of Finnish, German, and Spanish to compare against other methods of paradigm learning. In the second, we use a more comprehensive and complex dataset we have developed for 8 additional languages. This new dataset is less regular and intended to be more realistic in that it also features defective or incomplete inflection tables and inflection tables containing various alternate forms, naturally making the classification task substantially more difficult.[1]

Overall, supervised and semi-supervised learning of morphology by generalizing patterns from inflection tables is an active research field. Recent work sharing our goals includes Toutanova and Cherry (2009), Dreyer and Eisner (2011), which works with a fully Bayesian model, Dinu et al. (2012), Eskander et al. (2013), which attempts to learn lexicons from morphologically annotated corpora, and Durrett and DeNero (2013), who train a discriminative model that learns transformation rules between word forms. We directly compare our results against the last using the same dataset.

The paper is organized as follows: section 2 contains the experimental setup, section 3 the datasets, and section 4 the results and discussion.

2 Method

As a first step, our system converts inflection tables into paradigms using a procedure given in Hulden (2014). The system generalizes concrete inflection tables by associating the common symbol subsequences shared by the words (the LCS) with variables. These variables represent abstractions that at table reconstruction time can correspond to any sequence of one or more symbols. As many inflection tables of different words are identical after assigning the common parts to 'variables,' this procedure results in a comparatively small number of paradigms after being input a large number of inflection tables. The process is illustrated in Figure 1. During generalization, the forms that gave rise to a particular paradigm are stored and later used for training a classifier to assign unknown base forms to paradigms. Having a number of paradigms at our disposal by this generalization method, the task of reconstructing an inflection table for an unseen base form in effect means picking the correct paradigm from among the ones generalized, a standard classification task of choosing the right/best paradigm.

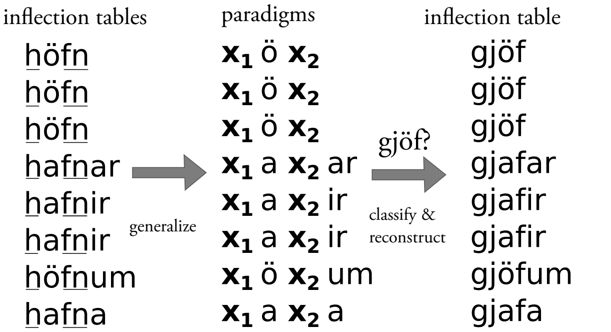

Figure 1: General overview of the system, exemplified using Icelandic nouns. First, a large number of inflection tables are generalized into a smaller number of paradigms; the generalization of the table for **höfn** 'harbor' into a paradigm is illustrated here. At classification time, an unknown base form is classified into one of the learned paradigms and its inflection table is reconstructed, illustrated here by **gjöf** 'present'.

After seeing a number of inflection tables generalized into abstract paradigms as described above, the task we evaluate is how well complete inflection tables can be reconstructed from only seeing an unknown base form. To this end, we train a "one-vs-the-rest" linear multi-class support vector machine (SVM).[2] For each example base form wb_i that is a member of paradigm p_j, we extract all substrings from wb_i from the left and right edges, and use those as binary features corresponding to the paradigm p_j.

[1] The data and the code is available at our website `https://svn.spraakbanken.gu.se/clt/naacl/2015/extract`

[2] Using LIBLINEAR (Fan et al., 2008) with L2-regularization.

For example, during training, the German verb **lesen** would have the following binary features activated: {#l, #le, #les, #lese, #lesen, #lesen#, lesen#, esen#, sen#, en#, n#}.

Before applying the classifier to an unseen base form and reconstructing the corresponding inflection table, many competing paradigms can be ruled out as being ill-matched simply by inspecting the base form. For example, the infinitive for the paradigm containing the English verb **sing** is generalized as x_1+**i**+x_2. At classification time of a verb like **run**, this paradigm can be ruled out due to incompatibility, as there is no **i** in **run**, and so the infinitive cannot be generated. Likewise, the Icelandic paradigm seen in Figure 1 can be ruled out for the base form **hest** 'horse', as the base form does not contain **ö**. The SVM-classifier may indeed suggest such paradigm assignments, but such classifications are ignored and the highest scoring compatible paradigm is selected instead. These additional constraints on possible base form-paradigm pairings are a general feature of the LCS-strategy and are not at all tied to the classification method here.

2.1 Feature selection

In order to eliminate noise features, we performed feature selection using the development set. We simultaneously tuned the SVM soft-margin penalty parameter C, as well as the length and type (prefix/suffix) of substrings to include as features. More concretely, we explored the values using a grid search over $C = 0.01 \ldots 5.0$, with a growing sequence gap (Hsu et al., 2003), as well as tuning the maximum length of anchored substring features to use $(3 \ldots 9)$, and whether to include prefix-anchored substrings at all $(0/1)$. In the second experiment, where cross-validation was used, we performed the same tuning procedure on each fold's development set.

3 Data

For the first experiment, we use the datasets provided by Durrett and DeNero (2013). This dataset contains complete inflection tables for German nouns and verbs (DE-NOUNS, DE-VERBS), Finnish verbs and nouns combined with adjectives (FI-VERBS, FI-NOUNADJ), and Spanish verbs (ES-VERBS). The number of inflection tables in this set ranges from 2,027 (DE-VERBS) to 7,249 (FI-VERBS). From these tables, 200 were held out for development and 200 for testing, following the splits that previous authors have used (Durrett and DeNero, 2013; Ahlberg et al., 2014) to ensure a fair baseline.[3]

For the second experiment, we collected additional inflection tables from Catalan (CA), English (EN), French (FR), Galician (GL), Italian (IT), Portuguese (PT), Russian (RU) (all from the FreeLing project (Padró and Stanilovsky, 2012)) and Maltese (MT) (Camilleri, 2013).[4] These inflection tables are often incomplete or defective and some contain very rarely occurring grammatical forms. Many alternate forms are also given. To avoid having to account for rare or historical forms, we filtered out grammatical forms (slots) that occur in less than $\sim 1\%$ of all inflection tables. We also performed an independent cross-check with Wiktionary and removed some inflection table slots that did not appear in that resource. We further limited the number of inflection tables to 5,000. In the second experiment, we also split each dataset into 5 folds for cross-validation (maximally 4,000 tables for training, 500 for development and 500 for testing for each fold).

4 Results and discussion

In the main results tables 1, 2, and 3 we report the per table accuracy and per form accuracy in reconstructing complete inflection tables from unseen base forms. The per table accuracy is the percentage of inflection tables that are perfectly reconstructed from the base form. The per form accuracy is the percentage of correct forms in the reconstructed table. The associated oracle scores, which independently provide a measure of generalization power of the LCS-method, represent the maximal percentage achievable by an oracle classifier that always picks

[3] The development and test data for the first experiment had been filtered to not contain any of the 200 most frequently occurring forms in the language (Durrett and DeNero, 2013); this may result in an easier classification task because the maneuver in effect ensures that words belonging to irregular paradigms—i.e. those which would otherwise be difficult to classify correctly—are never evaluated against.

[4] The FreeLing data also included Russian verbs. However, this data set was deemed too incomplete to be useful and was left out.

Data	Per table accuracy			Per form accuracy			Oracle acc. per form (table)
	SVM	AFH14	D&DN13	SVM	AFH14	D&DN13	
DE-VERBS	**91.5**	68.0	85.0	**98.11**	97.04	96.19	99.70 (198/200)
DE-NOUNS	**80.5**	76.5	79.5	**89.88**	87.81	88.94	100.00 (200/200)
ES-VERBS	**99.0**	96.0	95.0	**99.92**	99.52	99.67	100.00 (200/200)
FI-VERBS	**94.0**	92.5	87.5	**97.14**	96.36	96.43	99.00 (195/200)
FI-NOUNS-ADJS	**85.5**	85.0	83.5	**93.68**	91.91	93.41	100.00 (200/200)

Table 1: Results on experiment 1. Here AFH14 stands for Ahlberg et al. (2014) and D&DN for Durrett and DeNero (2013). The SVM-columns show the results of the current method.

the best learned paradigm for an unseen base form. In experiment 2, where the correct forms may consist of several alternatives, we only count a form as correct if all alternatives are given and all are correct. For example, the verb **dream** in English lists two alternative past participles, **dreamed** and **dreamt**, which both must be reconstructed for the past participle form to count as being correct.

Experiment 1

The accuracies obtained on the first three-language comparison experiment are shown in Table 1. Here, we see a consistent improvement upon the *maxsuff*-strategy (AFH14) that simply picks the longest matching suffix among the base forms seen and assigns the unseen word to the same paradigm (breaking ties by paradigm frequency), as well as improvement over other learning strategies (D&DN13). Particularly marked is the improved accuracy on German verbs. We assume that this is because German verb prefixes, which are ignored in a suffix-based classifier, contain information that is useful in classifying verb behavior. German verbs that contain so-called inseparable prefixes like **miss-, ver-, wider-** do not prefix a **ge-** in the past participle form. For example: **kaufen** ∼ **gekauft**, **brauchen** ∼ **gebraucht**, **legen** ∼ **gelegt**, but **verkaufen** ∼ **verkauft**, **widerlegen** ∼ **widerlegt**, **missbrauchen** ∼ **missbraucht**, reflecting the replacement of the standard **ge-** by the inseparable prefix. There are many such inseparable prefixes that immediately trigger this behavior (although some prefixes only occasionally show inseparable behavior), yet this information is lost when only looking at suffixes at classification time. This analysis is supported by the fact that, during feature selection, German verbs was the only dataset in this first experiment where word prefixes were not removed by the feature selection process.

Experiment 2

The results of the second experiment are given in tables 2 (per table accuracy) and 3 (per form accuracy). The tables contain information about how many inflection tables were input on average over 5 folds to the learner (**#tbl**), how many paradigms this reduced to (**#par**), and how many forms (slots) each paradigm has (**#forms**). The **mfreq** column is a baseline where the classifier always picks the most populated paradigm, i.e. the paradigm that resulted from combining the largest number of different inflection tables by the LCS process. The **AFH14** shows the performance of a maximal suffix matching classifier, identical to that used in Ahlberg et al. (2014).

Discussion

Overall, the results support earlier claims that the LCS-generalization appears to capture paradigmatic behavior well, especially if combined with careful classification into paradigms. There is a clear and consistent improvement over baselines that use the same data sets. In addition, the SVM-classifier yields results comparable, and in many cases better, to using a maximum suffix classifier and additionally having access to raw corpus data in the language, a semi-supervised experiment reported separately in Ahlberg et al. (2014). In this work we have not attempted to extend the current method to such a semi-supervised scenario, although such an extension seems both interesting and possible.

Data	#tbl	#par	mfreq	AFH14	SVM	Oracle
DE-N	2,210	66	18.99	76.09	**77.68**	98.99
DE-V	1,621	125	52.77	65.02	**83.59**	95.45
ES-V	3,243	90	70.42	92.25	**93.48**	96.59
FI-N&A	4,000	233	26.52	**83.20**	82.84	98.12
FI-V	4,000	204	43.04	**91.88**	91.64	94.76
MT-V	826	200	10.68	18.83	**38.64**	85.63
CA-N	4,000	49	44.12	94.00	**94.92**	99.44
CA-V	4,000	164	60.44	90.76	**93.40**	98.48
EN-V	4,000	161	77.12	89.40	**90.00**	97.40
FR-N	4,000	57	92.16	91.60	**93.96**	98.72
FR-V	4,000	95	81.52	93.72	**96.48**	98.80
GL-N	4,000	24	88.36	90.48	**95.08**	99.80
GL-V	3,212	101	45.21	58.92	**60.87**	98.95
IT-N	4,000	39	83.84	92.32	**93.76**	99.40
IT-V	4,000	115	63.96	89.68	**91.56**	98.68
PT-N	4,000	68	74.52	88.12	**90.88**	99.04
PT-V	4,000	92	62.00	76.96	**80.20**	99.20
RU-N	4,000	260	15.76	64.12	**66.36**	96.80

Table 2: Per table accuracy results on the second experiment. 5-fold cross-validation is used throughout. The **#tbl**-column shows the number of inflection tables input to the LCS-learner and the **#par** column shows the number of resulting unique paradigms. The **mfreq**-column illustrates a baseline of simply picking the most frequent paradigm, while **AFH14** is the strategy of finding the longest suffix match to the base forms in the training data (Ahlberg et al., 2014). The **SVM**-column shows the results discussed in this paper.

Data	#forms	mfreq	AFH14	SVM	Oracle
DE-N	8	57.36	89.72	**90.25**	99.69
DE-V	27	87.35	**96.12**	95.28	99.20
ES-V	57	93.80	98.72	**98.83**	99.47
FI-N&A	233	52.15	91.03	**91.06**	98.95
FI-V	54	70.38	**95.27**	95.22	96.76
MT-V	16	39.75	54.66	**61.15**	95.49
CA-N	2	71.30	96.89	**97.33**	97.93
CA-V	53	86.89	98.18	**98.89**	99.77
EN-V	6	91.43	95.93	**96.16**	99.28
FR-N	2	93.24	92.48	**94.68**	99.08
FR-V	51	91.47	97.09	**98.33**	99.02
GL-N	2	91.92	92.82	**95.38**	99.78
GL-V	70	94.89	**98.48**	98.32	99.67
IT-N	3	89.36	93.38	**94.59**	97.44
IT-V	51	89.51	97.76	**98.21**	99.64
PT-N	4	83.35	89.78	**91.97**	98.60
PT-V	65	92.62	96.81	**97.20**	99.68
RU-N	12	25.16	88.19	**89.35**	99.15

Table 3: Per form accuracy results on the second experiment. 5-fold cross-validation is used throughout. The **#forms**-column shows the number of different slots in the paradigms. Other columns are as in table 2.

In some cases, we see a significant drop between the per-form and the per-table accuracy. For example, in the case of Russian nouns, per table accuracy is at 66.36%, while the per-form accuracy is 89.35%. This effect is explained—not only in the Russian case but in many others—by the existence of similar paradigms that differ only in very few forms. If the classifier picks an incorrect, but closely related paradigm, most forms may be produced correctly although the entire reconstructed table counts as wrong if even a single form is incorrect.

A few outliers remain. The Maltese verbs, which exhibit Semitic interdigitation in some paradigms, seem to generalize fairly well, and have a per form oracle score of 95.49 (shown in table 3). However, this is not reflected in the relatively low per form accuracy (61.15), which warrants further analysis. It may be an indication of that the correct paradigm is simply difficult to ascertain based only on the lemma form, or that additional features could be developed, perhaps ones that are discontinuous in the word.

An obvious extension to the current method is to inspect a suggested reconstructed table holistically, i.e., not relying only on base form features. That is, one could avoid making a commitment to a particular paradigm based solely on the features of the base form, and instead also include features from all the forms that a paradigm would generate. Such features are of course available in the training data in the various forms in an inflection table. Features from the seen forms could be used to rate compatibility since an incorrect reconstruction of an inflection table may likely be identified by its tendency to produce phonotactic patterns rarely or never seen in the training data.

With relatively few paradigms learned from collections of word forms, one can achieve fairly high coverage on unseen data. In principle, for example, the 13 most frequently used paradigms of Spanish verbs suffice to cover 90% of all verbs (per token). A useful application of this is rapid language resource development—one can elicit from a speaker a small number of well-chosen inflection tables, e.g. all forms of specific nouns, verbs, adjectives; generalize these inflection tables into paradigms; and use this information to deduce the possible morphological classes for a majority of unseen word forms.

References

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, Gothenburg, Sweden. Association for Computational Linguistics.

John J. Camilleri. 2013. A computational grammar and lexicon for Maltese. Master's thesis, Chalmers University of Technology. Gothenburg, Sweden.

Liviu P Dinu, Vlad Niculae, and Octavia-Maria Şulea. 2012. Learning how to conjugate the Romanian verb: rules for regular and partially irregular verbs. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 524–528. Association for Computational Linguistics.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 616–627. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of NAACL-HLT*, pages 1185–1195.

Ramy Eskander, Nizar Habash, and Owen Rambow. 2013. Automatic extraction of morphological lexicons from morphologically annotated corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1032–1043. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. A practical guide to support vector classification.

Mans Hulden. 2014. Generalizing inflection tables into paradigms with finite state operations. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 29–36. Association for Computational Linguistics.

Jackson L Lee and John A Goldsmith. 2014. Complexity across morphological paradigms: a minimum description length approach to identifying inflectional stems. In *Poster at the MorphologyFest: Symposium on Morphological Complexity, Indiana University, Bloomington*.

Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 67–74.

Kristina Toutanova and Colin Cherry. 2009. A global model for joint lemmatization and part-of-speech prediction. 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: Volume 1-Volume 1*, pages 486–494. Association for Computational Linguistics.

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In *Proceedings of the fourth SIGHAN workshop on Chinese language processing*, pages 32–39.