Evaluation of Finite State Morphological Analyzers Based on Paradigm Extraction from Wiktionary

Ling Liu and **Mans Hulden**
Department of Linguistics
University of Colorado
first.last@colorado.edu

Abstract

Wiktionary provides lexical information for an increasing number of languages, including morphological inflection tables. It is a good resource for automatically learning rule-based analysis of the inflectional morphology of a language. This paper performs an extensive evaluation of a method to extract generalized paradigms from morphological inflection tables, which can be converted to weighted and unweighted finite transducers for morphological parsing and generation. The inflection tables of 55 languages from the English edition of Wiktionary are converted to such general paradigms, and the performance of the probabilistic parsers based on these paradigms are tested.

1 Introduction

Morphological inflection is used in many languages to convey syntactic and semantic information. It is a systematic source of sparsity for NLP tasks, especially for languages with rich morphological systems where one lexeme can be inflected into as many as over a million distinct word forms (Kibrik, 1998). In this case, morphological parsers which can convert the inflected word forms back to the lemma forms, or the other way around can largely benefit downstream tasks, like part-of-speech tagging, language modeling, and machine translation (Tseng et al., 2005; Hulden and Francom, 2012; Duh and Kirchhoff, 2004; Avramidis and Koehn, 2008). Various approaches have been adopted to tackle the morphological inflection and lemmatization problem. For example, Durrett and DeNero (2013) automatically extracts transformation rules from labeled data and learns how to apply these rules with a discriminative sequence model. Kann and Schütze (2016) proposes to use a recurrent neural network (RNN) encoder-decoder model to generate an inflected form of a lemma for a target morphological tag combination. The SIGMORPHON 2016 shared task (Cotterell et al., 2016) of morphological reinflection received 11 systems which used various approaches such as conditional random fields (CRF), RNNs, and other linguistics-inspired heuristics. Among all the methods, one standard technology is to use finite-state transducers, which are more interpretable and manually modifiable, and thus more easily incorporated into and made to assist linguists' work. Hulden (2014) presents a method to generalize inflection tables into paradigms with finite state implementations and Forsberg and Hulden (2016) subsequently introduce how to transform morphological inflection tables into both unweighted and weighted finite transducers and apply the transducers to parsing and generation, the result of which is very promising, especially for facilitating and assisting linguists' work in addition to applications to morphological parsing and generation for downstream NLP tasks. However, the system was evaluated with only three languages (German, Spanish, and Finnish), all with Latin script. This paper intends to carry out a more extensive evaluation of this method.

Wiktionary[1] provides a source of morphological paradigms for a wide and still increasing range of languages, which is a useful resource of crosslinguistic research. The data in this work also originate with Wiktionary.

In this paper we evaluate the cross-linguistic performance of the paradigm generalization method on 55 languages, of which inflection tables have been extracted from the Wiktionary data. All the languages are consistently annotated with universal morphological tags (Sylak-Glassman et al., 2015) and are in the native orthography. In particular, we evaluate the accuracy on the ability to lemmatize previously unseen word forms and the

[1]http://www.wiktionary.org

Proceedings of the 13th International Conference on Finite State Methods and Natural Language Processing, pages 69–74,
Umeå, Sweden, 4–6 September 2017.
https://doi.org/10.18653/v1/W17-4009

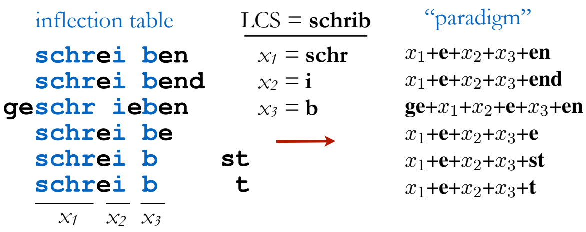

Figure 1: Illustration of the paradigm extraction mechanism. An inflection table is given as input, and the longest common subsequence (LCS) is extracted and assigned to “variable parts” of a more abstract paradigm, based on discontinuities in the LCS. Several inflection tables may yield the same “paradigm” in which case paradigms are collapsed, and information about the shape of the variable strings x_i is retained for statistical modeling.

ability to assign correct morphosyntactic tags to a word form.

2 Paradigm Extraction

The paradigm extraction method is based on the idea of finding, among a list of related word forms, the longest common subsequence (LCS) shared by the forms. After the extraction, the LCS is marked in each word form and assigned to “paradigm variables”. These variables are parts that are mutable in a paradigm, i.e. may change when going from one lemma to another, while the remaining, non-variable parts represent inflectional information. Figure 1 illustrates this process by showing a few forms of the German verb **schreiben**, the extraction of the LCS, and the assignment of the LCS into variable parts.

After such a generalization process, many paradigm representations which were generated from inflection tables turn out to be identical—indicating that the participating lemmas inflect according to the same pattern. Identical paradigms are collapsed, and the information about what strings were witnessed in the variable slots is stored for creating a probabilistic model of inflection. The reader is referred to (Hulden, 2014; Ahlberg et al., 2014; Ahlberg et al., 2015) for details.

This model already provides a method for performing morphological analysis when previously unseen word forms are encountered. One can create a transducer based on the paradigms that maps entries in a paradigm back to their lemma form in such a way that the variable parts x_i may correspond to any arbitrary string. For example, the paradigm in figure 1 would yield a lemmatizing transducer that would map e.g. **geliehen** $\mapsto$ **leihen** since the **l** can be assumed to match the variable x_1, the **i** x_2, and the **h** x_3. Forsberg and Hulden (2016) developl a model that creates such lemmatizing transducers from inflection tables, which also return the inflectional information of the source word form.

2.1 Analyzing word forms

This model has the disadvantage of often returning a large number of plausible analyses due to the fact that an unseen word form may fit many different learned paradigms, and also fit them in many different slots. One can, however, induce a language model of each variable part x_i in the paradigms and create a probabilistic model which favors production of such analyses where variable parts resemble those that have been seen in the training data. An n-gram model over the variables seen in each paradigm can be formulated as follows. Many paradigms have been collapsed from a large number of inflection tables which provide us with statistics over the shape of the x_i parts. We can, when trying to fit an unseen word form with a variable x_i consisting of letters $v_1, \ldots v_n$ into a paradigm and slot to produce its lemma, calculate the joint probability using an n-gram approximation of the letters according to the expression:

$$P(v_1, \ldots, v_n) = \prod_{i=1}^{n} P(v_i | v_{i-(n-1)}, \ldots, v_{i-1}) \tag{1}$$

These quantities can be estimated by maximum likelihood from the the training data as:

$$P(v_i | v_{i-(n-1)}, \ldots, v_{i-1}) = \frac{\#(v_{i-(n-1)}, \ldots, v_{i-1}, v_i)}{\#(v_{i-(n-1)}, \ldots, v_{i-1})} \tag{2}$$

Such a model is induced for each variable $x_1, \ldots, x_n$ in a paradigm, and when a proposed analysis is evaluated, the quantity $p(x_i, \ldots, x_n) = p(x_1) \times \ldots \times p(x_n)$ is evaluated to give a score of fit of a proposed variable assignment for a word to be analyzed.

For example, to calculate the fit of **geliehen** into the slot **ge**+x_1+x_2+**e**+x_3+**en** (Figure 1), we would evaluate $p(x_1)$ = **l**, $p(x_2)$ = **i** and $p(x_3)$ = **h** based on the above, yielding a probability estimate of the slot and paradigm matching the word form **geliehen**. Likewise, every possible assignment of variable parts in every paradigm will be calculated. This process can be encoded into a weighted finite state transducer (WFST) following Forsberg and Hulden (2016).

Language	Language group	Script
Adyghe	Northwest Caucasian	Cyrillic
Albanian	IE other	Latin (Albanian alphabet)
Armenian	IE other	Armenian
Asturian	IE/Italic/Romance/Western	Latin
Bashkir	Turkic	Cyrillic, Latin, Arabic
Basque	Language Isolate	Latin (Basque alphabet)
Bengali	IE/Indo-Iranian/Indo-Aryan	Eastern Nagari script (Bengali alphabet)
Bulgarian	IE/Balto-Slavic/Slavic	Cyrillic (Bulgarian alphabet)
Catalan	IE/Italic/Romance/Western	Latin (Catalan alphabet)
Danish	IE/Germanic/North G	Latin (Dano-Norwegian alphabet)
Dutch	IE/Germanic/West G	Latin (Dutch alphabet)
Esperanto	Created	Latin (Esperanto alphabet)
Estonian	Uralic/Finnic	Latin (Estonian alphabet)
Faroese	IE/Germanic/North G	Latin (Faroese orthography)
Finnish	Uralic/Finnic	Latin (Finnish alphabet)
French	IE/Italic/Romance/Western	Latin (French alphabet)
Friulian	IE/Italic/Romance/Western	Latin
Galician	IE/Italic/Romance/Western	Latin (Galician alphabet)
Georgian	Kartvelian	Georgian
German	IE/Germanic/West G	Latin (German alphabet)
Greek	IE/Hellenic	Greek
Hebrew	Afro-Asiatic/Semitic/Central S	Hebrew
Hindi	IE/Indo-Iranian/Indo-Aryan	Devanagari
Hungarian	Uralic/Finno-Ugric	Latin (Hungarian alphabet)
Icelandic	IE/Germanic/North G	Latin (Icelandic alphabet)
Italian	IE/Italic/Romance/Italo-Dalmatian	Latin (Italian alphabet)
Ladin	IE/Italic/Romance/Western	Latin
Latin	IE/Italic/Latino-Faliscan	Latin
Latvian	IE/Balto-Slavic/Baltic	Latin (Latvian alphabet)
Lithuanian	IE/Balto-Slavic/Baltic	Latin (Lithuanian alphabet)
Lower Sorbian	IE/Balto-Slavic/Slavic/West S	Latin (Sorbian alphabet)
Luxembourgish	IE/Germanic/West G	Latin (Luxembourgish alphabet)
Macedonian	IE/Balto-Slavic/Slavic/South S	Cyrillic (Macedonian alphabet)
Navajo	Da-Dene/Athabaskan	Latin
Northern Sami	Uralic/Sami	Latin (Northern Sami alphabet)
Norwegian Bokmal	IE/Germanic/North G	Latin (Norwegian alphabet)
Norwegian Nynorsk	IE/Germanic/North G	Latin (Norwegian alphabet)
Occitan	IE/Italic/Romance/Gallo-R	Latin
Polish	IE/Balto-Slavic/Slavic/West S	Latin (Polish alphabet)
Portuguese	IE/Italic/Romance/Western R	Latin (Portuguese alphabet)
Quechua	Quechua	Latin (Quechua alphabet)
Romanian	IE/Italic/Romance/Eastern R	Latin (Romanian alphabet)
Russian	IE/Balto-Slavic/Slavic/East S	Cyrillic (Russian alphabet)
Sanskrit	IE/Indo-Iranian/Indo-Aryan	Brahmic
Scottish Gaelic	IE/Celtic	Latin (Scottish Gaelic orthography)
Slovak	IE/Balto-Slavic/Slavic/West S	Latin (Slovak alphabet)
Slovene	IE/Balto-Slavic/Slavic/South S	Latin (Slovene alphabet)
Spanish	IE/Italic/Romance/Western	Latin (Spanish alphabet)
Swahili	Niger-Congo	Latin (Roman Swahili alphabet)
Swedish	IE/Germanic/North G	Latin (Swedish alphabet)
Turkish	Turkic	Latin (Turkish alphabet)
Ukrainian	IE/Balto-Slavic/Slavic/East	Cyrillic (Ukrainian alphabet)
Urdu	IE/Indo-Iranian/Indo-Aryan	Extended Perso-Arabic (Urdu alphabet)
Venetian	IE/Italic/Romance/Italo-Western	Latin
Welsh	IE/Celtic	Latin (Welsh alphabet)

Table 1: Languages, Language Groups, and Scripts

2.2 Evaluation

The paradigm extraction and application method presented in the previous part is evaluated with 55 languages from the Wiktionary Morphological Database[2] (Kirov et al., 2016) as part of the UniMorph project[3] which includes data for 350 languages at the time we downloaded it.[4] For each of the 55 languages, we learn paradigms from a random selection of 90% of the available inflection tables and leave 10% of the tables as held-out data. Tables for different parts-of-speech are generalized identically and the system does not keep these separate (although it is unlikely that a noun would inflect like a verb, for example). This means that the POS information is treated as a normal tag and we may receive analyses with different parts of speech.

The evaluation task is to convert the inflected word form back to its lemma form and assign it morphological tags. At test time, we analyze each form in the held-out tables separately and evaluate the accuracy on the highest scoring analysis. We report accuracies on several combinations of lemmatization correctness, morphosyntactic tag correctness and POS tag correctness.

The 55 languages fall into 19 language groups: Caucasian, Indo-European other, Italic (Romance, with the exception of Latin), Turkic, Language isolated, Indo-Aryan, Slavic, Baltic, Germanic, Uralic, Celtic, Semitic, Hellenic, Kartvelian, Da-Dene, Quechua, Turkic, Niger-Congo, and an artificial language—Esperanto. The data for each language is in its native script, which consists of 10 different scripts: Latin, Cyrillic, Armenian, Eastern Nagari, Georgian, Greek, Hebrew, Devanagari, Brahmic, and Pero-Arabic. The criterion for selecting the 55 languages is that each language has 8,000 or more entries in the original data in UniMorph Wiktionary Morphological Database. Languages with less entries are selected to increase the representativeness of the data. A summary of the languages, language groups and scripts is presented in table 1. The data for each language is used just at it is from the database. Little work is done to improve the quality of the data. Therefore, if there are misspellings or incorrect inflections in the data set, the paradigms are extracted and tested with any errors uncorrected.

The number of inflection tables may not be the same as the number of lemmas, because in cases where there are alternative inflected forms for one morphosyntactic description of a lemma in the UniMorph database, each form is represented in a separate table.

[2]https://github.com/ckirov/UniMorph/tree/master/data

[3]http://www.unimorph.org

[4]The 55 languages are: Adyghe, Albanian, Armenian, Asturian, Bashkir, Basque, Bengali, Bulgarian, Catalan, Danish, Dutch, Esperanto, Estonian, Faroese, Finnish, French, Friulian, Galician, Georgian, German, Greek, Hebrew, Hindi, Hungarian, Icelandic, Italian, Ladin, Latin, Latvian, Lithuanian, Lower Sorbian, Luxembourgish, Macedonian, Navajo, Northern Sami, Norwegian, Bokmal, Norwegian Nynorsk, Occitan, Polish, Portuguese, Quechua, Romanian, Russian, Sanskrit, Scottish Gaelic, Slovak, Slovene, Spanish, Swahili, Swedish, Turkish, Ukrainian, Urdu, Venetian, Welsh.

Language	LT	LPOS	LEMMA	TNum	PAllNum	P1Ex	P2Ex	PMEx	P0Var	TopFreq	WNum	LNum
Adyghe	0.6012	0.6012	0.7887	1,593	16	5	1	10	0	550	18,874	1,593
Albanian	0.7016	0.7100	0.7179	616	68	36	6	21	5	121	37,411	589
Armenian	0.8994	0.9010	0.9631	14,905	305	104	32	158	11	1,730	703,902	7,040
Asturian	0.8840	0.8933	0.9140	1,304	113	58	14	39	2	235	41,835	938
Bashkir	0.8816	0.9245	0.9245	773	36	6	2	28	0	92	8,981	773
Basque	0.0	0.0	0.0021	45	44	38	0	0	6	1	13,627	45
Basque2	0.1276	0.3111	0.3588	620	504	9	0	2	493	52	27788	620
Bengali	0.7234	0.7255	0.7255	225	74	55	3	15	1	21	5,691	136
Bulgarian	0.7429	0.7445	0.7627	2,912	225	98	25	96	6	315	55,523	2,471
Catalan	0.8550	0.8894	0.8894	1,558	107	59	11	33	4	779	83,182	1,557
Danish	0.6826	0.6945	0.7075	3,180	248	180	11	57	0	658	28,584	3,180
Dutch	0.6993	0.7110	0.7323	4,985	274	112	34	128	0	959	60,437	4,979
Esperanto	0.7628	0.7638	0.7700	23,687	46	28	2	13	3	11,652	98,565	23,687
Estonian	0.5958	0.5970	0.5970	887	779	776	2	1	0	19	39,102	886
Faroese	0.5121	0.5297	0.5702	3,333	621	419	62	135	5	217	52,836	3,077
Finnish	0.6803	0.6823	0.7633	20,000	2,403	862	477	1,064	0	685	627,085	14,274
French	0.7412	0.7466	0.8556	19,937	1,248	604	201	428	15	2,241	387,669	7,555
Friulian	0.7844	0.7844	0.7929	155	37	22	2	8	5	73	6,593	145
Galician	0.7389	0.7485	0.7613	486	63	32	7	22	2	184	29,843	472
Georgian	0.6623	0.6630	0.7828	3,784	63	49	3	11	0	2,117	78,196	3,782
German	0.6221	0.6510	0.8216	17,749	851	512	100	239	0	1653	197,080	15,059
Greek	0.6555	0.6619	0.6938	9,861	2,072	1,426	212	377	57	468	149,187	8,780
Hebrew	0.8217	0.8263	0.8288	867	555	437	58	55	4	9	24,247	510
Hindi	0.9692	0.9696	0.9746	852	33	10	6	16	1	184	72,424	788
Hindi2	0.9522	0,9553	0.9597	788	31	11	4	15	1	181	58,856	788
Hungarian	0.8550	0.9197	0.9277	15,838	4,136	3,640	37	173	286	1,009	597,744	13,952
Icelandic	0.6917	0.7158	0.7233	5,082	384	155	46	170	14	377	86,864	4,769
Italian	0.8972	0.8973	0.8973	10,009	374	238	32	98	6	3919	519,571	10,009
Ladin	0.8400	0.8595	0.8875	813	111	72	11	20	8	97	17,830	512
Latin	0.8155	0.8167	0.8406	25,079	2,463	1,602	275	567	19	3,643	967,062	20,497
Latvian	0.7943	0.8269	0.8450	10,067	519	275	66	167	11	1,722	216,420	7,558
Lithuanian	0.7786	0.7820	0.7969	1,925	515	298	79	133	5	130	58,814	1,458
Lower Sorbian	0.6098	0.6265	0.6273	1,224	351	260	18	70	3	57	25,062	994
Luxembourgish	0.7422	0.7554	0.7591	1,277	222	172	12	34	4	318	48,175	1,276
Macedonian	0.7319	0.7563	0.7837	10,313	272	134	16	114	8	869	178,363	10,310
Navajo	0.2649	0.2746	0.2912	675	455	398	22	17	18	61	13,059	674
N Sami	0.4377	0.4414	0.4476	2,203	1,343	1,250	41	52	0	44	66,344	2,107
N Bokmal	0.5866	0.6170	0.6733	5,750	355	211	42	99	3	1,097	24,704	5,518
N Nynorsk	0.5959	0.6242	0.6720	5,041	380	237	44	94	5	1,333	23,093	4,677
Occitan	0.9474	0.9474	0.9555	173	28	19	3	6	0	101	5,162	172
Polish	0.6967	0.7059	0.7138	10,698	1,300	737	161	398	4	566	202,927	10,179
Portuguese	0.9053	0.9062	0.9369	5,088	283	162	11	107	3	2,400	309,084	4,001
Quechua	0.4513	0.4515	0.9128	1,244	17	8	0	9	0	372	181,248	1,006
Romanian	0.6258	0.6380	0.6989	3,504	671	447	76	144	4	205	68,174	3,479
Russian	0.7100	0.7615	0.7749	28,017	1,669	825	231	612	1	1,898	460,981	27,924
Sanskrit	0.6847	0.7029	0.7998	924	79	43	7	26	3	243	27,031	924
S Gaelic	0.9231	0.9692	0.9692	101	74	61	9	3	1	4	1,208	70
Slovak	0.5582	0.5899	0.5912	1,093	250	137	42	68	3	116	16,435	1,046
Slovene	0.4383	0.4646	0.4886	2,601	949	747	62	115	25	150	64,504	2,554
Spanish	0.9037	0.9038	0.9355	6,355	420	194	52	169	5	1,977	389,308	5,460
Swahili	0.8550	0.8559	0.8567	74	42	40	0	2	0	20	12,365	74
Swedish	0.6928	0.7156	0.7615	10,833	458	243	57	155	4	1,188	92,890	10,503
Turkish	0.8286	0.8414	0.8480	3,572	237	97	29	111	0	155	278,624	3,572
Ukrainian	0.5645	0.5900	0.5948	1,494	248	155	30	62	3	142	22,398	1,493
Urdu	0.8784	0.8827	0.8827	182	51	43	1	7	0	40	12,755	182
Venetian	0.8160	0.8342	0.9327	959	112	73	10	22	7	193	29,790	607
Welsh	0.2378	0.2478	0.2478	183	158	155	1	2	1	4	10824	183

Table 2: Result Summary. LT–Lemma+Tags, LPOS–Lemma+POS, LEMMA–Lemma, TNum–The number of inflection tables in the data, PAllNum–The number of all paradigms for each language, P1Ex–The number of paradigms with only 1 example for the variable(s), P2Ex–The number of paradigms with only 2 examples for the variable(s), PMEx–The number of paradigms with more than 2 examples for the variable(s), P0Var–The number of paradigms with no variable in it, TopFreq – The number of examples for the most popular paradigm, WNum–The number of words, LNum–The number of lemmas, N Sami–Northern Sami, N Bokmal–Norwegian Bokmal, N Nynorsk–Norwegian Nynorsk, S Gaelic–Scottish Gaelic, Basque2–result of Basque with more data, Hindi2–result of Hindi without alternative inflections

3 Result and Discussion

More abstract paradigms are extracted successfully from all the morphological tables in the training set for each of the 55 languages. Lemmatization correctness ranges from a low end of 0% (Basque) to a high end of 96.3% (Armenian), 97.5% (Hindi). The lemma-POS accuracy ranges from 0% (Basque) to 97.0% (Hindi). The joint lemma-tag accuracy ranges from 0% (Basque) to 96.9% (Hindi).

One advantage of the probabilistic model is that it can rank the parsing and generation through the language model over variables. For the evaluation, we use only the most likely one. However, there can still be alternatives for the top ranking result. For example, the French word **écris** gets five analyses: **[écrir V; IND; PRS; 1; SG]**, **[écrir V; IND; PRS; 2; SG]**, **[écrir V; IND; PST; 1; SG; PFV]**, **[écrir V; IND; PST; 2; SG; PFV]** and **[écrir V; POS; IMP; 2; SG]**, with the same lemma form and part-of-speech, but different tags. We evaluate the recall of the lemma and all tags, lemma and only part-of-speech, and only lemma. Table 2 presents a summary of the evaluation result, as well as the data size for each language in terms of word counts, lemma counts and table counts, the number of abstract paradigms extracted for each language, the paradigm distribution as to their instantiation case numbers, and the number of instances for the most popular paradigm.

The results seem to correlate strongly with the amount of available data. For languages where only a few inflection tables are seen, or where all inflection tables represent the same paradigm, accuracies are low. For example, the initial evaluation on Basque is only 45 lemmas (and the number of inflection tables is the same), with 44 as training and 1 held-out for test. Each of the inflection table represents a distinct paradigm as is reflected by the fact that the number of abstract paradigms is the same as the training data size, i.e. 44. Therefore, the result of Basque is very low. A second round of evaluation was conducted with more data and the result is added to Table 2 as Basque2. The increased Basque data is 620 lemmas, with 558 used for training and 62 for testing. The result is still low but better than the initial one, because the paradigms become more representative with a larger coverage, which is testified by the most representative paradigm getting 52 instantiations from the training data. The result of Navajo is close to that of Basque2. The data sizes of the two languages are similar (Basque2 620, Navajo 675). For Navajo, the number of paradigms is 455, and for Basque2, the number of paradigms is 504. Basque2 has 493 paradigms without variables (i.e. 493 paradigms with only one instantiations), and Navajo has 398 paradigms with only one instantiation in the training data. The similarity of the results coincides with the fact that both languages are morphologically complex, and Basque morphology has even more variation. Conversely, languages where many different types of inflection tables are seen and the inflection tables are representative of the language morphology (reflected in a higher paradigm count and a lower ratio of paradigm counts to table counts), produce analyzers that can perform quite robustly. For example, the Hindi data produces 33 paradigms out of the 767 inflection training tables, resulting in a coverage where the recall for each of the three tests is over 95%. As alternative inflections of a lemma are represented with different tables, the recall may be higher than the case where each lemma gets only one inflection table, i.e. where no related form of the associated lemma has been witnessed. However, as alternative inflections are limited, keeping them with different tables should not influence the result by a large extent. Hindi2 is the result for Hindi using only one inflection table for each lemma and ignoring alternative inflections.

4 Conclusion

Generalized paradigms are successfully extracted for all the 55 languages from 19 language groups in 10 different scripts. For languages with a large size of representative data, the recalls of the lemma, lemma plus part-of-speech, and lemma plus all tags can be as high as over 95%. However, for languages for which the data is limited or less representative, the recalls are very low. This indicates that the method to extract generalized paradigms from morphological inflection tables works well despite linguistic diversity and script variations. The probabilistic model can yield good predictions and analyses when the available data for a language is sufficient and representative.

Acknowledgements

This work has been partly sponsored by DARPA I20 in the LORELEI program.

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.

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2015. Paradigm classification in supervised learning of morphology. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1024–1029, Denver, Colorado, May–June. Association for Computational Linguistics.

Eleftherios Avramidis and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *ACL*, pages 763–770.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. *ACL 2016*, page 10.

Kevin Duh and Katrin Kirchhoff. 2004. Automatic learning of language model structure. In *Proceedings of the 20th international conference on Computational Linguistics*, page 148. Association for Computational Linguistics.

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

Markus Forsberg and Mans Hulden. 2016. Learning transducer models for morphological analysis from example inflections. *ACL 2016*, page 42.

Mans Hulden and Jerid Francom. 2012. Boosting statistical tagger accuracy with simple rule-based grammars. In *LREC*, pages 2114–2117.

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.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. *arXiv preprint arXiv:1606.00589*.

Aleksandr E. Kibrik. 1998. Archi. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*, pages 455–476. Oxford: Blackwell Publishers.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In Cerstin Mahlow and Michael Piotrowski, editors, *Proceedings of the 4th Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, pages 72–93. Springer, Berlin.

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.