Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages

Niraj Aswani and Robert Gaizauskas
Department of Computer Science, University of Sheffield

Most dictionaries have words in their root Forms. Use it to obtain most common suffixes of root forms

Use common suffixes on Monolingual corpus to find words that do not exist in the monolingual dictionary

Hindi Monolingual Dictionary

EMILLE Corpus

3

1

2

Common root form suffixes

Verbs	%*	Nouns	%*	Adjectives	%*	Adverbs	%*
ा(*aa*)	99.81	ा(*aa*)	19.99	त (*ta*)	17.06	र (*ra*)	14.29
ना (*naa*)	99.71	ी(*ee*)	16.62	ी(*ee*)	14.89	ा(*aa*)	13.21
ाना (*aanaa*)	44.21	र (*ra*)	9.15	ा(*aa*)	10.76	:	11.52
वाना (*vaanaa*)	12.19	न (*na*)	8.06	य (*ya*)	9.20	त :(*taha*)	10.09
कना (*kanaa*)	8.03	-	-	क (*ka*)	7.87	क (*ka*)	8.13
राना (*raanaa*)	7.98	-	-	ित (*ita*)	8.74	त (*ta*)	6.43
लना (*lanaa*)	5.03	-	-	न (*na*)	6.50	-	-
-	-	-	-	िर (*ira*)	6.90	-	-

∗ percentage of number of words that have the listed suffix in that category.

6

If the new word is found in the dictionary, create a rule update counter. If not try next suffix.

Remove one character at the end and attach a high-frequency suffix

5

⊆ ⊆

4

Take one inflected word at a time

D1 ∩ D2 = ∅

7

Suffix	Replacement	Count	Example	Category
े (*e*)	ा (*aa*)	1561	लडके (*ladake, boys*) - लडका (*ladakaa, a boy*)	noun
े (*e*)	ा (*aa*)	1561	रोने (*rone, for crying*) - रोना (*ronaa, to cry*)	verb
ो (*o*)	ी (*ee*)	492	नौकरों (*naukaron, servants*) - नौकरी (*nakaree, a job*)	noun
े (*e*)	ी(*ee*)	483	कटोरे (*katore, bowls*) - कटोरी (*katoee, a bow*)	noun
ों (*on*)	-	476	परिवारों (*parivaaron, families*) - परिवार (*parivaar, a family*)	noun
-	ना (*naa*)	474	फिसल (*phisal, to slip*) - फिसलना (*phisalanaa, to slip*)	verb
ों (*on*)	ा	466	नमूनों (*namoonon, samples*) - नमूना (*namoonaa, a sample*)	noun
ते (*te*)	ना (*naa*)	461	जाते (*jaate, while going*) - जाना (*jaanaa, to go*)	verb
यों (*iyon*)	ी(*ee*)	432	नदियों (*nadiyon, rivers*) - नदी (*nadi, a river*)	noun
ता (*taa*)	ना (*naa*)	371	देता (*dete, while giving*) - देना (*denaa, to give*)	verb

Suffix replacement rules

Order of the rules

1. Given two rules, the rule with the longer suffix is executed first. This is to make sure that more specific rules are given priority over generalized ones.
2. If they have suffixes of the same length the rule with the higher frequency is executed first.
3. If they have suffixes of the same length and the frequencies of the suffixes are same, frequencies of the rules are looked up and the rule with higher frequency is executed first.
4. If the frequencies of replacements are same, the one with the smallest replacement string is executed first. This makes sure that the minimum change is applied to the word.
5. For the inflected words whose base-forms do not exist in the dictionary use the output of the first rule that matches the inflected word.
6. There are lots of rules – how to decide which ones to keep? Go to step 9.

8

Still not found? continue step 5 until there is one letter in the word remaining and then continue step 1. If no more words go to step 8.

D2

Dataset 2 consists of 2500 inflected words with their manually obtained root forms

9

1. Add a suffix replacement rule and execute it over D2.
2. If the F-measure increases, add it to the final rule set.
3. Otherwise, check the examples collected for this rule and decide.

10

Final Rule set

11

Dataset 1 consists of 2500 inflected words with their manually obtained root forms

D1

12

RFList	(Shrivastava et al., 2005)						Extended Ruleset					
	Non-derivational			Derivational			Non-derivational			Derivational		
	P	R	F	P	R	F	P	R	F	P	R	F
Wordnet	0.736	0.683	0.708	0.735	0.683	0.708	0.743	0.717	0.730	0.743	0.717	0.730
Extended	0.817	0.768	0.792	0.817	0.768	0.792	**0.821**	**0.803**	**0.812**	0.820	0.803	0.812

Results of experiments on the Hindi language

- To our knowledge, Shrivastava et al. (2005) is the best morpher available for the Hindi language.
- We extend their ruleset by adding missing rules using our approach.

Morpher based on Prefixes

Repeat steps 1 and 4 to 7 for collecting prefixes. Step 5 is continued until the remaining string was found as another individual word in the dictionary.

Prefix replacement rules

Prefix	Count	Example	Decomposition
अ (*a*)	1616	अपरिचित (*aparichit, unknown*)	अ (*a*) + परिचित (*parichit, known*)
वि (*vi*)	307	विदेशी (*videshee, foreign*)	वि (*vi*) + देशी (*deshee, local*)
अन् (*an*)	239	अनावश्यक (*anaavashyak, unnecessary*)	अन् (*an*) + आवश्यक (*aavashyak, necessary*)
सु	197	सुपात्र (good character)	सु + पात्र (character)

Experiments with the Gujarati Language

Gujarati Dictionary

3

EMILLE

1

2

Verbs	%	Nouns	%	Adjectives	%	Adverbs	%
વું (*vun*)	92.57	ી(*ee*)	12.27	ં (*n*)	22.36	ં (*n*)	22.82
વવું (*vavun*)	17.91	ા(*aa*)	11.27	ું (*un*)	22.95	ું (*un*)	17.41
ાવવું (*aavavun*)	12.95	ં (*n*)	8.86	ક (*ka*)	14.22	ા(*aa*)	12.00
ાવું (*aavun*)	9.90	ર (*ra*)	8.62	ી(*ee*)	9.50	ે(*e*)	11.52
ડવું (*davun*)	9.68	ઓ (*o*)	8.05	િત (*ita*)	7.48	ર (*ra*)	7.64
રવું (*ravun*)	9.23	ું (*un*)	7.99	િક (*ika*)	7.46	ક (*ka*)	7.29
ળવું (*lavun*)	6.41	તા (*taa*)	6.04	ર (*ra*)	7.20	ત (*ta*)	7.29
કવું (*kavun*)	5.29	નું (*nun*)	0.14	ળુ (*lu*)	6.91	થી (*thee*)	7.05
-	-	ાલ (*aal*)	0.14	વાળુ (*vaalu*)	6.09	ાં (*aan*)	6.17
-	-	વાદ (*vaad*)	0.13	ાળુ (*alu*)	6.09	-	-

Common root-form suffixes

6

4

⊆ ⊆

D2

5

9

Finalizing ruleset

Suffix	Replacement	Count	Example	Category
ી(*ee*)	વું (*vun*)	392	પીગળી (*peegalee, melting*) - પીગળવું (*pigalavu, to melt*)	verb
ના (*naa*)	-	358	સલાહકારના (*salaahakaaranaa, advisor's*) - સલાહકાર (*salaahakaar, advisor*)	noun
ની (*nee*)	-	347	સલાહકારની (*salaahakaaranee, advisor's*) - સલાહકાર (*salaahakaar, advisor*)	noun
ને (*ne*)	-	318	સલાહકારને (*salaahakaarane, to advisor*) - સલાહકાર (*salaahakaar, advisor*)	noun
ા(*aa*)	ું (*un*)	278	બતાવવા (*bataavavaa, for showing*) - બતાવવું (*bataavavun, to show*)	verb
ા(*aa*)	ુ (*un*)	275	જોડાયેલા (*jodaayelaa, connected(p)*) - જોડાયેલું (*jodaayelu, connected(s)*)	adj
માં (*maan*)	-	273	વિચારમાં (*vichaarmaan, in thought*) - વિચાર (*vichaar, thought*)	noun
ે(*e*)	વું (*vun*)	269	જાળવે (*jaalave, preserve*) - જાળવવું (*jaalavavun, preserve*)	verb
તા (*taa*)	વું (*vun*)	258	બતાવતા (*bataavataa, showing*) - બતાવવું (*bataavavun, to show*)	verb
નો (*no*)	-	258	કટોકટીનો (*katokateeno, of urgency*) - કટોકટી (*katokatee, urgent*)	noun

Suffix replacement rules

D1 ∩ D2 = ∅

10

Final Rule set

11

D1

12

P=0.83, R=0.70, F=0.76

Acknowledgement: This work is partially supported by the EU-funded MUSING project (IST-2004-027097).