A Constraint-based Morphological Analyzer for Concatenative and Non-concatenative Morphology

Farrah Cherry Fortes-Galvan[1] and Rachel Edita Roxas[2]

[1] BitMicro Networks International, Inc., Net Square Center, Bonifacio Global City, Taguig City, Philippines
[2] Software Technology Department, 2401 Taft Avenue, Malate, Manila, Dela Salle University-Manila, Philippines 1000
{cherry_fortes@yahoo.com, roxasr@dlsu.edu.ph}

Abstract. Morphological analysis in the current methods, such as finite-state and unification-based, are predominantly effective for handling concatenative morphology (e.g. prefixation and suffixation), although some of these techniques can also handle limited non-concatenative phenomena (e.g. infixation and partial and full-stem reduplication). A constraint-based method to perform morphological analysis that handles both concatenative and non-concatenative morphological phenomena is presented, based on the optimality theory framework and the two-level morphology rule representation. Although optimality theory has been proven effective in handling non-concatenative phenomena, it has been applied for the generation process, and in this study, it has been shown to be effective also in morphological analysis. The method was tested on 1,600 Tagalog verb forms (having 3 to 7 syllables) from 50 Tagalog roots which exhibit both concatenative and non-concatenative morphology. The resulting method is able to produce the correct underlying forms of the surface forms of 96% of the test data, having a 4% error, which is attributed to d-r alternation.

Keywords: morphological analysis, morphological generation, optimality theory.

1 Introduction

Morphology is the area of linguistics that deals with the internal structure of words – how they are systematically formed from smaller units. Computational morphology deals with the processing of words and word forms in both written (graphemic) and spoken (phonemic) form. A morpheme is the smallest meaningful unit of a language. It could be a word stem or affixes. A free morpheme can stand alone, while bound morpheme functions only as part of a word. Affixation is the process of adding a bound morpheme to a free morpheme attached in front (prefix) or at the end (suffix) of a stem. The combination of a prefix and a suffix is called a circumfix. The infix is an affix where the placement is defined in terms of some phonological conditions. Reduplication requires copying of some portion of the stem. It can be complete or partial (prefixal, infixal, or suffixal reduplications). Infixation and reduplication are non-concatenative morphological processes.

Much of the work in Philippine linguistics focused on the Tagalog language [1]. Tagalog language exhibits a rather complex verbal system. A single verb may contain reduplicated syllables, prefixation or suffixation and at the same time infixation. Reduplication in Tagalog can be partial or full. There could also be a combination of such phenomena in one word especially seen in Tagalog verbs. For instance, the word *pinanglilibang-libang* has root word *libang* and prefix *pang* (with infix *in*), partial reduplication of syllable *li* and full word reduplication of *libang*.

Research on computational morphology has been predominantly on concatenative morphology and on finite-state models of morphotactics [2]. Although attempts were made to handle non-concatenative phenomena, it has been on a limited capacity only [3]. Optimality Theory is a phonological approach that is proven effective in handling non-concatenative morphology [4]. It has been applied for the generation process and never been used in morphological analysis.

2 Optimality Theory for Morphological Generation

The optimality theory architecture for morphological generation (as cited by [5], Fig. 1) has several components: the GEN and EVAL functions, the lexicon of root words, and the databases for the rules and constraints.

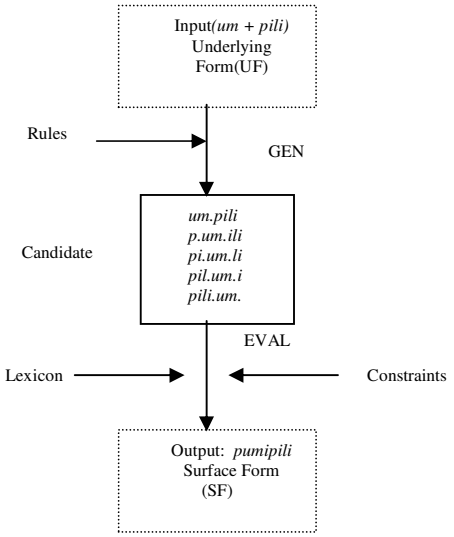

Fig. 1. The Generation Process in OT.

GEN *function* when applied to some input, produces a set of candidates, all of which are logically possible analyses of this input. EVAL *function* when applied to a set of candidates, produces an output, the optimal analysis of the input. LEXICON stores root words of words that can be inputted to GEN. Rules as applied to GEN are restrictions imposed, made of licit elements from universal vocabularies of linguistic representation, such as segmental structure, prosodic structure, morphology and syntax. Constraints applied to EVAL are structural requirements of a candidate that may be either satisfied or violated by an output form. Ranking of these constraints are imposed. While constraints are universal, rankings are not: differences in ranking are the source of cross-linguistic variation.

A theoretical study was conducted by Fosler in exploring the possibility of parsing through the reversal of the generation process in OT [5]. The architecture is shown in Fig. 2. It attempts to recover the underlying form from a surface form through reversing the generation process. Fosler made extensions to Ellison's approach of converting the constraints to finite-state transducers, showing the constraints to be regular.

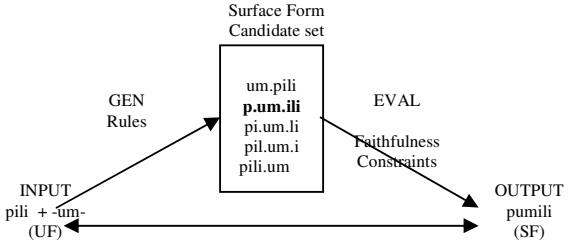

Fig. 2. Parsing in OT: Reversal of the Generation Process.

Fosler also employed the *um + gradwet* sample (which incidentally is in Tagalog) and illustrated in Table 1. NoCoda ranks higher than the Align constraint. NoCoda requires that syllables must be open, while Align requires that the prefix (-*um*-can appear as a prefix or move slightly into the word to which it is attached [6]) should remain as close to the front of the word as possible. The winning candidate, *gru.mad.wet,* violated the Align and NoCoda constraints twice. The final candidate violated NoCoda twice also, but it was pruned since it violated Align constraints more times than the winner. The winning candidate has the least number of violations.

Table 1. Fosler's representation on Tagalog infixation.

Candidates	NoCoda	Align
um.grad.wet	3	
gum.rad.wet	3	1
gru.mad.wet	2	2
gra.dum.wet	2	4

3 Optimality Theory for Morphological Analysis

Input in surface form is fed to the GEN function. The GEN function produces an underlying candidate set based on the rules and patterns of the verb. This candidate set will be validated by the EVAL function and an optimal candidate will be determined, based on a set of constraints. The flowchart (in Fig. 3) shows the processes in parsing.

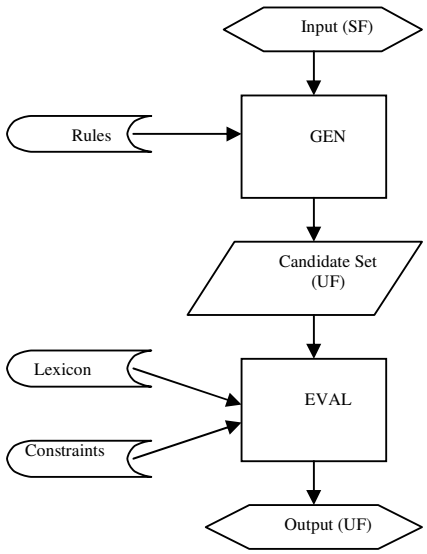

Fig. 3. TagMA: System Data Flow.

GEN function produces a candidate set. A candidate set is composed of underlying form candidates derived from the assumption that affixed and reduplicated verbs follow certain patterns. The GEN function produces any conceivable, possible underlying forms for a given input, under the morphological assumptions that are applicable. The candidate set shall be divided into separate paradigms. A paradigm is composed of candidates belonging to the same word structure classification. The following are the paradigms in this algorithm: Prefixed, Infixed, Partially Reduplicated, Full-Stem Reduplicated, and Suffixed.

There is a set of rules employed to produce each candidate for each paradigm. These rules account for the morphological analysis. The rules tokenize the input surface form into separate morphemes that represent the underlying form. Reduplication patterns according to French's work [6] was employed which includes also the infixation phenomena. Prefixation and suffixation configuration is based on [7], [8].

Evaluation through the application of constraints, is the determining engine of the optimal UF candidate. Evaluation of candidates is done by paradigm. Constraints shall only be applied to an appropriate candidate. The EVAL function will determine the winner (a candidate with no violations of any constraint) and extract its morphological categories from the lexicon and pass it to a syntactic parser. There are four types of ranked constraints that are used by EVAL:

Co-occurrence: certain root words could only be combined with certain set of affixes. Violation score: 4

Lexical Lookup: affixes and stems/root words should be in the dictionary. Violation score: 3

Affixal /Reduplication constraints: for each of the paradigm, certain characteristics must be possessed by the candidate. Violation: 2

Morphophonemic/Alternation Rules: these are the spelling-change rules that may change/delete/insert certain characters in order to make sure that the words are actually following the correct alternation rules, before going up the hierarchy. In this case, it can be said that the success of the rest of the constraints in the upper hierarchy depends on this constraint. Violation: 1

TagMA system is the application of the theories taken from OT. The system can handle both concatenative and non-concatenative morphologies. To measure the accuracy and efficiency of the system, 1,600 Tagalog verb forms (having 3 to 7 syllables) from 50 Tagalog roots [8] which exhibit both concatenative and non-concatenative morphology, particularly infixation and reduplication, and performance evaluation was measured in terms of actual runtime, number of rules, and number of constraints used based on the sample set of Tagalog verbs. The evaluation is further subdivided by paradigm (that is, prefix, infix, partial reduplication, full-stem reduplication, and suffix).

4 Results

The morphological analyzer accurately and efficiently outputs the underlying form for the specific test data of 1,600 Tagalog verbs. Overall, the resulting method is able to produce the correct underlying forms of the surface forms of 96% of the tokens in the test data, having a 4% error which is due to the d-r alternation rule.

To measure the efficacy of the system for performing morphological analysis on Tagalog verbs based on the number of syllables of the verbs that were taken from [8], the number of the rules and constraints those verbs undergo were also captured for the GEN and EVAL functions. In Table 2, the variance and standard deviation of the average number of candidates produced are presented.

Prefix and full-stem paradigms have 0 variance and standard deviation, which means that the number of candidates is predictable. In prefix, the number of syllables of a given input verb represents the number of conceivable candidates. Full-stem candidate depends upon the number of prefix candidates. The number of prefixed candidates is also the total candidates for full-stem. The rest of the paradigms have unpredictable number of candidates.

Table 2. Average number of candidates with variance and standard deviation.

SYL	PRE	INF	RED	FS	SFX
3					
AVE	3	3.08	1.64	3	4.99
VAR	0	0.74	1.86	0	4.12
STDD	0	0.86	1.36	0	2.03
4					
AVE	4	5.09	4.81	4	11.03
VAR	0	0.61	3.94	0	16.40
STDD	0	0.78	1.99	0	4.05
5					
AVE	5	7.28	8.89	5	21.12
VAR	0	0.86	3.65	0	50.05
STDD	0	0.91	1.91	0	7.07
6					
AVE	6	9.56	13.33	6	34.45
VAR	0	1.51	3.52	0	68.38
STDD	0	1.23	1.88	0	8.27
7					
AVE	7	11.34	17.05	7	46.30
VAR	0	1.33	2.01	0	55.58
STDD	0	1.15	1.42	0	7.46

Rules or patterns are applied by GEN to produce the candidate set. Any conceivable forms within the range of these rules are produced. As shown on Table 3, the number of rules applied in prefix is also predicable and stable. While rules being applied to the rest of the paradigms vary, but generally the relationship of number of syllables to the number of applied rules is approximately linear. This is due to the fact that the length of the candidate's affixes and stem are being considered before applying the rules.

Table 3. Average number of applied rules.

SYL	PRE	INF	RED	FS	SFX
3	2	8.86	9.23	4.88	3.22
4	2	12.38	18.21	7.45	7.67
5	2	15.39	28.97	11.64	14.57
6	2	18.50	40.32	17.20	23.32
7	2	21.74	49.34	22.43	30.87

EVAL function makes use of constraints to evaluate the candidate set. These constraints are only applied if certain conditions are met by the candidates. Table 4 shows that the relationship of the number of syllables and the number of applied constraints are linear.

Table 4. Average number of applied constraints.

SYL	PRE	INF	RED	FS	SFX
3	7.92	12.12	8.97	11.03	28.36
4	11.06	20.50	26.47	15.01	67.53
5	14.31	30.04	43.30	19.03	135.14
6	17.35	41.44	47.35	23.03	210.32
7	20.13	49.10	54.80	27.13	272.70

Overall, the resulting method is able to produce the correct underlying forms of the surface forms of 96% of the tokens in the test data, having a 4% error which is due to the d-r alternation rule. An example of an d-r alternation is in the word *lakaran*, which is from the root word *lakad* and suffix *an*, but *d* is changed to *r*.

In contrast, Cheng [9] applied an extension of Wicentowski's wordframe model [10] that captures both infixation and reduplication, showed a 90% accuracy of the morphological analysis of a corpus of 4,034 words for testing and 36,242 words for learning. Integration of other methods [11, 12] could also be considered to improve the analyzer's performance.

Although the system is intended for Tagalog verbs but can also be applied to some full-stem or complete reduplication phenomena in Tagalog non-verbs or other languages.

5 Conclusions

The approach has shown to effectively capture both concatenative (e.g. prefixation and suffixation) and non-concatenative morphological phenomena (e.g. infixation, partial and full-stem reduplication) with 96% accuracy.

It is recommended that the execution time of the method could be improved since the approach generates all possible candidate surface forms and eliminates those which violate any of the rules or constraints.

Acknowledgments. This project is funded by the Philippine Council for Advanced Science and Technology for Research and Development, Department of Science and Technology, Philippine Government.

References

1. De Guzman, V.: Syntactic Derivation of Tagalog Verbs. Honolulu. The University Press of Hawaii (1978)
2. Koskenniemi, K.: Two-level Morphology: A General Computational Model for Word-Form Recognition and Production. Ph. D. Thesis, University of Helsinki (1983).
3. Beesley, Kartunnen & Kartunnen, Lauri: "Finite – State Non- Concatenative Morphotactics" (2000).
4. Kager, R.: Optimality Theory. Cambridge University Press, United Kingdom (1999). Reprinted 2001
5. Fosler, J. E.: On Reversing the Generation Process in Optimality Theory (1996)
6. French, K. M.: Insights into Tagalog Reduplication, Infixation and Stress from Nonlinear Phonology. Texas, USA. The Summer Institute of Linguistics and the University of Texas at Arlington (1988)
7. Garcia, L.: Makabagong Gramar ng Filipino. Rex Printing Company, Quezon City, Philippines (1999)

8. Ramos, T. & Bautista, M.: Handbook of Tagalog Verbs Inflections, Modes, and Aspects. Honolulu, USA. University of Hawaii Press (1986)
9. Cheng, C. & See, S.: The Revised Wordframe Model for the Filipino Language. To appear in Vol. 3, No. 2, Journal for Research in Computing, Science and Engineering, Dela Salle University, Manila (2006)
10. Wicentowski, R.: Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework. PhD Thesis. Johns Hopkins University (2002)
11. Gelbukh, A. & Sidorov, G.: Approach to construction of automatic morphological analysis systems for inflective languages with little effort. Lecture Notes in Computer Science, N 2588, Springer (2003) 215-220
12. Gelbulk, A., Alexandrov, S. Y. Han.: Detecting Inflection Patterns in Natural language by Minimization of Morphological Model. Lecture Notes in Computer Science, N 3287, Springer (2004) 432-438