Linguistically-Motivated Automatic Morphological Analysis for Wordnet Enrichment

Tom Richens

Aston University
Aston Triangle, Birmingham B4 7ET, England

Abstract. Performance of NLP systems can only be as good as the lexical resources they employ. By modelling the evolved structure of language, there is scope for morpho-semantic enrichment of these resources. A set of linguistically-informed morphological rules is formulated from the CatVar database, implemented in a Java model of WordNet and tested on suffixation and desuffixation. Overgeneration and undergeneration are measured and an approach to improving these by using multilingual resources is proposed.

1 Introduction

The developers of statistically-based NLP techniques frequently report results with precision figures up to around 80%, without asking how far the qualitative limitations of the lexical resources employed might be degrading their results. The research presented here forms part of a project to address perceived inadequacies with respect to the representation of *morpho-semantic relations* in one of the most widely-used lexical resources namely *WordNet* (http://wordnet.princeton.edu/), [4].

1.1 Derivational Morphology as a Tool for Enriching Lexical Resources

However complex the mapping between morphology and semantics might be, *derivationally* related words must also be *semantically* related. Otherwise their morphological resemblance is co-incidental. The prerequisites for enriching lexical databases with morpho-semantic relations are correct identification of those morphological resemblances which are semantically significant and translation of these as semantic relations [1], [8]. Automated morphological relation discovery risks *overgeneration* (discovery of false relations) and *undergeneration* (failure to discover valid relations). To avoid these pitfalls requires linguistic rigour which can be applied by formulating language-aware morphological rules. If suffixed and non-suffixed forms, either of which can be generated from the other by the application of a well-informed rule both occur in the lexicon, then a derivational relation, however remote, exists between them, but if the rule is ill-informed, then there will be exceptions where the resemblance between them is co-incidental. A morphological rule can be formulated as a transformation between wordforms. In order to serve as a *semantic* tool it needs to define a meaning transformation, which is a *semantic relation* [1], [2].

Richens T. (2009).
Linguistically-Motivated Automatic Morphological Analysis for Wordnet Enrichment.
In *Proceedings of the 6th International Workshop on Natural Language Processing and Cognitive Science* , pages 36-45
DOI: 10.5220/0002171900360045

1.2 Previous Related Research

[5] proposes a mathematically-complex but language-ignorant approach to automatic affix identification from corpora. It can identify common English inflectional suffixes and might be an aid to deciphering text in a forgotten language, but does not yield any semantic information.

[6] present their *categorial variation* database, CatVar (http://clipdemos.umiacs.umd.edu/catvar/), which consists of clusters of morphologically related {word : part-of-speech} pairs, such that the same wordform may occur more than once in the same cluster as a different part of speech. Although they list their data sources, they say little about how these were combined. The approach to evaluation lacks clarity, so the reported 91.82% precision could be considered optimistic, though an independent calculation (table 2) gives 90.77%. CatVar was intended for WordNet enrichment along with other applications.

[13] suggests that the system of representation is more stable and explicit in Chinese than in languages with phonetic orthography, where the morphemic structure of one language may depend on another. The research presented here was largely motivated by the challenge of demonstrating that the representation of a European language with a multilingual dimension to its morphology is computationally tractable.

[1] suggest that *morphological* relations discovered in one language can be exported as *semantic* relations to enrich a wordnet in another language.

[2] propose the formulation of morphological rules to allow the automatic encoding of such relations. They observe that overgeneration but not undergeneration can be addressed by automatic cross-reference to a lexicon.

[9] proposes a *Morphodynamic WordNet*, connecting morphologically related words. He defines the morphogenesis of semantic forms as the generation of senses from a semantic nucleus or lexical root. This *tree* representation is superior to the *cluster* representation [6], in that it shows that there is always a rooted derivational hierarchy in any set of morphologically related forms.

[7] proposes a graph-based approach to discovery of morphological relations from a machine-readable dictionary which dispenses with the concepts *morpheme* and *affix*. He uses semantic information from dictionary definitions to support this, rather than inferring semantic information from morphology as proposed here.

1.3 Plan of this Paper

§2 reviews the CatVar database and introduces a lexicon based on WordNet 3.0. §3 describes the formulation and implementation of morphological rules and establishes density metrics for morpho-semantic relations. §4 presents the results from morphological rule application on 2 datasets, showing an improvement over CatVar clusters and using another dataset for comparison; the density of morpho-semantic relations discovered is compared with that in WordNet demonstrating the scope for enrichment. §5 analyses the causes of overgeneration and undergeneration in the results. §6 concludes with the proposal for the deployment of multilingual resources to improve on these results.

2 Experimental Resources

2.1 WordNet Model and Lexicon

WordNet is a lexical database consisting of *word senses,* connected by *lexical relations*, grouped into sets of synonyms (*synsets*), connected by *semantic* relations. This research was made possible by the development of a Java model of WordNet 3.0 in which synsets, word senses and relations are represented by instances of corresponding Java classes and appropriate subclasses. The data sources were the WordNet Prolog files downloaded from http://wordnet.princeton.edu/obtain. The model allows the interrogation and modification of the data in ways not possible with standard interfaces. While a word's absence from any lexicon is not conclusive, a single lexicon must be used for comparative measurements of overgeneration. The model includes a *lexicon* generated automatically from the WordNet data.

2.2 CatVar Sample Dataset

From the CatVar database a random sample was taken of 521 clusters of at least 3 pairs, comprising 2417 pairs altogether. The lexicon excludes 248 wordforms (10.3%) as the given part of speech. Of these 74 are legitimate uses of participles as adjectives or nouns, leaving 7.2% overgeneration. Morphologically unrelated examples are {chilli (n.): chilly (adj.)}, {compass (n.): compassion (n.)} and {stud (n.): student (n.)}. 49 words (2.0%) are morphologically unrelated to the headwords (verified against [3], [10] and [12]). For comparative purposes overgeneration is then 9.2%. Examples of undergeneration are failure to capture pairs {facial: face}, {quarterly: quarter} and {ripen : ripe}. 75 such related words were not found in the appropriate cluster.

3 Methodology

3.1 Formulation of Morphological Rules

Apart from 2 prefixations and a few abbreviations, the morphological transformations exhibited by the CatVar dataset are all cases of suffixation or of identical wordforms being used as different parts of speech. There are sufficient examples for rules to be formulated to encapsulate the morphological transformations between pairs of cluster members, on the understanding that they apply only where the wordforms generated can be validated against the lexicon.

Overgeneration can be a consequence of attempting to encode derivational morphology without reference to etymology. Correctly encoding morphological data requires correctly decoding derivational history, by unravelling language back through its evolution. Correct formulation of English suffixation rules presupposes an understanding that different rules apply depending on etymology. English words ending in "-ion" are generally derived from frequently irregular Latin *passive* participles. Consequently correct morphological rules require reference to Latin grammar. The derivation of English words from Latin *active* participles is complicated because some words of Latin origin have come into English directly while others have come via French: whereas Latin active participles end in "-ans" or "-ens" from which we get English adjectives in "-ant" or "-ent", French active participles always end in "-ant", resulting in English adjectives in "-ant" even when one would expect "-ent" from the Latin origin.

By examining the pairwise transformations exhibited by each cluster in the dataset, excluding overgenerations, a set of rules was formulated to encapsulate the transformations involved. Some rules referring to other languages have been formulated in such a way that a transformation from one English word to another can be applied, while others could not be implemented without reference to other languages' lexical resources. Some generalised spelling rules are included for the application and removal of suffixes. These do not apply to suffix substitutions. The ruleset is available from http://www.rockhouse.me.uk/Linguistics/Morphology/.

The rules apply to suffixation or desuffixation or to semantic relation identification. They comprise 5 fields (represented as for suffixation in table 1). A rule can only apply where the input matches the source part of speech. Where both wordform fields are empty, no morphological change applies but only a part of speech change; where the source wordform field is empty and the target is non-empty, the target wordform is a suffix which is appended to the input, subject to the generalised spelling rules; where both are non-empty, the rule only applies to an input whose wordform ends with the source wordform, replacing it with the target wordform, without reference to general rules. The Target part of speech is associated with the output. A desuffixation application needs simply to swap the Source and Target fields. The content of the *Relation* field expresses the semantic transformation which applies from Source to Target. The first example in table 1 is a *monolingual* rule to which generalised spelling rules do not apply; the second is a *multilingual* rule implemented monolingually, to which generalised spelling rules apply.

Some of the semantic relations, including those most frequently exhibited by the rules e. g. *pertainym* (23 rules) and *participle* (18), correspond to WordNet relations. The next most frequent is *gerund* (18 rules). The extensive set of nouns ending in "-ion" generally mean the same as an active (or occasionally passive) gerund. Despite their usually active meaning, these words are derived from the Latin passive participle, where a corresponding Latin verb exists. Where no Latin verb exists, they are most usually generated by appending the suffix "-ation". 191 rules have been formulated of which 156 have been implemented. The remainder require multilingual resources.

Table 1. Representation of morphological rules.

Source		Target		Semantic Relation
Wordform	POS	Wordform	POS	
ate	VERB	ative	ADJECTIVE	Participle
	VERB	ant	ADJECTIVE	Participle

3.2 Implementation of Morphological Rules

3.2.1 Autogeneration of Suffixed Forms

Suffixation and desuffixation algorithms were developed to apply the rules, each of which was defined as a transformation between two {morpheme: part-of-speech} pairs, to be applied to a {word : part-of-speech} pair to generate a second {word : part-of-speech} pair. Every input word is confronted with every rule. Where the rule is applicable to the input word, another is output. The algorithm exploits the lexicon for validation and the irregular inflection data from the WordNet exception files. The words output by the rules applied to a *seed* input word are re-cycled as input until no further valid output is generated. The total output from each seed is directed to a cluster of {word: part-of-speech} pairs, structurally identical to a CatVar cluster.

3.2.2 Application of Morphological Rules

To compare the output with CatVar itself, the algorithm was applied using the shortest word in each CatVar cluster as seed. Where there was more than one shortest word, all were used. The rules were also tested on a word list generated from the lexicon. Because the applicability of the ruleset might vary according to word length, random word lists were generated of word lengths from 4 to 14 characters. These lists were concatenated to form a list of 1108 wordforms from which 96 hyphenated forms were removed leaving 1012. This word length range facilitated a desuffixation experiment. The generalised spelling rules were adapted as desuffixation rules, similar to [11], though derived independently.

3.3 Potential for Enrichment of WordNet Relations

To explore the scope for morpho-semantic enrichment of WordNet, the proportion of morphological relations already encoded in WordNet, whether as derivational pointers or as other types of relation, needed investigation.

The functionality of the class used to represent a {word: part-of-speech} pair was extended to store the relations in which the WordNet senses of its wordforms participate. From this data the number of WordNet relations between all senses of the members of a cluster was calculated. WordNet derivational pointers were counted separately.

Since it is possible for more than one WordNet relation to exist between two synsets, the number of duplicate relations was also calculated. Assuming that each cluster member represents a unique sense, then the *maximum* possible number of relational pairings for any cluster (excluding duplicates), where there is a relation between each member of the cluster and every other member and **n** = the number of cluster members is given by $\mathbf{(n^2 - n) / 2}$.

Derivation being a directional phenomenon, viewing a cluster as a tree, while all members are related indirectly, each member is directly derived from at most 1 other member. The *correct* number of relations within each tree comprising unique senses, where **n** = the number of nodes is then given by **n - 1**.

Table 2. Comparison of Autogenerated Results with CatVar data.

Dataset	CatVar sample dataset	Autogeneration from CatVar sample dataset		CatVar sample dataset only	Auto-generation only	Common to both
Ruleset	n/a	Full	Restricted	Full	Full	Full
Not in lexicon	174	0	0	174	0	0
In lexicon but unrelated	49	70	0	44	65	5
In lexicon and related	2194	2432	2151	183	421	2011
Overgeneration	9.2%	2.88%	0%	n/a	n/a	n/a
Recall	Baseline	+3.52%	-11.01%	n/a	n/a	n/a
Precision	90.77%	97.20%	100%	n/a	n/a	n/a
TOTAL	2417	2502	2151	401	486	2016

4 Results

With the full ruleset, table 2 shows a 3.52% improvement in recall and a 7.08% improvement in precision over the CatVar baseline. Most of the 70 unrelated outputs were generated from an unrelated input, so that within any output cluster, one error would be the source of consequent errors. The adjective "moral" was incorrectly generated from "more" and led to 10 consequent overgenerations such as "moralise" and "morality". There were 3 {word : POS} pairs in the seed set which were not in the lexicon, 22 initial errors in applying the rules and 45 consequent errors.

In an attempt to eliminate all overgeneration, the 21 overgenerating rules were removed and the experiment was repeated with the *restricted ruleset*. 100% precision was achieved representing 10.17% improvement over the baseline at the price of a 11.01% deterioration in recall. 190 wordforms in the CatVar dataset were no longer represented. Of these only 3 were morphologically unrelated. Results achieved with the word list data are shown in table 3.

Table 3. Performance on Suffixation and Desuffixation with word list.

	Word list	Suffixation	Desuffixation	
Ruleset	n/a	Full	Full	Restricted
In lexicon but unrelated	n/a	19	39	14
In lexicon and related	n/a	768	887	729
Wordforms generated	1012	787	926	743
Recall	Baseline	+77.77%	+91.50%	+73.41%
Precision	n/a	97.59%	95.78%	98.11%
Overgeneration	n/a	2.41%	4.21%	1.88%
TOTAL	1012	1799	1938	1755

Table 4. WordNet relations between members of morphological clusters.

	CatVar dataset		Word list suffixation		Word list suffix stripping	
	Total	**Cluster average**	**Total**	**Cluster average**	**Total**	**Cluster average**
WN DERIV relations	1963	3.77	664	0.60	1008	0.91
All WN relations	2366	4.54	827	0.75	1278	1.15
DERIV as % of WN	82.97%		80.29%		78.87%	
Duplicate relations	86	0.17	26	0.02	34	0.03
Synsets / cluster		9.01		3.12		4.30
Max. relation count		70.98		18.54		27.95
% of max. realised	6.17%		3.90%		4.02%	
Correct relation count		8.01		2.12		3.30
% of correct realised	54.64%		34.14%		34.00%	

Table 4 correlates the WordNet relations between members of CatVar and output clusters, compared to the maximum and correct values for unique senses (§3.3). There is little variance between experiments in the proportion of the WordNet relations which are derivational pointers. However, using the original CatVar clusters yields a significantly higher relation count. This suggests that CatVar has already been used for WordNet enrichment [6], and that this enrichment has not been confined to derivational pointers. Given that the maximum and correct relation count would be greater if multiple senses are involved, the figures confirm the potential for further enrichment.

5 Analysis

5.1 Productivity and Overgeneration

Productivity was measured by lexicon-validated rule executions including duplicates generated by more than one rule. Table 5 shows the rules for which the ratio of initial and consequent overgenerations to rule applications >= 0.5 for both word list experiments, such that the rule is generating more wrong data than right data. With suffix stripping, the worst overgenerating rule was a monolingual implementation of a multilingually-formulated rule. Correct multilingual application of such rules could yield an improvement in performance.

Table 6 shows all the rules which overgenerated in both word list experiments. None of these rules are multilingual. Further investigation into the circumstances in which the worse-performing rules overgenerate may enable these rules to be re-formulated. Certain rules overgenerate below a threshold word-length [11], producing false associations such as between "fin" and "fine"; "read" and "ready", and between unrelated meanings of "still" as different parts of speech.

Table 5. Rules generating more wrong than right data on word list dataset.

	Source		Target			
	Word form	POS	Word form	POS	Over-generations per rule execution	Languages in formulation
Suffixation		V	ative	Adj.	3	1
		V	ed	N	1	1
	al	Adj.	ate	Adj.	1	1
	e	N	y	Adj.	0.75	1
		V	ant	Adj.	0.67	> 1
		V	ee	N	0.5	1
Suffix stripping	age	N		V	1.33	> 1
	ed	N		V	1	1
	en	V		N	1	1
	al	N		V	0.57	1
	eer	N		N	0.5	1
	man	N		N	0.5	1

5.2 Undergeneration

183 related wordforms in CatVar were not autogenerated: 28 plurals in "-s" were outside the scope of the rules; 20 undergenerations arose from non-implementation of rules requiring reference to Latin passive participles. Implementing these rules is the most important single improvement that could be made to the ruleset. 11 forms were not generated because no consistent rule could be found for the application of the "e-" suffix; 6 words were not generated because the rule required a different part of speech for either source or target; 5 root forms including "biology" and "vertebra" are missing from the CatVar dataset and consequently their derivatives were not generated.

Table 6. Persistently overgenerating rules.

				Output overgeneration / rule productivity	
Unsuffixed POS	Suffix	Suffixed POS	Langs. in formulation	Suffixation	Suffix stripping
N.	y	Adj.	1	0.14	0.09
V.	ed	N.	1	1	1
V.	ed	Adj.	1	0.02	0.11
Adj.	ly	Adv.	1	0.01	0.03

69 cases of undergeneration in desuffixation were identified plus 6 cases of consequent undergeneration. These include 6 POS mismatches and 5 singulars not generated from plurals; 12 undergenerations (17.39%) involve an unimplemented rule involving Latin passive participles; in 5 cases, both words have a French derivation, but the spellings do not correspond because they were imported probably at different

times from a language whose spelling was not yet standardised. 40.58% of undergenerations in desuffixation involve other languages.

6 Conclusions and Proposed Future Research

A linguistically-motivated and multilingually aware approach to discovering morpho-semantic relations has been demonstrated, which outperforms the CatVar database and can be applied directly to any lexicon without other resources.

There is plenty of scope for enriching WordNet with data relating to derivational morphology. The Java model of WordNet is a firm foundation for implementing and demonstrating this enrichment. A set of new types of relation has been proposed to capture the semantics . Further research will verify their applicability.

Some morphological rules are unreliable as implemented, and need more rigorous formulations. Implementation of appropriate word length thresholds would allow the automatic processing of regular longer words while shorter words are checked manually. Further rules could be formulated by examining suffixes in the lexicon without CatVar.

Some of the most important morphological rules have not been implemented, for lack of multilingual resources. Others have been implemented monolingually, accounting for much overgeneration. The most important cause of undergeneration is non-implementation of multilingual rules, especially with reference to Latin participles. Implementing these rules is the most important single enhancement that could be made. This will be a significant area of further research, leading to a fully enriched morpho-semantic database.

References

1. Bilgin, O., Çetinoğlu, Ö. & Oflazer, K. (2004). Morphosemantic Relations In and Across Wordnets, Proceedings of the Second International WordNet Conference, Brno, Czech Republic, January 20-23, 2004, 60-66.
2. Bosch, S., Fellbaum, C. & Pala, K. (2008). Enhancing WordNets with Morphological Relations: A Case Study from Czech, English and Zulu, Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, Jan. 22-5 2008, 74-90.
3. COED (1971-80). The Compact Edition of the Oxford English Dictionary, Complete Text Reproduced Micrographically, Oxford University Press.
4. Fellbaum, C. (ed.) (1998). WordNet: An Electronic Lexical Database, Cambridge, MA., MIT Press.
5. Goldsmith, J. (2001). Unsupervised Learning of the Morphology of a Natural Language, Computational Linguistics, 27, 153-198.
6. Habash, N. & Dorr, B. (2003). A Categorial Variation Database for English, Proceedings of the North American Association for Computational Linguistics, Edmonton, Canada, 96-102.
7. Hathout, N. (2008). Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy, Proceedings of the 3rd Textgraphs workshop on Graph-based Algorithms for Natural Language Processing, 22nd International Conference on Computational Linguistics , Manchester, 24 August, 2008.

8. Koeva, S., Krstev, C. & Vitas, D. (2008). Morpho-semantic Relations in WordNet: A Case Study for two Slavic Languages, Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, Jan. 22-5 2008, 239-253.
9. Mbame, N. (2008). Towards a Morphodynamic WordNet of the Lexical Meaning, Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, Jan. 22-5 2008, 304-310.
10. Onions, C. T. (Ed.) (1966). The Oxford Dictionary of English Etymology, Oxford, Clarendon Press.
11. Porter, M. F. (1980). An algorithm for suffix stripping, Program, 14(3). 130-137.
12. Simpson, D. P. (1966). Cassell's New Latin Dictionary, London, Cassell, 4th. Edition.
13. Wong, S. H. S. (2004). Fighting Arbitrariness in WordNet-like Lexical Databases. A Natural Language Motivated Remedy, Proceedings of the Second International WordNet Conference, Brno, Czech Republic, January 20-23, 2004, 234-241.