Reconciliation of Unsupervised Clustering, Segmentation and Cohesion

David M. W. Powers
Department of Computer Science
The Flinders University of South Australia

powers@acm.org

Abstract

This extended abstract examines the progress of a project on unsupervised language learning, and focuses on two different approaches to segmentation, as well as how cohesion may be generalized from it definitive morpho-syntactic instantiation. It is intended as a discussion paper, and outlines the specific hypotheses currenlty being tested.

1. Introduction and Motivation

This extended abstract summarize recent and current work being carried out by my group in relation to unsupervised learning of language.

The work described is unsupervised in the sense that

- there is no human preprocessing of the raw corpora;
- there is no preconceived target grammar;
- there is a minimum of imposed formalism;
- there is no tuning for particular languages/datasets.

The methodology used has been inspired by linguistic, psycholinguistic and cognitivist research in language and vision. Whilst a connectionist framework is not in general used, the approach is intended to be neurologically plausbile, although there is more direct influence from probability theory and information theory, and some experiments have used self-organizing neural nets (Powers, 1989; Schifferdecker, 1994).

Our earliest work focussed on learning syntax by finding statistical correlations or using self-organizing time-delay networks (Powers, 1989). Hierarchical grammars were produced by introducing newly found relationships as new candidates for correlation. Reasonable grammars were produced only for training date of consistent short length phrases or sentences, but the experiment lead to significant insights: the closed class elements (function words in the initial experiments) were learned first, and these acted like seeds which expanded into larger and larger grammatically meaningful units.

This research was subsequently generalized to a binary clustering approach inspired by Pike's Phonemic (1949) and Tagmemic methodologies (1977). Tokens that function similarly in some sense (phonological, morphological, syntactic or semantic) but represent systematic rather than free variation, will form Complementary Distributions or classes. However, free distributions and distributions based on correlations with information which is unavailable at the current level of analysis will not be distinguishable, and thus complementarity will not be established, and may only be assumed under the hypothesis that there is no such thing as free variation or arbitrariness, and that apparent free variation always has causal roots. The complementary method of Pike is Contrast in Identical or Analogous Environments (CIE/CAE) and *assumes* that there is extraneous information about whether or not the units belong to the same emic unit or not. If we do *not* make any use of extra-sentential information, and rather *assume* that all units with distinct forms at the current level convey distinct information, but where they occur in the same set of contexts (coset) they are treated as similar and belong to the same class and convey some common information. The set of units which can occur in a contextual coset are Contextual Distributions and have the same basic character as Complementary Distributions, and are thus treated as such.

2. Speech and Phonology

Whilst the outlawing of free variation looks like heresy at the speech and phonetic levels, it simply means that information that is not relevant to a phonemic transcription is discarded as useless in models that allow free variation, while in deeper models, explanations should be available. One of our current projects is focussing precisely on this supraphonemic information, both from the perspective of capturing supralinguistic information (speaker attributes/mood etc.) for its own sake and with a goal of tracking speakers against complex auditory backgrounds.

Schifferdecker (1994) has successfully used the technique to produce phonemes from raw speech data as well as from raw phonetic transcriptions, although this work did not explore the hierarchical aspects of the technique (except as a consequence of dendritic representation of the classification space).

3. Semantics and Synonymy

At the semantic level, it supports our denial of the

David M. W. Powers (1998) **Reconciliation of Unsupervised Clustering, Segmentation and Cohesion.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning,* ACL, pp 307-310.

existence of pure synonymy. Thus words like 'too' and 'also' which are apparent synonyms have quite different syntactic constraints, while words like 'small' and 'little' which are apparent synonyms and appear to occupy the same same syntactic role, actually have quite different connotations. Thus 'a small boy' is small for his age, whilst 'a little boy' is a young child, and 'a small little boy' combines these implications; the tendency for 'little' to prefer and be preferred when a metaphorical interpretation is appropriate is confirmed by idioms like 'a little while' and 'a little bit' whereas 'small' tends to have more direct connection to the underlying spatial interpretation, and when used in a metaphorical or temporal context it thus tends to reinforce the metaphor and supply additional emphasis — contrast 'except for one little detail' and 'except for one small detail'. Of course, any examples of this sort are highly influenced by the specific language, dialect and idiolect of the speaker and may vary at each level.

Many researchers have used clustering techniques to induce semantic classes (e.g. Finch, 1993), although these have tended again to be non-hierarchical except to the extent that a pairwise clustering technique induces a dendritic structure on the semantic space (although Finch did perform two levels of analysis in one experiment).

4. Morphology and Syntax

At the syntactic level the assumption of absence of free variation is not so controversial, and although generative grammarians have tended to treat some choices as arbitrary, e.g. the choice between active and passive, which is probably more a function of their ignoring pragmatics to focus on grammar. At the level of morphology, what may appear to be free variation synchronically usually has a diachronic explanation, and invariably involves clear complementarity in terms of the distribution of allomorph according to the syntactic role of the embedding word.

The experiments of Powers (1992) demonstrated both learning of classes and hierarchical rules from the character level up to the level of simple noun phrases and simple clauses. As is the case with subjacency, noun phrases and clauses tend to act similarly, and indeed we propose that they themselves form a complementary distribution (involving their multiple forms, including nominalized clauses and verbs: 'he wanted the girl to come', 'the girl must come', 'he decided that the girl should come', 'he decided the girl could come') and suggests a generalization of the finiteness feature of verbs should apply to both nouns and verbs and their dominated structures ('to' and 'that' are both optional markers of the infinite form; the finite form would appear to be the default role of a verb and the unmarked form).

5. Segmentation and Grammar

Both Powers (1989) and Powers (1994) depend for their hierarchical organization on a fuzzy approach to segments. At the word level, Powers (1989) allowed four hypotheses: a word should group with the word to the left or the word to the right, or with a phrase to the left of a phrase to the right, where a phrase has previously been recognized as a candidate group. Hypotheses were rated according to their usage, and those involved in the most highly rated overall parse were reinforced. Powers (1992) allowed one or two (or in some experiments three) given or induced units to operate as a putative unit for the purposes of distributional analysis. Apart from thresholding (to eliminate noise, and to make it amenable to the small computer available), frequency information was ignored and each context was associated with a coset of (one to three) units on either side. Classes were formed by a technique which turns out to be clustering using a Hamming distance of 2 (or 3 in some experiments), in which classes can be merged (union) and the common coset determined (intersection).

The size and coverage of the individual left and right cosets and their union and intersection gave eight measures of the strength of a class, and in all cases identified the vowels as the strongest class for the original dictionary corpus, and for most other corpora tried, with right context appearing more useful than left, coset size being more accurate than coset coverage, union size being more reliable than intersection size. Note that Powers (1997a) generalizes the approach and considers a multitude of different clustering metrics and methods, introducing a pair of goodness measures which allow a more principled approach to closing and evaluating clusters (rather than closing at a specific cluster, you close when the goodness measure reaches its first local maximum).

In the Powers (1992) experiments, classes were added as new units and the process was repeated. The fuzzy variable size candidate units for the next level meant that hyperclasses of context-free rules were learned. However the grammar led to high levels of ambiguity using non-deterministic parsing, and the presented hierarchy is based arbitrarily on a simple greedy approach, but (for this reason) performance as a recognizer/parser was not evaluated.

Though in this work phonologically, morphologically and grammatically meaningful classes and structure were formed, up to phrase/clause level, no interpretation of the structures or classes was offered, and no attempt was made to discover or propose cohesive constraints or semantic relationships. At the same time however, Entwisle and Groves (1994), Powers (1997b) and Entwisle and Posers (1997) have produced a constraint parser which uses precisely the kind of morphological and grammatical classes which are thrown up by the self-organizing and clustering experiments, and have started to address how one develop meaningful statistics for a true grammar learning system without any preconceived notions of what the correct parse/phrase structure is (if any). In particular Powers (1997b) performed experiments in the context of grammat checking application, using automatic segmentation techniques based on those of Harris (1960) and similar to those used by Brent (1997), but combined with context-conditioned probabilities which were used to decide between confusable words. The same technique has been applied in a Loebner Prize entry by Bastin and Cordier (1997).

This gives us two competing approaches to segmentation. In the first, segmentation is a side effect of the fuzzification of input units during classification (the segments chosen are those which give the best classification according to some metric). Incidentally, Powers (1992) also reports work in which hyphenation points were marked, thus introducing an element of supervision, but it did not improve performance (which agained suffered from ambiguity and thus didn't produce definite results, being non-probabilistic, although a greedy algorithm performed quite reasonably). The second (Harris) approach examines the conditional information or perplexity for each possible prefix/suffix to determine likely segmentation points — which is expected to show a local maximum in the perplexity.

6. Reconciling the Methods

The Harris (1960) approach works on the insight that within a unit, particularly a closed class functional unit such as an affix, there is less freedom of choice than at the boundary of units. This depends strongly on the fact that the number of affixes is much lower than individual characters, whilst their frequency is so much higher than the morphs they collate with. Viz. they define large cosets.

The Powers (1992) approach works by finding the groups of segments which have the largest cosets, and thus have high frequency and low information, their information content tending to be more syntactic than semantic. The segmentation and classification occur simultaneously, and it seems there is no advantage to doing perplexity-based segmentation before doing the classification, although this has not yet been investigated.

The segmentation process may however be repeated, finding the subsequent perplexity or information maxima. In addition, even the initial functional segments found may be used directly to learn or check a grammar (Entwisle and Groves, 1994; Powers, 1997b), although this already makes use of the known word segmentation and the assumption, which is for English is an excellent first approximation, that affixes are either word initial or word final, and that it is this prefixes and suffixes which determine the syntactic roles of the words.

7. Augmenting the Methods

The approach used by Entwisle and Groves (1994) is only semi-automatic, and wasn't originally conceived as a learning system. When a sentence fails to parse, it means that a constraint must be relaxed, and this constraint is identified manually—being a system which involves no statistics, which is being trained on text which may contain errors (e.g. one error was discovered in the first chapter of the Alice Corpus, Carroll, 1865), and where the relaxation may involve the supplying of new roles or the removal of a constraint at any one of a number of possible points.

The approach used by Powers (1997b) is only intended to identify typing errors and substitution errors (e.g. 'there' for 'their') and builds and stores a differential grammar only when the word can be disambiguated from its closed-class context, but already constraints based on the closed class words and functional affixes suffices to perform better than commercial grammar checkers.

The segmentation and classification methods on their own do not attempt to check cohesive constraints, such as agreement, but doing so could be expected to reduce the ambiguity which is so rife. Powers (1992) reports one *word* with around 5000 different 'parses'.

The specific approach we are using in our current work is to extend the structure determined by a version of the approach of Powers (1992 and 1997a) which produces binary grammar rules. The extended structure augments a higher level unit with features constructed from or inherited from the lower level units. This construction is being carried out virtually at present, while we examine the best way to propogate information, and we investigate and seek to differentiate the specific hypotheses that (a) the more frequent, or (b) the higher perplexity, segments play the morpho-syntactic cohesive roles, whilst their binary siblings hold the primary content to be retained and passed on.

Whilst this strategy is the one suggested by the primary morphological cohesion, and could straightforwardly be applied after a single segmentation pass, using the hierarchical classification approach produces a far stronger hypothesis, predicting that vowels in English, where they are strongest under both conditions (a) and (b), play a primarily structural or phonological role, and that affixes, prepositions, articles, relatives, conjunctions and the like act as the heads of their superordinate structures.

An additional aim of the present project is to seek to tease apart homonyms and their manifestations at the other levels, including the dual role of the letter 'y' (sometimes clearly vowel as in 'xylophone', sometimes ambiguously consonantal as in 'play, playing, played'), the suffix '-s' and the word 'to'. In Powers (1997a) both 'y' and space were identified as vowels using certain clustering techniques and methods (and the issues are discussed in that paper). We are generalizing the approach of identifying a class, such as the vowels, and then identifying those units, such as 'y', which atypically have a larger coset than the class which has been selected as having maximal coverage (resolving the Powers (1992) dilemma in favour of coverage as the preferred metric).

8. Discussion and Conclusion

This extended abstract documents work in progress, contrasting existing approaches in recent publications and setting out the direction we are following. Preliminary results should be available at the workshop, but the paper is mainly intended to provoke discussion of the pro's and con's of the two approaches to segmentation.

9. References

Bastin, V. and Cordier, D. (1998). Methods and tricks used in an attempt to pas the Turing Test, *NeMLaP3/CoNLL98 Workshop on Human Computer Conversation.*

Brent, M.R. (1997). a Unified Model of Lexical Acquisition and Lexical Access. *Journal of Psycholinguistic Research* **26**:363-375.

Carroll, L. (1865). *Alice's Adventures in Wonderland.* The Millennium Fulcrum Edition 2.9, Gutenberg Project.

Entwisle, J. and Grovers, M. (1994). A Method of Parsing English Based on Sentence Form, *NeMLaP,* 116-122.

Entwisle, J. and Powers, D.M.W. (1998). The Present Use of Statistics in the Evaluation of NLP Parsers. *NeMLaP3/CoNLL98 Joint Conference.*

Finch, S. (1993) *Finding Structure in Language.* Ph.D Dissertation, University of Edinburgh

Harris, Z. (1960) *Structural Linguistics.* University of Chicago Press

Pike. K. (1949) *Phonemics*, University of Michigan Press

Pike, K. and Pike, E. (1977) *Grammatical Analysis*, SIL

Powers, D.M.W. and Turk, C.C.R. (1989). *Machine Learning of Natural Language.* Springer-Verlag

Powers, D.M.W. (1992). On the significance of Closed Classes and Boundary Conditions: Experiments in LExical and Syntactic Learning,. *SHOE* 245-266. ITK, University of Tilburg, NL

Powers. D.M.W. (1997a). Unsupervised Learning of Linguistic Structure: An Empirical Evaluation, *Journal of Corpus Linguistics*

Powers, D.M.W. (1997b). Learning and Application of Differential Grammars. *CoNLL97*, ACL, Madrid, Spain

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* Singapore:World Scientific

Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication.* Urbana: U. Illinois Press

Schifferdecker, G. (1994) *Finding Structure in Language.* Diplom Thesis, University of Karlsruhe.

Zipf, G.K. (1949) *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* AW