Learning Feature-Value Grammars from Plain Text

Tony C. Smith
Department of Computer Science, University of Waikato
Hamilton, New Zealand
tcs@cs.waikato.ac.nz

Abstract

This paper outlines preliminary work aimed at learning Feature-Value Grammars from plain text. Common suffixes are gleaned from a word suffix tree and used to form a first approximation of how regular inflection is marked. Words are generalised according to these suffixes and then subjected to trigram analysis in an attempt to identify agreement dependencies. They are subsequently labeled with a feature whose value is given by the common suffix. A means for converting the feature dependencies into a unification grammar is described wherein feature structures are projected on to unlabeled words. Irregularly inflected words are subsumed into common categories through the process of unification.

1 Motivation

Unification grammars (UGs) have become the established formalism for natural language understanding systems, primarily because of their clean denotational semantics and their ability to capture complex grammatical constraints through feature dependencies (Uszkoreit & Zaenen, 1996). But engineering even modestly sized UGs can take a very long time, making the idea of constructing a comprehensive, robust, competent UG by hand virtually intractable. Recent advances in stochastic language modeling, however, have made it possible to incorporate statistical information into UGs (Abney, 1996 and Smith & Cleary, 1997), thus giving access to the complexity estimates now widely regarded as essential for automatically learning adequate grammars from positive data alone. But this still leaves open the question of exactly how such learning can be achieved for UGs.

A probabilistic unification grammar (PUG) has three principal components: 1) a context-free account of linear precedence relations, 2) a set of features for expressing grammatical dependencies, and 3) probability distributions for the rules and features. Methods for unsupervised learning of the first and last of these components have already been suitably worked out. For example, the context-free description can be addressed with solutions borrowed from work in learning PCFGs (Jelinek et al, 1992), and the distribution can be estimated by training on sample data (Eisele, 1994 and Brew, 1995). The outstanding problem then is how to derive a satisfactory set of features in the absence of overt semantic information.

This paper describes preliminary work aimed at learning a Feature-Value Grammar from plain text. It is based on the generally held notion that syntactic agreement and morphological inflection are closely related (Abney, 1987 and Fukui & Speas, 1986). Morphological clues about inflectional affixes are gleaned from the vocabulary of a language using a word suffix tree. Common suffixes are assumed to identify related semantic elements undergoing the same inflectional process, allowing the contexts in which they occur to be generalised through the creation of feature structures. Feature values are set according to the common suffix and projected on to unlabeled words. The contexts and the agreement constraints are thereafter expressed using a unification formalism. Irregularly inflected words are subsumed into existing categories by unifying the contexts in which they occur with those established for regularly inflected words.

Tony C. Smith (1998) **Learning Feature-Value Grammars from Plain Text.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98 Workshop on Paradigms and Grounding in Language Learning,* ACL, pp 291-294.

2 Feature identification

A UG encodes lexical properties as feature structures (specifying such things as part-of-speech, number, tense, person, thematic role, etc.) whose values percolate up through a subsumption hierarchy by the process of unification (Sanfilippo, 1993). Syntactic constraints are imposed by forcing agreement between features of grammatically related structures.

Kazman (1994) argues that features correspond to semantic properties associated with thematic categories (e.g. nouns, verbs and adjectives) and that learning syntax is equivalent to figuring out how these properties impose constraints on the functional categories (e.g. determiners, auxiliaries, and complementizers) of a particular language. This study takes the slightly stronger position that the process by which thematic and functional categories are combined is mediated by morphological inflection. Like Kazman's system, *Babel*, the focus is on the role of inflectional affixes in the acquisition of agreement. But unlike *Babel*, which makes inferences over semantically related words identified through set operations on input already tagged with attributes, this work addresses feature identification as a bootstrapping problem, where inflectional affixes and the constraints they impose are inferred from plain text.

A first approximation

The first objective is to detect when and how inflection is manifest. This is addressed through generalisation on a word suffix tree (WST) constructed for the vocabulary of the language. A WST is a derivative of a letter-based multiway trie built from an ordered set of words. Each distinct sequence of characters along a path in the trie is collapsed into a single node, resulting in a WST for which all leaf nodes are common suffixes to the prefix terminated by their parent node (Andersson, 1996). A sample portion of a WST is shown in Figure 1. Note that the symbol *$* is a kind of NULL suffix, which shows that the parent node is itself a suffix and thus corresponds to the end of an actual word. It follows that its leaf nodes correspond to genuine morphological suffixes.

Given that regular inflection is largely realised through suffixion on root categories, a first approximation of these categories may be given by assigning a common lexical identity to words that share the same set of suffixes. That is, it is assumed that words which inflect in the same ways likely belong to the same syntactic category. Clearly not all suffixes are inflectional. Therefore, some general restrictions

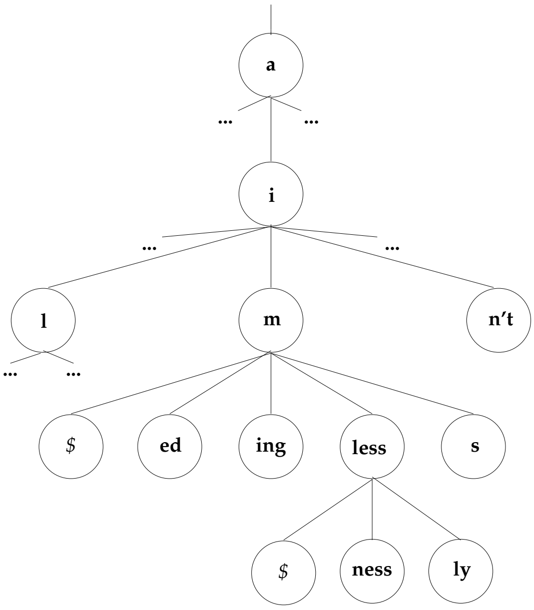

Figure 1: Portion of a word suffix tree.

are applied in an analysis of the WST in an effort to garner a set of possible inflectional suffixes. First, any suffix which has a suffix itself cannot be inflectional, based on the assumption that inflectional suffixes always occur at the end of a word. Second, root categories must have at least two inflected forms, thus a prefix may only be a possible root category if it appears to have at least two inflectional suffixes. The corollary is that a suffix is not inflectional if it is the only inflectional suffix. Under these restrictions, the suffix set for "aim" in Figure 1 would be {*$* , *ed*, *ing*, *s*}.

Inflection in context

To identify which suffixes are grammatically significant, a contextual analysis of how they are used must be carried out. This can be done by generalising over the trigrams of a large sample of text in which each word has been replaced by its corresponding suffix as given by the WST analysis. Almost all functional categories and irregularly inflected words have no inflectional suffixes associated with them and are therefore left unchanged.

The trigrams are sorted and processed in decreasing order according to their frequency. Feature structures are hypothesised to reconcile trigrams that differ in only one term. For example, some of the most frequent trigrams might be as follows:

1)	the	*-s*	were
2)	the	*-s*	*-$*
3)	the	*-$*	was
4)	the	*-$*	*-s*
5)	the	*-$*	*-ed*
6)	the	*-s*	*-ed*

Assume in this instance that *-s* has replaced *dogs* in phrase 1, 2 and 6, and *$* has replaced *dog* in phrases 3, 4, and 5. In addition, assume that the prefix is *walk* for the suffixes given in the third position for phrases 2, 4, 5 and 6.

These six related trigrams imply an agreement constraint that can be captured with a feature structure. For example, *were* and *-$* appear after the context *the -s*, but not after *the -$* , and *was* and *-s* occur after *the $* but not after *-s*, indicating a possible dependency between the last two terms. Phrases 5 and 6 imply a common syntactic role for *$* and *-s*, thus we might infer that the dependency is one of feature agreement. As the second term is uniformly a suffix, we might assume that it projects the agreement and is therefore inflectional. To characterise this, we associate a feature with the words appearing in the second position, and assign it the value of the suffix in each instance, giving the following lexical entries:

dog($F1$ =*$*)
dogs($F1$ =*-s*)

To characterise the dependency in the first four phrases, we project the feature structure on to the words in the third position, assigning the corresponding feature value needed to preserve the dependency, as follows:

dogs($F1$ =*-s*)
dog($F1$ =*$*)
were($F1$ =*-s*)
walk($F1$ =*-s*)
was($F1$ =*$*)
walks($F1$ =*$*)

Phrases 5 and 6 must be made to have the same feature structure, but this appears to entail assigning two different values to the feature structure for *walked*. However, given that *walked* is not constrained by this particular feature, its value can be left ungrounded, giving:

dogs($F1$ =*-s*)
dog($F1$ =*$*)
were($F1$ =*-s*)
walk($F1$ =*-s*)
was($F1$ =*$*)
walks($F1$ =*$*)
walked($F1$ =X)

From this limited set of phrases, it appears unnecessary to extend the inflectional constraint to the word *the*. However, given a trigram of the form "*a $ was*" without the complementary trigram "*a $ were*", agreement would force projection of the feature structure on to the determiner.

Once a word has been identified as an inflected form, this provides additional information for the generalisation of subsequent trigrams. If a term is known to project an agreement constraint in one instance, this curtails the number of hypotheses that must be tested to determine the source of any new constraints. That is, if *were* and *walk* come up in another set of related trigrams, the existing feature $F1$ can be trialed first as a possible explanation.

Capturing the syntactic constraints

Deriving features in the manner described in the previous section provides an account of inflectional agreement. To translate this into syntactic constraints requires the addition of corresponding unification rules. Thus, as each trigram is processed, any changes to the feature structure must generate a rule that captures the linear precedence relation. This can be done efficiently with logic programs, such as Prolog DCGs. Initially, the grammar is formed by generating clauses to cover dependencies between pairs of terms, annotated with the appropriate feature structures and values. The grammar is built up by combining adjacent clauses and unifying their variables (i.e. features). The unification also allows rules for irregular inflections to be transformed into a more general form.

Irregular inflection

Irregular words may follow some of the same inflectional patterns as regular words, such as the present and gerundive forms in English verbs, and thus can be generalised with the same mechanism. In other instances, they may force the creation of a new feature structure which captures the same agreement constraint. To avoid this, every new rule is compared against existing rules to see if they have a common structure which can be generalised. Only rules which differ by a single term need to be examined, and only if the features of that term are grounded in the established case. If the new rule can be unified with an old rule by a consistent change in its corresponding feature values, then the lexicon is adjusted and the new rule is discarded. Since irregular forms do not differ in their usage, sufficiently large samples of text (enough to cause a match between rules) will allow the same agreement constraint to be captured by one rule. This solution also applies to words whose suffixing is irregular because of orthographic conventions, as when *abated* and *abates* are categorised by the suffix set $\{d, s\}$ instead of the more common $\{ed, s\}$.

3 Remarks

To the extent that inflectional agreement morphology and syntactic agreement structures are linked, generalisation over inflectional suffixes is likely the only means by which a unification grammar can be learned from plain text. This work represents an initial attempt at doing just that.

The WST is a suitable data structure for uncovering suffixes, but is insufficient for identifying those which mark inflection. This requires a characterisation of how individual suffixes are used contextually, and identification of instances where they appear to impose agreement constraints.

Limiting context analysis to trigrams has the obvious disadvantage that long distance dependencies cannot be reliably inferred unless they happen to percolate up through a series of unification operations between smaller phrases. It is possible that some statistical techniques for finding lexical dependencies, such as those used in constructing link grammars, would be a more effective way to build feature structures and the grammar.

Perhaps the most appealing aspect to this approach is that it attempts to combine morphological constraints and syntactic constraints within a single model for grammar induction. In so doing it has uncovered a number of interesting problems and ideas which should generate interesting discussions in a language learning workshop.

References

Steven Abney. *The Noun Phrase in its Sentential Aspect*. PhD thesis, MIT, 1987. unpublished.

Steven Abney. Stochastic attribute-value grammars. *The Computation and Language E-Print Archive*, page 21, October 1996. 9610003.

A. Andersson, N. Jesper Larsson, and Kurt Swanson. Suffix trees on words. In D. Hirschberg and G. Myers, editors, *Lecture Notes in Computer Science 1075, Combinatorial Pattern Matching*, pages 102–115. Springer-Verlag, 1996.

Chris Brew. Stochastic hpsg. In *Proceedings of EACL-95*, 1995.

S. F. Chen. *Building probabilistic models for Natural Language*. PhD thesis, Harvard University, Cambridge, Massachusetts, Cambridge, Mass., 1996.

Andreas Eisele. Towards probabilistic extensions of constraint-based grammars. Deliverable r1.2.b, DYANA-2, September 1994.

Naoki Fukui and Peggy Speas. Specifiers and projection. *MIT Working Papers in Linguistics*, 8:128–172, 1986.

F. Jelinek, J. D. Lafferty, and R. L. Mercer. Basic methods of probabilistic context-free grammars. In *Speech Recognition and Understanding: Recent Advances, Trends and Applications. Proceedings of the NATO Advanced Study Institute*, pages 345–360, 1992.

Rick Kazman. Simulating the child's acquisition of the lexicon and syntax—experiences with *babel*. *Machine Learning*, 16:87–120, 1994.

A. Sanfilippo. Lkb encoding of lexical knowledge. In T. Briscoe, A. Copestake, and V. de Paiva, editors, *Default Inheritance within Unification-Based Approaches to the Lexicon*. Cambridge University Press, 1993.

Tony C. Smith and John G. Cleary. Probabilistic unification grammars. In *Workshop Notes: ACSC '97 Australasian Natural Language Processing Summer Workshop*, pages 25–32, Macquarie University, February 1997.

Hans Uszkoreit and Annie Zaenen. Grammar formalisms. In Ron Cole, editor, *A Survey of the State of the Art in Human Language Technology*, chapter 3.3. Center for Spoken Language Understanding, University of Pisa, Italy, 1996.