TALN 2004, Fez, 19-22 April 2004

Clustering of Word Types and Unification of Word Tokens into Grammatical Word-Classes
Le Regroupement de Types de Mots et l'Unification d'Occurrences de Mots dans des Catégories grammaticales de mots

Eric Atwell
School of Computing – University of Leeds
Leeds LS2 9JT, England
eric@comp.leeds.ac.uk

Résumé – Abstract

Ce papier discute la Néoposie: l'inférence auto-adaptive de catégories grammaticales de mots de la langue naturelle. L'inférence grammaticale peut être divisée en deux parties : l'inférence de catégories grammaticales de mots et l'inférence de la structure. Nous examinons les éléments de base de l'apprentissage auto-adaptif du marquage des catégories grammaticales, et discutons l'adaptation des trois types principaux de marqueurs des catégories grammaticales à l'inférence auto-adaptive de catégories grammaticales de mots. Des marqueurs statistiques de n-grammes suggèrent une approche de regroupement statistique, mais le regroupement n'aide ni avec les types de mots peu fréquents, ni avec les types de mots nombreux qui peuvent se présenter dans plus d'une catégorie grammaticale. Le marqueur alternatif d'apprentissage basé sur la transformation suggère une approche basée sur la contrainte de l'unification de contextes d'occurrences de mots. Celle-ci présente un moyen de regrouper des mots peu fréquents, et permet aux occurrences différentes d'un seul type de mot d'appartenir à des catégories différentes selon les contextes grammaticaux où ils se présentent. Cependant, la simple unification de contextes d'occurrences de mots produit un nombre incroyablement grand de catégories grammaticales de mots. Nous avons essayé d'unifier plus de catégories en modérant le contexte de la correspondance pour permettre l'unification des catégories de mots aussi bien que des occurrences de mots, mais cela entraîne des unifications fausses. Nous concluons que l'avenir peut être un hybride qui comprend le regroupement de types de mots peu fréquents, l'unification de contextes d'occurrences de mots, et le 'seeding' avec une connaissance linguistique limitée. Nous demandons un programme de nouvelles recherches pour développer une valise pour la découverte de la langue naturelle.

This paper discusses Neoposy: unsupervised inference of grammatical word-classes in Natural Language. Grammatical Inference can be divided into inference of grammatical word-classes and inference of structure. We review the background of supervised learning of Part-of-Speech tagging; and discuss the adaptation of the three main types of Part-of-Speech tagger to unsupervised inference of grammatical word-classes. Statistical N-gram taggers suggest a statistical clustering approach, but clustering does not help with low-frequency word-types, or

with the many word-types which can appear in more than one grammatical category. The alternative Transformation-Based Learning tagger suggests a constraint-based approach of unification of word-token contexts. This offers a way to group together low-frequency word-types, and allows different tokens of one word-type to belong to different categories according to grammatical contexts they appear in. However, simple unification of word-token-contexts yields an implausibly large number of Part-of-Speech categories; we have attempted to merge more categories by "relaxing" matching context to allow unification of word-categories as well as word-tokens, but this results in spurious unifications. We conclude that the way ahead may be a hybrid involving clustering of frequent word-types, unification of word-token-contexts, and "seeding" with limited linguistic knowledge. We call for a programme of further research to develop a Language Discovery Toolkit.

Keywords – Mots Clés

Corpus, marquage des catégories grammaticales, regroupement, unification, catégories de mots, type/occurrence, évaluation.
Corpus, Part-of-Speech tagging, clustering, unification, word classes, type/token, evaluation

1 Neoposy: unsupervised inference of grammatical word-classes in Natural Language

A first stage in Unsupervised Natural Language Learning (UNLL) should be the partitioning or grouping of words into word-classes. According to the Collins English Dictionary, "neology" is: ***a newly coined word, or a phrase or familiar word used in a new sense; or the practice of using or introducing neologies***. In the language of computational linguists, "PoS" has evolved from an acronym, into a full word-lemma participating in compound terms such as pos-tagger, pos-tagged corpus, bi-pos model, uniposy/polyposy … So, as a logical extension, (Atwell 2003) proposed "neoposy" as a neology meaning "***a newly coined classification of words into Parts of Speech; or the practice of introducing or using neoposies***".

Neoposy could be a useful first phase in discovering structural patterns in Grammar Inference research. Neoposy may prove interesting in its own right, helping linguists to discover new perspectives on Part-of-Speech analysis, particularly in studies of languages very different from English. Traditional PoS (based on grammatical-word classes in Latin, Greek, and Sanskrit) may not fit some languages: "Conventional notions of 'parts of speech', in particular, really belong to Indo-European and can be highly confusing when imposed on Asian languages such as Malay… The alternative is a genuine data-driven approach, identifying the categories that actually appear in the texts." (Knowles and Don 2003)

2 Clustering words into word-classes

A range of approaches to clustering words into classes have been investigated (eg Atwell 1983, Finch and Chater 1992, Hughes and Atwell 1994). In general these researchers have tried to cluster word-types whose representative tokens in a Corpus appeared in similar contexts, but varied what counts as "context" (eg all immediate neighbour words;

neighbouring function-words; wider contextual templates), and varied the similarity metric and clustering algorithm.

This approach ultimately stems from linguists' attempts to define the concept of word-class in term of syntactic interchangeability; the Collins English Dictionary explains "part of speech" as: ***a class of words sharing important syntactic or semantic features; a group of words in a language that may occur in similar positions or fulfil similar functions in a sentence.*** For example, the previous sentence includes the word-sequences ***a class of*** and ***a group of*** ; this suggests ***class*** and ***group*** belong to the same word-class as they occur in similar contexts.

However, the three criteria traditionally used to determine Part of Speech can conflict: a word TYPE may fit more than one category, because individual TOKENS behave differently. Furthermore, the three criteria can be problematic to define as Machine Learnable features:

Semantic feature: noun = thing. Schoolchildren are often told that a noun is a "thing", a verb is an "action or state", an adjective is a "describer"; these are really semantic rather than syntactic classifications. A computer analysis would require knowledge of the features characterizing and differentiating "things", "actions" etc; these are hard to pin down.

Syntactic feature: noun can inflect sing v plural. An inflectional affix may be a clue to part-of-speech. However, to use this clue, a system requires knowledge of the internal morphology of words: whether inflections are marked by suffixes (eg English), prefixes (eg Japanese), word-internal variation (eg Arabic) or some complex combination of these (eg Maltese). At minimum, a system would need knowledge of contrasting pairs of words with/without each feature, and there may not be representative examples in a training corpus. Most Grammar Inference systems assume that words are atomic units to be joined into phrases and clauses, and that internal analysis of words is outside their scope.

Position/function: noun fits "a X of" This works for X when it appears in a pre-defined context pattern, but frequently words appear in new patterns not seen before. For Machine Learning, it may be possible to "relax" the patterns: instead of matching on exact words before and after, match on the word-classes before and after.

Clustering algorithms are not specific to Unsupervised Natural Language Learning: a range of generic clustering algorithms for Machine Learning can be found in the literature. These generic clustering systems require the user to formalise the problem in terms of a feature-space: every instance or object to be clustered must be characterised by a set of feature-values, so that instances with same or similar feature-values can be lumped together. Generally clustering systems assume each instance is independent; whereas when clustering words in a text, it may be helpful to allow the "contextual features" to either be words or be replaced by wordclass-labels as clustering proceeds. Every instance (word-type) must be characterised by a vector of feature-values (neighbour word-types and coocurrence frequencies). Instances with similar feature-vectors are lumped together or merged, so words are merged into word-classes. The words' feature-vectors are also merged; furthermore all other feature-vectors where merged words appear in the context must be updated. Feature-values (context-word types) must be updated after each iteration; this is not part of standard clustering.

3 Problems with clustering of word-types

These clustering algorithms assume a word-type can belong to only one class; for English at least, this turns out to be reasonable for function words (articles, prepositions, personal pronouns) but not a good assumption for open-class categories. Also, in general, statistical clustering requires many examples of each instance to be clustered this can be achieved for high-frequency word-types, but many types are sparsely-represented. Zipf's law of word-frequency distribution suggests that the great majority of word-types in a corpus occur rarely, and about half of all word-types occur only once.

This results in neoposy which passes a linguist's "looks good to me" evaluation for some small word-clusters corresponding to closed-class function-word categories (articles, prepositions, personal pronouns): the author can claim that at least some of the machine-learnt word groupings "look good" because they appear to correspond to linguistic intuitions about word-classes. However, statistical clustering does not classify the majority of word-types in the training corpus. Furthermore, the basic assumption that every word belongs to one and only one class does not allow existing word-clustering systems to cope adequately with words which linguists and lexicographers perceive as syntactically ambiguous. This is particularly problematic for isolating languages, that is, languages where words are generally not inflected for grammatical function and may serve more than one grammatical function; for example, in English or Malay many nouns can be used as verbs, and vice versa.

The root of the problem is the general assumption that the word-type is the atomic unit to be clustered, using the set of word-token contexts for a word-type as the feature-vector to use in measuring similarity between word-types, applying standard statistical clustering techniques. For example, (Atwell 1983) assumed that a word-type can be characterised by its set of word-types and contexts in a corpus, where the context is just the immediately preceding word: two word-types are merged into a joint word-class if the corresponding word-tokens in the training corpus show that similar sets of words tend to precede them. Subsequent researchers have tried varying clustering parameters such as the context window, the order of merging, and the similarity metric; but this does not allow a word to belong to more than one class.

4 Unification of word tokens

One answer may be to try clustering word **tokens** rather than word types. In the earlier example, we can say that the specific word-tokens *class* and *group* in the given sentence share similar contexts and hence share word-class. We need not generalise this to all other occurrences of *class* or *group* in a larger corpus, only to occurrences which share similar context. However, a statistical clustering algorithm needs a large number of examples to compare; it cannot apply to pairs of tokens, where N=1.

It may be possible to infer at least some limited grammatical word-classification if we focus on low-frequency words, and just look for words which share a grammatical context. A simple Prolog implementation assumes "relevant context" is just the preceding word:

```
?- neoposy([the,cat,sat,on,the,mat],Tagged).

Tagged = [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat, T2]]
```

Prolog variable Tagged is instantiated to a list of [word, Tag] pairs, where Tag has an arbitrary name generated by the program, letter T followed by an integer. These integers increment starting from T1, unless a word has a "context" seen earlier, in which case it repeats the earlier tag. In the example sentence "the cat sat on the mat", word "mat" has the same context (preceding word "the") as earlier word "cat", so it gets the T2 tag instead of a new tag T6.

We see that the two tokens *the* have distinct tags T1 and T5 since they have different contexts; but the token *mat* is assigned the same tag as token *cat* because they have the same context (preceding word-type). This also illustrates an interesting contrast with word-type clustering: word-type clustering works best with high-frequency words for which there are plenty of example tokens; whereas word-token clustering, if it can be achieved, offers a way to assign low-frequency words and even hapax legomena (words occurring only once in the whole corpus) to word-classes, as long as they appear in a context which can be recognised as characteristic of a known word-class. In effect we are clustering or grouping together word-contexts rather than the words themselves.

5 Relaxed unification of word-tokens.

This prototype demonstrator also illustrates some problems with classification based on tokens. If we no longer assume all tokens of one type belong to one class, do we allow as many classes as there are tokens?

Presumably not, as there would then be no point in classification; in fact it would be hard to justify even calling this classification. In the above example sentence there are 6 words ad 5 Tags: at least two words share a word-class, because they share a context. This means there are as many classes as there are "contexts"; in the above case, a context is the preceding word-TYPE; this implies we will get as many classes as there are word-types. A million-word corpus such as the Lancaster-Oslo/Bergen (LOB) corpus yields about 50,000 word-types, so our simple neoposy word-unification algorithm yields about 50,000 word-classes. This is a lot less than a million (one per token), but arguably still too many to be useful.

It is tempting to follow the analogy of the clustering algorithm in Figure 1, iterating the pattern-match, using classes instead of words in contexts; this "relaxation" should allow more token-contexts to be unified, reducing the number of word-classes. This is illustrated in our modified neoposy2 algorithm, applied to the sentence "the cat sat on the mat and went to sleep".

?- neoposy2 ([the,cat,sat,on,the,mat,and,went,to,sleep], Tagged).

Tagged= [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat,T2], [and,T6], [went,T7], [to,T8], [sleep,T9]];

Tagged= [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat,T2], [and,T3], [went,T7], [to,T8], [sleep,T9]];

Tagged= [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat,T2], [and,T3], [went,T4], [to,T8], [sleep,T9]];

Tagged= [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat,T2], [and,T3], [went,T4], [to,T5], [sleep,T9]];

Tagged= [[the,T1], [cat,T2], [sat,T3], [on,T4], [the,T5], [mat,T2], [and,T3], [went,T4], [to,T5], [sleep,T2]]

As before, we see that at first [cat,T2] and [mat,T2] share word-class T2 because they share the "context", preceding word-type. The neoposy2 algorithm has been modified to allow

unification on either preceding word-type or word-class. So, a second iteration produces a solution with [sat, T3] unified with [and, T3] because they share the preceding word-class T2; a third iteration unifies [on, T4] and [went, T4] after T3; a fourth iteration unifies [the, T5] and [to, T5] after T4; and a fifth iteration unifies [mat, T2] (already unified with [cat, T2]) and [sleep, T2] after T5. This is clearly contrary to linguistic intuition: it is evident that relaxation of unification or merging by allowing word-class context may be too powerful.

6 Conclusions and future research

It seems we can infer grammatical Parts of Speech for high-frequency words by statistical clustering, but this excludes the great majority of word-types in a training set. Clustering does not help with low-frequency word-types, or with the many word-types which can appear in more than one grammatical category. The alternative constraint-based approach of unification of word-token contexts offers a way to group together low-frequency word-types, and allows different tokens of one word-type to belong to different categories according to grammatical contexts they appear in. However, simple unification of word-token-contexts yields an implausibly large number of Part-of-Speech categories; we have attempted to merge more categories by "relaxing" matching context to allow unification of word-categories as well as word-tokens, but this results in spurious unifications.

Wholly unsupervised machine learning via either word-type clustering or word-token unification may be unachievable, but perhaps some hybrid of token-based and type-based learning, combined with limited sensible "learning hints" may be more manageable. (Atwell 2003) calls for a programme of further research to develop a Language Discovery Toolkit, to include a hybrid solution: use word-type statistical clustering for high-frequency words, mainly closed-class function words with only one part-of-speech; then word-token constraint-based unification for rare words and hapax legomena; and as a backup if and when this fails, allow "learning hints": "seed" with a limited PoS-lexicon provided by linguists.

Références

E Atwell, 1983. Constituent-Likelihood Grammar. ICAME Journal Vol.7

E Atwell, 2003. A new machine learning algorithm for Neoposy: coining new Parts of Speech. In D Archer, P Rayson, A Wilson, T McEnery (eds), Proceedings of the Corpus Linguistics 2003 conference, UCREL technical papers, volume 16 pp43-47, Lancaster University

S Finch, N Chater,1992. Bootstrapping Syntactic Categories Using Statistical Methods. in Proceedings of 1st SHOE Workshop, Tilburg University, The Netherlands

J Hughes, E Atwell, 1994. The automated evaluation of inferred word classifications. in A Cohn (ed), ECAI'94: Proceedings of the 11th European Conference on Artificial Intelligence", pp535-539, Chichester, John Wiley

G Knowles, Z Don, 2003. Tagging a corpus of Malay texts, and coping with 'syntactic drift'. In D Archer, P Rayson, A Wilson, T McEnery (eds), Proceedings of the Corpus Linguistics 2003 conference, UCREL technical papers, volume 16 pp422-428, Lancaster University