Building Part-of-speech Corpora through Histogram Hopping

Marc Vilain

The MITRE Corporation
206 Burlington Rd, Bedford MA 01730, USA
mbv@mitre.org

Abstract

This paper are concerned with lowering the cost of producing training resources for part-of-speech taggers. We focus primarily on the resource needs of unsupervised taggers, as these can be trained with simpler resources than their supervised counterparts. We introduce *histogram hopping*, a new approach for developing the central training resources of unsupervised taggers, and describe a simple annotation prototype that implements the approach. We then discuss the applicability of histogram hopping to the development of resources for supervised taggers. Finally, we report on a preliminary pilot study for French that validates this work.

Motivation and Preliminaries

Part-of-speech tagging is both valuable and ubiquitous. Many current approaches to language processing assume as *de rigueur* an initial labeling of lexemes with their part of speech, and indeed, trainable techniques to achieve this end are both numerous and mature.

Unfortunately, the range of languages for which part-of-speech taggers are available is narrow. The training techniques for most taggers are exceptionally data-hungry, and the creation of adequately-sized training corpora is therefore expensive, requiring extensive human effort to annotate upwards of a million words of text.

The present work is concerned with lowering the effort required to create useful training resources for part-of-speech tagging. Our hope is that this will lead to timely development of part-of-speech taggers in languages that have been hitherto poorly served by computational linguistic research.

Two key insights underlay this work. First, the presentation order by which annotators label the training corpus need not be linear. In this paper, we introduce an approach that focuses the annotation effort on all the instances of a lexeme, rather than on the linear order in which lexemes appear in the corpus. This results in huge reductions in the time required to annotate a corpus.

A second key insight is that there are high-utility resources for part-of-speech training besides fully-annotated corpora. Judiciously done, a partial annotation of a corpus may yield enough lexical information to support unsupervised training methods.

Data Requirements for P-O-S Taggers

Some years ago, Patrick Winston described the then-current blocks world task as the *E. Coli* of AI planning problems: its simplicity and transparent relevance made it the touchstone of much planning research. The same can be said of part-of-speech tagging for the area of corpus-based natural language processing. There is hardly a training regimen of currency today that hasn't been applied to the task of identifying parts of speech.

Among today's readily available taggers, we find strongly statistical systems based on maximum likelihood (Church, 1988), hidden Markov models (Brants, 2000) or maximum entropy modeling (Ratnaparkhi, 1996). Artificial neural networks have been applied in a number of ways, most notably by Schmid (1994). The most widely-disseminated approach of all, however, is most likely that of Brill's, which relies on transformational rule sequences (Brill 1992, 1993, Ramshaw & Marcus 1994, and others).

The Penn Treebank

Propelling all these efforts has been the availability of large-scale corpora tagged for part of speech. Key among these corpora is the widely-distributed Penn Treebank [Marcus et al ????]. The Treebank gathers three major sources: a revised version of the seminal Brown corpus, the ATIS spoken dialogues, and an extensive sampling of the Wall Street Journal. Together, these account for over a million words of hand-annotated training data.

To state the obvious: this is a very large data set. The long list of annotators who contributed to Treebank testifies to the effort required to craft such an artifact. Equally obvious is the hugely positive influence that Treebank and related corpora have exercised on computational linguistic research.

But large-scale corpora such as Treebank have also had some unintended effects. Indeed, the very size of these resources has afforded the CL community the luxury to experiment with learning techniques that perform extremely well, but demand huge amounts of training data. For part-of-speech tagging, these techniques have set exceptionally high performance standards, but have had the unintended effect of narrowing the scope of most P-O-S research to English.

The issue is the high cost of reproducing the research in other languages. By definition, the research cannot proceed without the foreign equivalent of a Penn Treebank. Unfortunately, individual research efforts typically lack the financial wherewithal to create such resources. The problem is made worse by the perception that part-of-speech tagging is an essentially solved problem. There is little motivation to researchers of creating non-English P-O-S corpora, if these are not likely to cause any fundamental rethinking or breakthrough in part-of-speech tagging. For this reason, where P-O-S-tagged corpora have been created outside of English, it has typically been through the auspices of national infrastructure efforts and consortia, such as the Chinese "PKU" corpus or the pan-European MULTEXT project.

Supervised and Unsupervised Approaches

A common denominator among most approaches to P-O-S tagging is the use of a supervised training regimen. This strategy requires individually labeled training exemplars, and is thus well matched to large corpora such as the Penn Treebank. Indeed, since supervised methods guide their decisions by statistical estimates gathered from these exemplars, there is a natural premium on having very large corpora to ensure statistical reliability.

For resource-poor languages, alternatives to supervised methods are likely to be more practical. In the particular case of part-of-speech tagging, a particularly promising approach is the unsupervised technique first investigated by ??? and later revised by Brill (1995).

What makes this technique unsupervised is that it does not require individually annotated training exemplars to estimate the space of possible tagging decisions. Instead, this decision space is modeled with a lexicon. For a given lexeme, the lexicon indicates those parts of speech that may legitimately be assigned to that lexeme. In French, for example, the space of decisions for an unambiguous lexeme like “maison” is the singleton “NN” (common noun) whereas an ambiguous case like “la” is assigned a range, *e.g.*, “DT PRP CL” (determiner, pronoun, clitic).

To disambiguate such cases, a part-of-speech tagger will require normative estimates of the relative likelihood of each case. In this unsupervised framework, these are modeled with the additional assistance of an un-annotated corpus. The probability of a given tag □ being assigned to an ambiguous lexeme $\square_0$ is estimated by considering the relative distribution in the corpus of those unambiguous lexemes $\square_1 \ldots \square_n$, for which the lexicon only assigns the same part of speech □.

Brill (1995) shows that this technique achieves near state-of-the-art performance with no reliance whatsoever on supervised training data. He goes on to show that what error term remains vis-à-vis supervised methods can be erased by follow-on supervised training on very small fully-annotated data sets (on the order of 20,000 words).

Lexica for Unsupervised Tagging

Our main concern here, however, is not so much with performance as with data preparation. To this extent, the key observation is that the data requirements for unsupervised P-O-S tagging differ considerably from those for supervised methods. Although a large corpus of text is still required for normative estimation, it need not be laboriously annotated. Instead, the labor-intensive part of the data preparation is the creation of the lexicon. This task, however, is a considerably smaller one. As opposed to a *token*-level Treebank, a lexicon is organized at the *type* level, and thus only requires one entry per lexeme, regardless of frequency of occurrence. For instance, for our sample training corpus drawn from the ACL-ECI distribution of French news stories, 20,295 lexicon entries cover the 189,119 non-numeric tokens in the corpus.

For this approach to succeed, however, a tagging lexicon $\mathcal{L}$ must be complete with respect to its lexical inventory. More precisely, for each lexeme □ □ $\mathcal{L}$, and for a given universe of discourse $\mathcal{U}$ (i.e., a corpus), if □/ $\square_i$ is a legitimate tagging in $\mathcal{U}$, then $\square_I$ must be among those tags $\square_1 \ldots \square_n$ assigned to □ by $\mathcal{L}$.

Counterintuitive as this might seem, the completeness of a lexicon relative to its lexical inventory is more important than the completeness of the lexical inventory overall. This is because unsupervised tagging regimens such as Brill’s operate by selecting from among the ambiguous choices available for a lexeme. If a relevant choice $\square_I$ is not available among the tags assigned by $\mathcal{L}$ (for a given □), this regimen provides no option for introducing $\square_I$ into its decision space. In contrast, lexemes that are outright out of vocabulary are obvious to identify, and can be treated by one of a number of back-off strategies: Brill (1995) used an ancillary supervised tagging pass.

Are Found Lexica an Option?

An interesting option under the unsupervised regimen might be to exploit found resources for the tagging lexicon. A particularly attractive proposition would be to make direct use of machine-readable dictionaries, *i.e.*, digital versions of desktop dictionaries that are intended for human use.

To assess the feasibility of this approach, we attempted to compile an English tagging lexicon from the *Collins English Dictionary*, for which an electronic version has been made available by the ACL. Several impediments arose to this approach.

(1) Incompleteness of lexical variants. A dictionary intended for human consumption will typically not list the morphological variants of a stem.

(2) Grammar-school tag sets. The traditional parts of speech in dictionaries (noun, adj, *etc*.) make fewer distinctions than typical computational tag sets.

(3) Coverage relative to corpora. Readily-available MRD’s may not be well-matched to the genre and vocabulary of training or application corpora.

To minimize the first two impediments, we designed a reverse stemmer that generated the morphological variants of *Collins* head stems. This generator also normalized the *Collins* part-of-speech tags from grammar school categories (*e.g., noun*) to their Penn Treebank equivalents (*NN, NNS, etc*). This greatly minimized these first two impediments, but at the cost of engaging in a significant exercise in computational morphology.

Nonetheless, our experience with this approach is that even our expanded version of *Collins* did not capture the full range of lexical usage in our corpora of interest, *i.e.*, it failed to meet our criterion for completeness with respect to its lexical inventory. The moral – once again – is that the best way to characterize actual language use is through analysis of a corpus. In this light, the major contribution of the present paper is a regimen for lexicon creation and P-O-S annotation that remains rooted in a corpus, but that also yields unsupervised tagging resources at minimal annotation cost.

Annotation through Histogram Hopping

The prevalent approach to annotating parts of speech is to present the annotator with a sequence of sentences. Sometimes, the sentences are pre-annotated by applying a preliminary tagger that is trained on whatever portion of the corpus has been annotated to date. Either way, the annotator proceeds through the text in standard reading order, considers each lexeme on its own, and either adds a tag de novo or corrects that which may have been pre-annotated.

In our own experience with annotating corpora for parts of speech,[1] we found this method to be most appropriate for

[1] We developed an 80,000-word P-O-S-tagged corpus of Spanish in support of our MET-1 entity tagger (Aberdeen *et al*, 1996).

```
              la charge de la gestion des (0)  ports fluviaux aux
                  dire sur le passage des (1)  bateaux .
           sur la notion de propriété des (2)  rives du Lez
     négociations , tant les intérèts des (3)  deux communes sont
                  o˘ l' on pense avoir des (4)  moyens de rétorsion
  fleuve pour faciliter la circulaton des (5)  bateaux et n'
  processus des photocopies ainsi que des (6)  pièces détachées pour
         qui entre dans le processus des (7)  photocopies ainsi que
```

Figure1: KWIC display for “des.”

unambiguous lexemes. For ambiguous lexemes, such as closed-class words that can be either articles, pronouns, and so forth, it is usually much easier to decide what tag to assign to a given instance when it is viewed in comparison to other cases. To this extent, the annotation approach taken here departs radically from the standard linear presentation strategy. We begin by histogramming the lexical inventory of the training corpus. The annotator is then presented with each lexeme in order of histogram appearance. We show all instances of the lexeme simultaneously through a keyword-in-context (KWIC) display. For obvious reasons, we call this approach histogram hopping.

Histogram Hopping for Lexicon Creation

A key advantage is that the KWIC display makes obvious to the annotator the different contexts in which a lexeme can be used. This greatly simplifies the determination of which parts of speech are valid for the lexeme, and which contexts are applicable for each potential part of speech. Figure 1, for instance, shows an annotation turn drawn from a French P-O-S annotation experiment (for the lexeme “des”). Along with the KWIC display, the annotator is presented with a very simple command-line interface into which to enter the valid parts of speech for this item. This particular word is ambiguous, and in this instance the annotator would enter the parts of speech DT and INDT. The former is for the use of “des” as an article meaning “some” (case 4) and the second is for the use of “des” as an elision for the preposition-determiner pair “de les” (al the other cases). For the purpose of lexicon creation, the annotator does not need to indicate which part of speech is to be assigned to which item in the KWIC display. Nonetheless, having the full range of alternatives available greatly eases the P-O-S selection task.

Clearly, this kind of interaction mode is well suited to creating lexica for unsupervised P-O-S tagging, as it provides a type-level interface for a type-level resource. Indeed, if one traverses the histogram to completion, the result will be to produce a lexicon that completely covers the training corpus. Note also that this approach addresses the issue of lexicon completeness relative to a lexeme, and does so in a much more natural way than linear presentation, as all corpus-anchored instances of a lexeme are considered at once.

In addition, the fact that the lexical histogram is traversed in order leads to a number of potential annotation speed improvements. In particular, relatively early stages of histogram traversal will typically nominate both high-frequency ambiguous words and high-frequency unambiguous words. This allows for preliminary tagger training to proceed even before the lexicon is fully created. For inflected languages in particular, one of the most useful products of preliminary training is a morphological guesser. By pre-tagging a lexeme with the results of morphological guessing, the annotator’s task is often further reduced to simply vetting the pre-tagged guesses.

We have experimented with two modes of morphological guessing: *winner-takes-all* and *all-are-winners*. In the former mode, we run a rule-sequence guesser similar to Brill’s (1992), which chooses a single best guess for a given lexeme. In our French experiments, we found that even small amounts of training data yielded a guesser that produced consistently rational part-of-speech estimates. The following random sample is typical: aside from “écartèlement,” the guessed parts of speech are legitimate for some use of the corresponding word (if not necessarily the most common use).

écarts	NNS
écartèlement	RB
échafaudages	NNS
échalote	NN
échangeant	VBG
échanger	VB
échangé	NN
échangée	VBN
échangés	NNS
échappaient	VBP
échappait	VBP

The *winner-take-all* strategy, however, is limited in its value to the purpose of lexicon creation, as what we care about is not so much obtaining a single – or even a best – part of speech, but eliciting all the valid ones. Towards this end, we have been experimenting with a morphological guesser that produces statistically valid alternatives. Although this is still work in progress, it appears to be a promising approach.

Histogram Hopping for Supervised Data

We have also experimented with histogram hopping as a means for creating token-level annotations, *i.e.*, the traditional kind of P-O-S corpora modeled on the Penn Treebank. The method proceeds as before, by traversing a histogram of lexemes, whose uses are visualized in a KWIC display. As before, at each annotation turn, the annotator indicates the parts of speech that are valid for a lexeme, and in addition links these to the appropriate cases in the KWIC display. In the example in Figure 1, the annotator would link the DT label to case 4 (“des moyens”) and the INDT label to all the others.

In our command-line interface the user would record this by typing “`IN 4 INDT`” – this causes “des” in case 4 to be tagged as a proposition, and all other cases to default to the INDT tag. Admittedly, this is a somewhat clumsy input method. The real advantage of this method, however, is that the KWIC display allows the annotator to compare any given use of a lexeme to its full range of linguistic behaviors. This has the same advantage of clarifying use at the token level as it does at the type level. In this approach, traversing the histogram in order eventually yields a fully-tagged corpus.

Effectiveness of Histogram Hopping

The primary reason for the effectiveness of histogram hopping is that it allows the annotator to form a differential diagnosis for a lexeme's valid parts of speech. By this term (borrowed from medicine), we mean the establishment in the annotator's mind of the full range of syntactic roles for a lexeme, along with justifications and contexts for each part-of-speech assignment. As noted, this differential diagnosis is easy to achieve with the KWIC display, but is elusive with linear presentation order.

When the annotator is engaged in a lexicon-building task, little more is needed in handling a lexeme than reporting this differential diagnosis. This allows the annotator to move rapidly through recurring words in a corpus, even words with significant degrees of part-of-speech ambiguity. But histogram hopping can also be a faster method for the creation of fully-annotated corpora than linear presentation modes. The reason for this appears mostly to have to do with the annotation of ambiguous words, especially function words. Even when one is only annotating one instance of this class of word, it is necessary to some degree to mentally recall the entire range of what is permissible for the word. This again is aided by the KWIC display.

Annotation Experiments

To validate the ideas presented here, we performed some annotation experiments on a corpus of French news stories. We used the ACL's *European Coding Initiative* distribution, and focused *Le Monde* articles from 1989 and 1990 (section fre01a01). We divided the complete data set into segments of various sizes, and annotated these by histogram hopping. We used the technique to build both type-level lexica and corpora that were fully annotated at the token level.

We targeted the annotations to a tag set that is only a minimal variant from that used in the Penn Treebank. We added a small number of additional tags to cover verbal inflections that are productive in French, but absent in English, e.g., VBF (future), VBC (conditional), and the like. We also added the INDT category for elided forms such as "des" (= "de les"), and various categories to cover clitics. We did not choose to expand the tag set to cover inflectional markers such as person, number, or gender, as was done by Sánchez León, (1994) for Spanish. In our experience, this approach does not add much predictive power to the determination of part of speech, and indeed, by creating large numbers of sparsely populated categories, tends to weaken it.

The implementation used in these experiments had the anachronistic line-oriented interface described earlier.

Empirical Measurements

We found that even with a clumsy non-graphical interface, significantly higher tagging rates could be obtained with KWIC presentation than with linear presentation. For the full annotation task, we measured tagging rates on the order of 1500 words/hr using histogram hopping and the KWIC display. This compares favorably to the 500-750 words/hr that we had come to expect from earlier efforts with linear tagging. For the lexicon-building task, we were able to reach rates of 2000-2500 words/hr. These assessments were all conducted prior to the integration of morphological guessing.

Conclusions and Future Work

We are encouraged by these results, although in our opinion they are still preliminary. We would certainly welcome wider use of this approach, especially should it involve a larger number of annotators (our study had a sole French speaker).

Among extensions of the present method, we would like to incorporate our ongoing work many-valued morphological guessing. Our expectation is that this would increase annotation throughput even further, especially for lexicon building.

We also need to assess more fully the quality of taggers built using our unsupervised resources. Along those lines, one interesting option is to limit lexicon creation to those items that are not singletons in the corpus. Singletons, as it happens can not participate in the derivation of unsupervised tagging models, but account for at least half the tokens of a given corpus. By ignoring the singleton tail of our lexical histograms, we have measured annotation rates of almost 8000 words/hr. Our hope is that this tail-limited approach will nonetheless yield P-O-S tagging performance that is comparable to full histogram traversal. If so, this could prove a very practical way to bring underserved languages into the computational fold.

References

Aberdeen, J, Burger, J, Day, D, Hirschman, L, Robinson, P, Vilain, M (1996) Description of the Alembic system as used in MET. In *Prcdgs. of the TIPSTER Workshop, Phase II*. San Francisco: Morgan Kaufmann.

Brants, T. (2000). Tnt – a statistical part-of-speech tagger. In *Prcdgs. of the 6th Applied Natural Language Processing Conference* (ANLP '00), Seattle, WA.

Brill, E. (1992) A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Brill, E. (1993) *A Corpus-Based Approach to Language Learning*. Doctoral Dissertation, Department of Computer and Information Science, Univ. of Pennsylvania.

Brill, E. (1995). Unsupervised learning of disambiguation rules for part-of-speech tagging. In *Prcdgs. of the 3rd Wkshp on Very large Corpora*, Cambridge, MA.

Ramshaw, L.A., and Marcus, M.P. (1994). Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging. In *Prcdgs. of the ACL Workshop on Combining Statistical and Symbolic Approaches to Language*, Las Cruces, NM.

Ratnaparkhi. A. (1996). A maximum entropy part-of-speech tagger. In *Prcdgs, of Empirical Methods in Natural Language Processing*, Philadelphia, PA

Sánchez León, F. (1994). Spanish tagset for the crater project. Manuscript from *The Computation and Language E-Print Archive* (http://xxx.lanl.gov/cmp-lg/).

Schmid, H. (1994).. Part-of-speech tagging with neural networks. In *Prcdgs. of the 15th International Conference on Computational Linguistics* (COLING94).