HOW DO WE COUNT?
THE PROBLEM OF TAGGING PHRASAL VERBS IN PARTS

Nava A. Shaked

The Graduate School and University Center
The City University of New York
33 West 42nd Street. New York, NY 10036
nava@nynexst.com

ABSTRACT

This paper examines the current performance of the stochastic tagger PARTS (Church 88) in handling phrasal verbs, describes a problem that arises from the statistical model used, and suggests a way to improve the tagger's performance. The solution involves a change in the definition of what counts as a word for the purpose of tagging phrasal verbs.

1. INTRODUCTION

Statistical taggers are commonly used to preprocess natural language. Operations like parsing, information retrieval, machine translation, and so on, are facilitated by having as input a text tagged with a part of speech label for each lexical item. In order to be useful, a tagger must be accurate as well as efficient. The claim among researchers advocating the use of statistics for NLP (e.g. Marcus *et al.* 92) is that taggers are routinely correct about 95% of the time. The 5% error rate is not perceived as a problem mainly because human taggers disagree or make mistakes at approximately the same rate. On the other hand, even a 5% error rate can cause a much higher rate of mistakes later in processing if the mistake falls on a key element that is crucial to the correct analysis of the whole sentence. One example is the phrasal verb construction (e.g. *gun down, back off*). An error in tagging this two element sequence will cause the analysis of the entire sentence to be faulty. An analysis of the errors made by the stochastic tagger PARTS (Church 88) reveals that phrasal verbs do indeed constitute a problem for the model.

2. PHRASAL VERBS

The basic assumption underlying the stochastic process is the notion of independence. Words are defined as units separated by spaces and then undergo statistical approximations. As a result the elements of a phrasal verb are treated as two individual words, each with its own lexical probability (i.e. the probability of observing part of speech i given word j). An interesting pattern emerges when we examine the errors involving phrasal verbs. A phrasal verb such as *sum up* will be tagged by PARTS as noun + preposition instead of verb + particle. This error influences the tagging of other words in the sentence as well. One typical error is found in infinitive constructions, where a phrase like *to gun down* is tagged as INTO NOUN IN (a prepositional 'to' followed by a noun followed by another preposition). Words like *gun, back, and sum*, in isolation, have a very high probability of being nouns as opposed to verbs, which results in the misclassification described above. However, when these words are followed by a particle, they are usually verbs, and in the infinitive construction, always verbs.

2.1. THE HYPOTHESIS

The error appears to follow from the operation of the stochastic process itself. In a trigram model the probability of each word is calculated by taking into consideration two elements: the lexical probability (probability of the word bearing a certain tag) and the contextual probability (probability of a word bearing a certain tag given two previous parts of speech). As a result, if an element has a very high lexical probability of being a noun (*gun* is a noun in 99 out of 102 occurrences in the Brown Corpus), it will not only influence but will actually override the contextual probability, which might suggest a different assignment. In the case of *to gun down* the ambiguity of *to* is enhanced by the ambiguity of *gun*, and a mistake in tagging *gun* will automatically lead to an incorrect tagging of *to* as a preposition.

It follows that the tagger should perform poorly on

phrasal verbs in those cases where the ambiguous element occurs much more frequenty as a noun (or any other element that is not a verb).The tagger will experience fewer problems handling this construction when the ambiguous element is a verb in the vast majority of instances. If this is true, the model should be changed to take into consideration the dependency between the verb and the particle in order to optimize the performance of the tagger.

3. THE EXPERIMENT

3.1. DATA

The first step in testing this hypothesis was to evaluate the current performance of PARTS in handling the phrasal verb construction. To do this a set of 94 pairs of Verb+Particle/Preposition was chosen to represent a range of dominant frequencies from overwhelmingly noun to overwhelmingly verb. 20 example sentences were randomly selected for each pair using an on-line corpus called MODERN, which is a collection of several corpora (Brown, WSJ, AP88-92, HANSE, HAROW, WAVER, DOE, NSF, TREEBANK, and DISS) totaling more than 400 million words. These sentences were first tagged manually to provide a baseline and then tagged automatically using PARTS. The a priori option of assuming only a verbal tag for all the pairs in question was also explored in order to test if this simple solution will be appropriate in all cases. The accuracy of the 3 tagging approaches was evaluated.

3.2. RESULTS

Table 2 presents a sample of the pairs examined in the first column, PARTS performance for each pair in the second, and the results of assuming a verbal tag in the third. (The "choice" column is explained below.)

The average performance of PARTS for this task is 89%, which is lower than the general average performance of the tagger as claimed in Church 88. Yet we notice that simply assigning a verbal tag to all pairs actually degrades performance because in some cases the content word is almost always a noun rather than a verb. For example, a phrasal verb like *box in* generally appears with an intervening object (*to box something in*), and thus when *box* and *in* are adjacent (except for those rare cases involving heavy NP shift) *box* is a noun.

Thus we see that there is a need to distinguish between the cases where the two element sequence should be considered as one word for the purpose of assigning the Lexical Probability (i.e.,phrasal verb) and cases where we have a Noun + Preposition combination where PARTS' analyses will be preferred. The "choice" in

PHRASAL VERB	IMP.	FREQ. DIST. (BROWN)	
date-from	1.00	date	NN/98 VB/6
flesh-out	0.59	flesh	NN/53 VB/1
bottle-up	0.55	bottle	NN/77 VB/1
hand-over	0.35	hand	NN/411 VB/8
narrow-down	0.23	narrow	JJ/61 NN/1 VB/1
close-down	0.22	close	JJ/81 NN/16 QL/1 RB/95 VB/40
drive-off	0.22	drive	NN/49 VB/46
cry-out	0.21	cry	NN/31 VB/19
average-out	0.20	average	JJ/64 NN/60 VB/6
head-for	0.18	head	JJ/4 NN/404 VB/14
end-up	0.16	end	NN/359 VB/41
account-for	0.15	account	NN/89 VB/28
double-up	0.14	double	JJ/37 NN/11 RB/4 VB/6
back-off	0.13	back	JJ/27 NN/177 RB/720 VB/26
cool-off	0.13	cool	JJ/49 NN/3 RB/1 VB/8
clear-out	0.12	clear	JJ/197 NN/1 RB/10 VB/15
calm-down	0.10	calm	JJ/22 NN/8 VB/7

Table 1: 10% or more improvement for elements of non verbal frequency.

Table 2 shows that allowing a choice between PARTS' analysis and one verbal tag to the phrase by taking the higher performance score, improves the performance of PARTS from 89% to 96% for this task, and reduces the errors in other constructions involving phrasal verbs. When is this alternative needed? In the cases where PARTS had 10% or more errors, most of the verbs occur much more often as nouns or adjectives. This confirms my hypothesis that PARTS will have a problem solving the N/V ambiguity in cases where the lexical probability of the word points to a noun. These are the very cases that should be treated as one unit in the system. The lexical probability should be assigned to the pair as a whole rather than considering the two elements separately. Table 1 lists the cases where tagging improves 10% or more when PARTS is given the additional choice of assigning a verbal tag to the whole expression. Frequency distributions of these tokens in the Brown Corpus are presented as well, which reflect why statistical probabilities err in these cases. In order to tag these expressions correctly, we will have to capture additional information about the pair which is not available from the PARTS statistical model.

pairs	parts	all verb	choice
account-for	0.84	1	1
aim-at	0.90	0.30	0.90
average-out	0.7	0.9	0.9
back-off	0.86	1	1
balance-out	0.92	0.84	0.92
bargain-for	0.87	0.58	0.87
block-in	0.97	0.02	0.97
book-in	1	0	1
bottle-up	0.36	0.90	0.90
bottom-out	0.8	0.85	0.85
box-in	1	0.02	1
break-away	1	1	1
call-back	0.96	0.84	0.96
calm-down	0.85	0.95	0.95
care-for	0.9	0.48	0.93
cash-in	0.95	0.25	0.95
change-into	0.85	0.89	0.89
check-in	0.96	0.48	0.96
clear-out	0.87	1	1
close-down	0.77	1	1
contract-in	1	0.02	1
cool-off	0.86	1	1
credit-with	1	0	1
cry-out	0.79	1	1
date-from	0	1	1
deal-with	0.96	0.92	0.96
demand-of	1	0.04	1
double-up	0.80	0.95	0.95
end-up	0.83	1	1
fall-in	0.92	0.29	0.92
feel-for	0.93	0.33	0.93
flesh-out	0.41	1	1
flow-from	0.94	0.42	0.94
fool-around	0.91	1	1
force-upon	0.84	0.61	0.84
gun-down	0.60	0.62	0.62
hand-over	0.65	1	1
head-for	0.63	0.81	0.81
heat-up	0.94	1	1
hold-down	0.92	1	1
lead-on	1	0.07	1
let-down	0.57	0.57	0.57
live-for	0.91	1	1
move-in	0.96	0.60	0.96
narrow-down	0.77	1	1
part-with	0.79	0.43	0.79
phone-in	0.91	0.12	0.91
TOTAL AVERAGE	0.89	0.79	0.96

Table 2: A Sample of Performance Evaluation

4. CONCLUSION: LINGUISTIC INTUITIONS

This paper shows that for some cases of phrasal verbs it is not enough to rely on lexical probability alone: We must take into consideration the dependency between the verb and the particle in order to improve the performance of the tagger.The relationship between verbs and particles is deeply rooted in Linguistics. Smith (1943) introduced the term *phrasal verb*, arguing that it should be regarded as a type of idiom because the elements behave as a unit. He claimed that phrasal verbs express a single concept that often has a one word counterpart in other languages, yet does not always have compositional meaning. Some particles are syntactically more adverbial in nature and some more prepositional, but it is generally agreed that the phrasal verb constitutes a kind of integral functional unit. Perhaps linguistic knowledge can help solve the tagging problem described here and force a redefinition of the boundaries of phrasal verbs. For now we can redefine the word boundaries for the problematic cases that PARTS doesn't handle well. Future research should concentrate on the linguistic characteristics of this problematic construction to determine if there are other cases where the current assumption that one word equals one unit interferes with successful processing.

5. ACKNOWLEDGEMENT

6. REFERENCES

K. W. Church. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proc. Conf. on Applied Natural Language Processing*, 136–143, 1988.

K. W. Church, & R. Mercer. Introduction to the Special Issue on Computational Linguistics Using Large Corpora. To appear in *Computational Linguistics*, 1993.

C. Le raux. On The Interface of Morphology & Syntax. Evidence from Verb-Particle Combinations in African. SPIL 18. November 1988. MA Thesis.

M. Marcus, B. Santorini & D. Magerman. First steps towards an annotated database of American English. Dept. of Computer and Information Science, University of Pennsylvania, 1992. MS.

L. P. Smith. *Words & Idioms: Studies in The English Language*. 5th ed. London, 1943.