Guaranteed Pre–Tagging for the Brill Tagger

Saif Mohammad and Ted Pedersen

University of Minnesota, Duluth, MN 55812 USA
moha0149@d.umn.edu, tpederse@umn.edu
http://www.d.umn.edu/{~moha0149,~tpederse}

Abstract. This paper describes and evaluates a simple modification to the Brill Part–of–Speech Tagger. In its standard distribution the Brill Tagger allows manual assignment of a part–of–speech tag to a word prior to tagging. However, it may change it to another tag during processing. We suggest a change that guarantees that the *pre–tag* remains unchanged and ensures that it is used throughout the tagging process. Our method of guaranteed pre-tagging is appropriate when the tag of a word is known for certain, and is intended to help improve the accuracy of tagging by providing a reliable anchor or seed around which to tag.

1 Introduction

Part–of–speech tagging is a prerequisite task for many natural language processing applications, among them parsing, word sense disambiguation, machine translation, etc. The Brill Tagger (c.f., [1], [2], [3], [5]) is one of the most widely used tools for assigning parts–of–speech to words. It is a hybrid of machine learning and statistical methods that is based on transformation based learning.

The Brill Tagger has several virtues that we feel recommend it above other taggers. First, the source code is distributed. This is rare, as most other part–of–speech taggers are only distributed in executable format. Second, the simplicity of the transformation based learning approach makes it possible for us to both understand and modify the process to meet our needs. Finally, the tagger is quite accurate, and consistently achieves overall accuracy of at least 95%.

Part–of–speech taggers normally assume that the sentence it is processing is completely untagged. However, if the tags for some of the words in a text are known prior to tagging, then it would be desirable to incorporate that information in such a way that the tagger can use it and hopefully improve its accuracy. The act of manually assigning tags to selected words in a text prior to tagging will be referred to as *pre–tagging*. The affected words are said to be *pre–tagged* and the actual tags assigned to them are known as *pre–tags*.

Pre–tagging is intended to take advantage of the locality of part–of–speech tags. The tag for any word is generally determined by one or two immediate neighbors. Pre–tagging can be thought of as the process of manually priming or seeding the tagging process with reliable prior information. If the part–of–speech of a word can be manually assigned prior to tagging, then the surrounding tags

may be tagged more accurately as a result of this additional information. Pre–tagging is possible because of this locality property; assigning a tag to a word does not affect the tagging of the entire sentence and can be thought of as introducing a very localized constraint on the tagging process.

We have developed a pre–tagging technique for the Brill Tagger that allows words to be assigned pre–tags, and then guarantees that the pre–tag will remain unchanged throughout tagging and will affect the tagging of its neighbors. Thus, if we are certain that a word should have a particular part–of–speech tag, we can provide that information and be assured that the pre–tag will remain in the final output and will have been used to determine the tags of neighboring words. While the Brill Tagger provides a form of pre–tagging, it gives no assurances that the pre–tag will actually be used in the tagging process. Thus our approach is distinct in that it guarantees that prior information about part–of–speech tags will be incorporated into the tagging process.

This paper continues with a short introduction to the Brill tagger and its existing form of pre–tagging. It goes on to introduce our guaranteed form, and then discusses an evaluation of the impact of this new form of tagging.

2 The Brill Tagger

The Brill Tagger proceeds in two phases. In the first phase, the Initial State Tagger assigns each word its most likely tag based on information it finds in a lexicon. In the second phase, a series of contextual rules are applied by the Final State Tagger to determine which of those initial tags should be transformed into other tags. Our experiments and modifications are based on the August 1994 version of the Brill Tagger, known as RULE_BASED_TAGGER_1.14.

2.1 Initial State Tagger

The first phase of tagging is performed by the Initial State Tagger, which simply assigns the most likely tag to each word it encounters. The most likely tag for a word is given in the lexicon. If the word is not in the lexicon it is considered as an unknown word and is tagged as a proper noun (NNP) if it is capitalized and as a noun (NN) if it is not.

The lexicon (LEXICON.BROWN.AND.WSJ) we use in our experiments is from the standard distribution of the Brill Tagger (1.14) and was derived from the Penn TreeBank tagging of the Wall Street Journal and the Brown Corpus. This lexicon consists of almost 94,000 words and provides their most likely part–of–speech, based on frequency information taken from the aforementioned corpora. It also lists the other parts–of–speech with which each word can be used. Note that there are separate entries for the different morphological and capitalized forms of a word. The lexicon shown in Table 1 follows the standard form of a Brill Tagger lexicon and is referred to in examples throughout this paper.

Table 1. Example Lexicon

Word	Most Frequent Tag	Other Possible Tags	
brown	JJ	NN VB	...(L1)
chair	VB	NN	...(L2)
evening	NN	JJ	...(L3)
in	IN	FW NN	...(L4)
meeting	NN	VB	...(L5)
pretty	RB	JJ	...(L6)
sit	VB	FW VB	...(L7)
the	DT	NNP PDT	...(L8)
this	DT	PDT	...(L9)
time	NN	VB	...(L10)
will	MD	VBP NN	...(L11)

Entry L1 tells us that *brown* is usually used as an adjective (JJ) but may also be used as a verb (VB) or a noun (NN). L2 shows that *chair* is most often a verb (VB) but can also be a noun (NN). Note that the order of the other possible tags is not significant, it simply indicates that the word was used in these parts–of–speech in the corpus the lexicon was induced from.

The tags assigned by the Initial State Tagger may be transformed following a set of lexical rules based on suffixes, infixes, and prefixes of the word. The tagger comes with predefined lexical rule files that have been learned from the same corpora used to learn the lexicon. The lexical rules file affects only the unknown words and as such is not directly involved in or affected by pre–tagging, so we will not discuss it any further.

2.2 Final State Tagger

The next stage of tagging determines if any of the tags assigned by the Initial State Tagger should be changed based on a set of contextual rules. These rules specify that the current tag of a word may be transformed into another tag based on its context. This context usually consists of one, two or three words (and their part–of–speech tags) to the left and right of the tagged word.

For our experiments we use a contextual rule file (CONTEXTUALRULEFILE.WSJ) provided in the standard 1.14 distribution of the Brill Tagger. This consists of 284 rules derived from the Penn TreeBank tagging of the Wall Street Journal. The examples in this paper rely on just a few contextual rules, and those are shown in Table 2.

The rule C1 indicates that if a word is tagged as a noun (NN) then its tag should be changed to verb (VB) if the part–of–speech of the next word is a determiner (DT). Rule C2 says that the tag of a word should be changed from adverb (RB) to adjective (JJ), if the next word is tagged as a noun (NN). Rule C3 says that an adjective (JJ) should be changed to a noun (NN) if the next word has a verb tag (VB). Finally, rule C4 is lexicalized, and says that a word

Table 2. Example Contextual Rules

Current Tag	New Tag	When	
NN	VB	NEXTTAG DT	...(C1)
RB	JJ	NEXTTAG NN	...(C2)
JJ	NN	NEXTTAG VB	...(C3)
NN	JJ	NEXTWD meeting	...(C4)

tagged as a noun (NN) should be changed to an adjective (JJ) if the next word is *meeting*.

3 Standard Pre-Tagging with the Brill Tagger

Pre–tagging is the act of manually assigning a part–of–speech tag to words in a text prior to that text being automatically tagged with the Brill Tagger. The following example will illustrate the general concept of pre–tagging, and show the limitations of pre–tagging as provided in the standard distribution of the Brill Tagger.

Suppose that it is critical to an application to know that *chair* is being used as a noun (NN) in the following context. We could apply a pre–tag as follows:

```
Mona will sit in the pretty chair//NN this time          (1)
```

The Initial State Tagger will assign the most likely tag to each word, except for *chair* which is pre–tagged as a noun (NN) and for *Mona* which is not in the lexicon but is tagged as a proper noun (NNP) since it is capitalized. The results of this initial tagging are as follows:

```
Mona/NNP will/MD sit/VB in/IN the/DT
pretty/RB chair//NN this/DT time/NN                      (2)
```

The Final State Tagger will look for contextual rules to apply, and will transform tags accordingly. It treats a pre–tagged word like any other, so the pre–tag may be changed during the course of tagging. While the standard distribution of the Brill Tagger allows a user to specify a different initial tag for a word via pre–tagging, it does not guarantee that this be used throughout tagging. Given the input above, the Brill Tagger will produce the following tagging:

```
Mona/NNP will/MD sit/VB in/IN the/DT
pretty/RB chair//VB this/DT time/NN                      (3)
```

Note that the tag of *chair* has been changed to a verb (VB). While *chair* can be a verb, as in *Mona will chair the meeting this time*, in this case it is not. In particular, *chair* was pre–tagged as a noun (NN) but this was overridden by the

Final State Tagger which mis–tagged it as a verb (VB). This change occurred due to the contextual rule C1 shown in Table 2. This rule says that a word that is tagged as a noun (NN) should be changed to a verb (VB) when it is followed by a determiner (DT). This error is compounded, since *pretty* was tagged by the Initial State Tagger as an adverb (RB), due to lexicon entry L6 in Table 1. Since *chair* is considered a verb, the initial tagging of *pretty* as an adverb (RB) will be allowed to stand.

In this example the erroneous tagging of *chair* causes the tag of *pretty* to remain unchanged. We can observe the opposite behavior with a simple change in the example. Suppose that Mona is sitting in a *brown* chair. We could again pre–tag *chair* to indicate that it is a noun:

```
Mona will sit in the brown chair//NN this time        (4)
```

The Initial State Tagger will assign the same tags as it did in Sentence 2, except that *brown* will be tagged an adjective (JJ) since that is its most likely tag, as shown in L1 in Table 1.

```
Mona/NNP will/MD sit/VB in/IN the/DT
brown/JJ chair//NN this/DT time/NN                    (5)
```

From this the Final State Tagger will produce the following:

```
Mona/NNP will/MD sit/VB in/IN the/DT
brown/NN chair//VB this/DT time/NN                    (6)
```

Here the pre–tag of *chair* is changed to a verb (VB) due to contextual rule C1. This triggers a change in the tag of *brown* due to rule C3, which says that an adjective (JJ) should be changed to a noun (NN) when it is followed by a verb (VB). Thus, the improper changing of the pre–tag of *chair* has resulted in the incorrect tag being applied to *brown* as well.

The standard distribution of the Brill Tagger provides relatively weak pre–tagging that simply overrides the Initial State Tagger. However, those pre–tags can be altered by the Final State Tagger, and such changes can trigger other transformations in the tags of neighboring words.

4 Guaranteed Pre–Tagging

Our objective is to guarantee that a manually assigned pre–tag be respected (and left unchanged) by both the Initial State Tagger and the Final State Tagger. We believe that there are cases when the pre–tag should be considered absolute and affect the outcome of the tags on surrounding words (and not vice–versa). If a contextual rule changes a pre–tag from what is known to be correct to something that is not, then the surrounding words may also be incorrectly tagged via applications of contextual rules that rely upon the improperly changed pre–tag.

To achieve guaranteed pre–tagging, we have made a simple change to the Brill tagger that prevents it from applying contextual rules that result in changes to a pre–tagged word. However, we still allow contextual rules to change surrounding tags based on the pre–tag. So while a pre–tag may not be changed, the tags of surrounding words may be changed based on that pre–tag.

Let's return to the examples of Sentences 1 and 4. In each a noun (NN) pre–tag was assigned to *chair* prior to tagging, but it was overridden. As a result *chair* was improperly tagged as a verb (VB) and this had an impact on the tagging of *pretty* and *brown*.

With guaranteed pre–tagging, the final output of the Brill Tagger for Sentence 1 is as follows:

```
Mona/NNP will/MD sit/VB in/IN the/DT

pretty/JJ chair//NN this/DT time/NN
```
(7)

Note that the pre–tag of *chair* remains unchanged. The contextual rule C1 is not applied due to our prohibition against changing pre–tags. Since *chair* remains a noun, contextual rule C2 changes *pretty* from having an adverb (RB) tag to having an adjective (JJ) tag.

In the case of Sentence 4, the output of the Brill Tagger is:

```
Mona/NNP will/MD sit/VB in/IN the/DT

brown/JJ chair//NN this/DT time/NN
```
(8)

Note that the pre–tag of *chair* has not been changed, and in fact no contextual rules have been triggered. All of the other words in the sentence retain the tags as assigned by the Initial State Tagger.

These simple examples show how guaranteed pre–tagging can affect the outcome of the Brill Tagger. Next, we describe an extensive experiment that we carried out in assessing the impact of pre–tagging in part–of–speech tagging text to be used in a series of word sense disambiguation experiments.

5 Impact of Guaranteed Pre–Tagging on Senseval-2 data

We evaluated the effect of guaranteed pre–tagging on a large corpus of data that we part–of–speech tagged. It was our experience with this data that actually motivated the development of the guaranteed pre–tagging approach.

5.1 Experiment

The English lexical sample data for the SENSEVAL-2 word sense disambiguation exercise includes 4,328 test instances and 8,611 training instances [4]. Each instance consists of a few sentences where a single target word within each instance is manually assigned a sense–tag that indicates its meaning in that particular context. There are 73 different nouns, verbs, and adjectives that are sense–tagged

and serve as target words. This data is typically used for corpus–based supervised learning experiments where a model of disambiguation is learned from the training instances and then evaluated on the test instances.

We part–of–speech tagged this data with the Brill Tagger in preparation for some word sense disambiguation experiments. This tagging was done with the posSenseval package, now available from the authors. The focus of the word sense experiment was on the utility of part–of–speech tags of words near the target word as features for disambiguation, so we were particularly concerned that the tagging be as accurate as possible. Since the crude part–of–speech of the target word is known (noun, verb, or adjective) we decided it would be worthwhile to manually tag all of the target words with their appropriate part–of–speech tag, so as to possibly improve the tagging of nearby words.

5.2 Results

The pre–tagging feature of the original Brill Tagger was used to specify the appropriate pre–tags of the target words. An analysis of the tagging results surprised us. Of the 4,328 target words in the test instances assigned pre–tags, 576 were changed. Of those, 388 were minor changes within a single part–of–speech (e.g., from a past tense verb to a present tense) and 188 tags had been changed to completely different parts–of–speech (e.g., from a verb to a noun). We call the latter *radical changes* since they pose a greater concern. It seems likely that the surrounding tags have a reasonable chance of being mis–tagged as a result of radical errors. Of the 8,611 target words in the training data that were pre–tagged, 1,020 of those were mis–tagged, with 291 radical errors and 729 minor errors. Since we were certain of the pre–tags we assigned, and since we were quite concerned about the negative impact of radical errors in the tagging of target words, we developed the guaranteed approach to pre–tagging described here.

The guaranteed pre–tagging prevented radical errors and ensured that target words retained their pre–tags. We noted that of the 291 sentences in the training data where radical errors had previously existed, 36 sentences now had a change of a neighboring tag due to the correctly applied pre–tag. In the 188 sentences from the test data where radical errors had occurred, 18 sentences had a change in a neighboring tag due to an erroneous change of a pre–tag.

5.3 Discussion

At first the number of changes in the neighboring tags struck us as rather small. However, upon reflection they appear reasonable, and we shall explain how we arrive at that conclusion.

There are approximately 529,000 tokens in the test data, and of those only 25,000 are changed from their initial state taggings via contextual rules. In the training data there are 1,059,000 tokens, where 48,000 are changed from their initial state taggings via contextual rules. Thus, in both cases only about 5% of the assigned tags are something other than what the initial state tagger decided.

Guaranteed pre–tagging corrected the 188 radical errors in the test data and the 291 radical errors in the training data. Since most contextual rules affect only one word to the left or one word to the right of the target word, we would expect that contextual rules might change the tags of adjacent neighbors of the target words about 5% of the time. Based on this rather loose analysis we would expect that (188*2)*.05 = 19 neighboring tags should change in the test data and (291*2)*.05 = 29 should change in the training data. It turns out that these estimates are not far off, as the actual number of changes were 18 and 36.

In fact the analysis can be made a bit more precise, by noting that contextual rules can be divided into those that are triggered by content words, e.g., nouns, verbs, adjectives, and adverbs, and those that are triggered by function words. In the test data 14,100 tokens were changed based on transformations triggered by the current tag being a function word, and 10,300 were based on it being a content word. In the training data, 27,800 were triggered by the current tag being a function word and 20,500 were based on it being a content word.

This distinction is relevant since we know that the target words in the SENSEVAL-2 data are content words, and these are the only words that have been pre–tagged. We can estimate the probability that a target word will trigger a contextual rule by determining the overall probability in the test and training data that a contextual rule will be triggered, given that the token under consideration is a content word. The number of content tokens in the test data is 273,000 and the number in the training data is 546,000. Based on the counts of the number of contextual rules triggered by content words already provided, we can determine that the expected probability of a contextual rule triggering when the given token is a content word is about 4%. Thus, the expected number of changes that we computed above can be refined slightly, (188*2)*.04 = 15 and (291*2)*.04 = 23. These are still reasonably close to the observed number of changes (18 and 36). This suggests that pre–tagging is having an impact on the assignment of tags, and that the rate of change to neighboring tags is consistent with the rates of change of 4% and 5% that we have derived above.

We are uncertain whether we should expect 95% of the tokens to retain their initial state tags in general. We suspect this figure would be lower if the SENSEVAL-2 data were more like the Wall Street Journal and Brown Corpus from which the contextual rule file was learned. However, the SENSEVAL-2 data is from varied sources and is much noisier than the TreeBank and Brown Corpus data.

6 An Anomaly in Lexicalized Contextual Rules

During the testing of guaranteed pre–tagging we noticed somewhat unusual behavior in the Brill Tagger. If a contextual rule is lexicalized with a word that has been pre–tagged, then that rule will not be applied under any circumstances. A lexicalized contextual rule is simply one where a specific word appears as a part of the rule, as in contextual rule C4 in Table 2. Consider the following case, where we pre–tag *meeting* as a noun (NN):

```
Mona will chair the evening meeting//NN
```
(9)

The Initial State Tagger of the standard distribution will assign the following tags:

```
Mona/NNP will/MD chair/VB the/DT evening/NN meeting//NN
```
(10)

We were surprised that in cases like these the Final State Tagger made no transformations. In particular, it surprised us that it did not apply rule C4, which says that if a noun (NN) precedes the word *meeting* then it should be tagged as an adjective (JJ). By all accounts this rule should be triggered.

However, after some investigation in the source code we determined that the Brill Tagger internally appends a backslash (/) to a word that has been pre–tagged, which makes it impossible for it to trigger any contextual rule that is lexicalized with that word. Thus in the above case the Brill Tagger viewed the word in the sentence as *meeting/*, whereas it viewed rule C4 as requiring *meeting*. But, as these are different the sentence does not trigger the contextual rule. We see no particular reason to avoid this behavior, so we overrode this particular feature and now allow lexicalized contextual rules to be triggered by pre–tagged words as well. With that change in place, the Brill Tagger uses rule C4 and produces the expected output:

```
Mona/NNP will/MD chair/VB the/DT evening/JJ meeting//NN
```
(11)

7 Conclusions

This paper describes an approach which guarantees that the pre–tagged words provided to the Brill Tagger will be unchanged throughout tagging and appear in the final output, and that they will affect the tags of neighboring words. We argue that this is a reasonable way to utilize prior knowledge that may be provided to the Brill Tagger, and showed via an extensive pre–tagging experiment with the SENSEVAL-2 English lexical sample data that pre–tagging has a reasonable impact on the tagging. We also show how the impact is commensurate with what we would expect any change in contextual rule processing to have. The authors have made a patch available to the Brill Tagger from their web sites that will implement guaranteed pre–tagging, and also correct a slight anomaly in handling lexicalized contextual rules.

8 Acknowledgments

We would like to thank Rada Mihalcea, Grace Ngai, and Radu Florian for useful discussions regarding pre–tagging. Grace Ngai suggested describing pre–tags as anchors or seeds, and we have adopted that terminology. We are grateful to

Eric Brill for making the source code to his tagger available, without which this project would not have been possible.

Ted Pedersen is partially supported by a National Science Foundation Faculty Early CAREER Development award (#0092784). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either express or implied, of the sponsors or of the United States Government.

References

1. E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Computational Linguistics*, Trento, Italy, 1992.
2. E. Brill. Some advances in rule-based part of speech tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994.
3. E. Brill. Transformation–based error–driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.
4. P. Edmonds and S. Cotton, editors. *Proceedings of the Senseval–2 Workshop*. Association for Computational Linguistics, Toulouse, France, 2001.
5. L. Ramshaw and M. Marcus. Exploring the statistical derivation of transformational rule sequences for part-of-speech tagging. In *ACL Balancing Act Workshop*, pages 86–95, 1994.