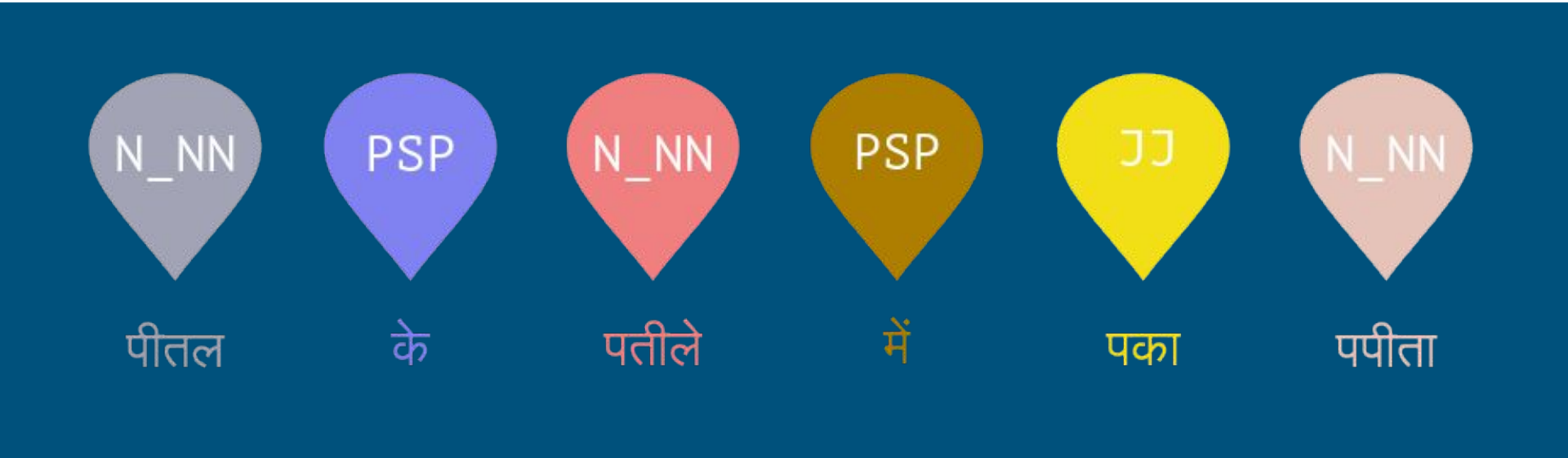
Brill’s Tagger for Hindi

A report by Abhigyan Ghosh

POS Tagging



About

Brill’s tagger is rule based tagger which depends on a dictionary to get possible tags for each word to be tagged. Disambiguation is done by analysing the linguistic features of the word, its preceding word, its following word and other aspects.

Process of Tagging

1. From a manually tagged Hindi corpus, first we extract the frequency of occurrence of each word
2. Then we run the parser back on the test sentences to find cases in which the tag for the word is not the most frequently occuring
3. We then formulate rules to fix those cases any try to increase the efficiency

Finding Rules

For the cases in which the original tag did not match with the statistically most probable tag, we track the tags before and after it and mark these as suspected rules. We also track how many times the tag did not match in this context and sort it in descending order.

From the check of suspected rules, we check for consistency of rules by doing a simple context consistency analysis.

Consistent Rules

From the previous method we get a list of rules. But the rules in which the original context is same are ignored as those cases are ambiguous for our tagger. For example we cannot have rules of the type in which:

X Y PREV-TAG A NEXT-TAG B

X Z PREV-TAG A NEXT-TAG B

i.e. A|X|B->A|Y|B and A|X|B->A|Z|B

where a,b is the context and x is the original tag and y is the new tag. So we remove all such rules from the list of rules. Also rules which are statistically unimportant can be ignored.

Some Rules Used

Successful

- ANY V_VAUX PREV-TAG V_VM NEXT-TAG V_VAUX
- V_VM V_VAUX PREV-TAG V_VM OR V_VAUX
- DM_DMD PR_PRP NEXT-TAG PSP PREV-TAG PSP
- DM_DMD|V_VAUX|CC_CCS -> PR_PRP|V_VM|CC_CCS

Not so Successful

- N_NN N_NNP PREV-TAG N_NNP OR NEXT-TAG N_NNP
- V_VAUX V_VM NEXT-TAG CC_CCS AND PREV-TAG V_VM

Problems

The extra 7-8% is difficult to overcome. The maximum which we were able to achieve is 92.969752% on 10% of the corpus.

Probable Solution

The automatically generated rules are good and can be expected to be better for larger test data.

Demo

Thank You

Review

Would you recommend this book?
Write your review here.