Sayar Ghosh Roy - 20171047

Bengali Training Data with tags obtained from the given data:

[For a word in the training set, all possible tags found in the given dataset have been listed. Note that, I found atmost one tag for each word in the given dataset.]

1) অনেকদিন/N_NN পর/N_NST বন্ধুর/N_NN সঙ্গে/PSP দেখা/V_VM ।

2) বিকেল/N_NN চারটা/N_NN নাগাদ/PSP অমল/N_NNP আমাদের/PR_PRP বাসায়/N_NN এল/V_VM_VF ।

3) তুমি/PR_PRP এতক্ষণ/N_NN কী/PR_PRQ করছিলে/V_VM_VF?

4) দারোগাবাবু/N_NN চেয়ার/N_NN ছেড়ে/V_VM উঠে/V_VM দাঁড়ালেন/V_VM_VF ।/RD_PUNC

5) কাঠবেড়ালিটা/N_NN লেজ/N_NN তুলে/V_VAUX ছুটছে/V_VM_VF ।/RD_PUNC ।

6) সবাইকে/PR_PRP সব/QT_QTF কথা/N_NN বলা/V_VM যায়/V_VAUX না/RP_NEG ।/RD_PUNC ।

7) তাঁর/PR_PR গল্প/N_NN যতই/PR_PRL শুনি/V_VM_VF ,/RD_PUNC ততই/PR অবাক/N_NN হই/V_VM_VF ।/RD_PUNC ।

৪) অন্য/QT_QTF অনেক/QT_QTF সমকালীন/JJ গায়কের/N_NN মতো/PSP ডিলান/N_NNP এখনও/PR অবসর/N_NN নেননি/V_VM_VF ।/RD_PUNC বা/CC_CCD অনুষ্ঠান/N_NN থেকে/PSP নিজেকে/PR_PRF সরিয়ে/V_VM রাখেননি/V_VAUX ।/RD_PUNC, এখনও/PR তিনি/PR_PRP পৃথিবীর /N_NNP বিভিন্ন/QT_QTF দেশে/N_NN সঙ্গীতসফর/N_NN করে/V_VM বেড়াচ্ছেনV_VAUX সমান/JJ উৎসাহ /N_NN নিয়ে/PSP।

Anaysis:

বন্ধুর has been tagged as N_NN. Note that, in the BIS tag set, there is no possessive marker such as the $ in the Brown Tags. It is fine for our purpose since we are only considering the head tags. There are other similar examples. The granularity for dividing the noun tags is not really high in the BIS tag set for Bengali.

The data provided gives us a unique tag for each word and that tag is actually correct. I used a Bengali dictionary to check if the words in the training data can have more than one parts of speech and only the following 4 words matched that prerequisite:

[Brief Information on the Dictionary used: Bangiya shabdakosh is a Bengali lexicon by Haricharan Bandyopadhyay. It was first published by Bishwakosh Press, Kolkata in the year 1934. Haricharan was a scholar in Biswabharati and was inspired and patronised by Rabindranath Tagore to complete this remarkable work. He also authored a book named Sanskrita Prabeshika which is a basic write up on Sanskrit grammar and language. The dictionary lists all possible POS forms of a word and provides examples for each case.The examples in old Bengali are also preserved thereby making it an ideal resource for word study in Bengali.]

পর – adjective, noun, preposition, verb

বাসা – noun, verb

বলা - noun, verb

সমান - adjective, noun

In the given data, only 1 tag has been found for each of these words.

Consider the word পর – which can have these four possible parts of speech according to the dictionary: adjective, noun, preposition, verb. The BIS tag in the given dataset for this tag in N_NST which is a noun for location. Although the dictionary lists preposition as a possible tag, this is actually N_NST in the BIS tag set as it considers Bengali to only have postpositions which is correct. The dictionary kind of tries to approximate the POS to corresponding English POS for which this ambiguity arises and can be prevented if we stick to BIS tags for Bengali. Note that পর can also be a NN or a JJ used to refer to some object or entity not belonging to the speaker or mark an object as one not belonging to the speaker respectively. It can also be a verb meaning 'to wear'. Note that the N_NST tag is the most frequent tag for the word and hence if we go by Brill's method, no disambiguation rules are required. In other cases, the semantic content of the words and the discourse will play a major role in determining the tags, apart from simply considering the tags of the surrounding words. e.g: In an utterance such as আমার পর; পর can be N_NST or JJ based on the meaning at the level of the discourse.

বাসা can be a verb in morphologically combined words such as ভালবাসা meaning to love. It can exist as standalones in poetry only and not in prose. Note that, the base form বাস is a verb from which the noun form বাসা is derived. Unless বাসা is followed by a V_AUX, it stays a N_NN. This rule takes care of ambiguous scenarios. Note that it is not applicable in our training set.

বলা which is typically a V_VM can be a N_NN meaning strength. This form is not used anymore in modern Bengali and is replaced by বল.

সমান tagged as a JJ can also be a N_NN referrring to the state of being equal. If it is not acting upon any noun, change the tag from JJ to N_NN. Again, this rule is not applicable here.

Now, we tag our testing data based on the BIS tags given in the dataset.

Testing Data:

এই পাঁচদিনে ফটিক তার কাজ বেশকিছুটা শিখে নিয়েছে। উপেনবাবু লোক ভাল হওয়াতে অবিশ্যি খুব সুবিধে হয়েছে। তিনি ফটিককে বারো টাকা মাইনে, থাকার জায়গা, আর খেতে দেবেন। এক মাসের মাইনে আগাম দিয়েছেন। উপেনবাবু যে

লোক ভাল, সেটা ফটিক সত্যি করে বুঝেছে গতকাল।কাছেই একটা পানের দোকান থেকে উপেনবাবুর জন্য পান আনতে গিয়ে বিশু নামে আরেকটা পানের দোকানের ছেলের সঙ্গে ফটিকের আলাপ হয়। বিশুও সবে মাসখানেক হল কাজে ঢুকেছে। ঢোকার দুদিনের মধে সে একটা চায়ের কাপ ভাঙে, আর সঙ্গে সঙ্গে তার মাথার একগোছা চুল মালিক বেণীবাবুর হাতে উঠে আসে।

- An excerpt from ফটিকচাঁদ (fotikchad) by সত্যজিৎ রায় (Satyajit Ray)

Testing Data tagged according to the given tags in the dataset:

এই/DM_DMD পাঁচদিনে/N_NN ফটিক/N_NNP তার/PR_PRP কাজ/N_NN ।/RD_PUNC বেশকিছুটা/JJ শিখে/V_VM_VNF নিয়েছে/V_VM_VF ।/RD_PUNC । উপেনবাবু/N_NNP লোক/N_NN ভাল/JJ হওয়াতে/V_VM_VINF অবিশ্যি/RB খুব/RP_INTF সুবিধে/N_NN হয়েছে/V_VM_VF ।/RD_PUNC । তিনি/PR_PRP ফটিককে/N_NN বারো/QT_QTC টাকা/N_NN মাইনে, থাকারV_VM_VNG জায়গা/N_NN, আর/CC_CCD খেতে/V_VAUX দেবেন/V_VAUX ।/RD_PUNC । এক/QT_QTC মাসের/N_NN মাইনে আগাম/JJ দিয়েছেন/V_VAUX ।/RD_PUNC । উপেনবাবু/N_NNP লোক/N_NN ভাল/JJ, সেটা/PR ফটিক/N_NNP সত্যি/RB করে/V_VM বুঝেছে/V_VAUX গতকাল/N_NN ।কাছেই/N_NST একটা/QT_QTF পানের/N_NN দোকান/N_NN থেকে/PSP উপেনবাবুর/N_NN জন্য/PSP পান/N_NN আনতে/V_VM গিয়ে/V_VAUX বিশু/N_NNP নামে/N_NN আরেকটা/QT_QTF পানের/N_NN দোকানের/N_NN ছেলের/N_NN সঙ্গে/PSP ফটিকের /N_NNP আলাপ/N_NN হয়/V_VAUX ।/RD_PUNC । বিশুও/N_NNP সবে/RB সখানেক/N_NN হলV_VM_VF --/CC_CCS কাজে ঢুকেছে/V_VM_VF ।/RD_PUNC। ঢোকার/V_VAUX দুদিনের/N_NN মধ্যে/N_NST সে/PR_PRP একটা/QT_QTC চায়ের/N_NN কাপ/N_NN ভাঙে/V_VM_VF, আর/CC_CCD সঙ্গে/PSP সঙ্গে/PSP তার/PR_PRP মাথার /N_NN একগোছা/JJ চুল/N_NN মালিক/N_NN বেণীবাবুর/N_NN হাতে/N_NN উঠে/V_VM_VNF আসে/V_VM_VF ।/RD_PUNC ।

Issues:

মাইনে and কাজে did not receive any tags. They are actually common nouns. In such cases we assign the tag N_NN to the entity and check whether it is a proper noun using NER (Named Entity Recognition). This is beyond the scope of POS tagging.

একগোছা did not receive a tag either. It should be a JJ. Similar words in similar contexts received the JJ tag. We can arrive at this by using a morph analyser. এক – which is a cardinal QTC combines with গোছা, a N_NN to form a JJ meaning piled up and it modifies the NP following it.

After this, I checked all the words in the testing set in the dictionary and I found one amiguous word root which has more than one POS:
বেশ – verb, adjective
It is used as a root in the word বেশকিছুটা tagged as JJ. বেশ as a root can be a JJ meaning good or as an old verb form meaning: to enter. Note that বেশ is not used as a verb anymore and has been replaced by প্রবেশ.

Conclusion:

The tags received by the training and testing data according to the given dataset was mostly unique and without disambiguates. A Bengali word typically exists in only one POS form and we morphologically create a new word form which marks the change in POS. There are exceptions to this statement and those ambiguities are addressed usually by simple rules and by looking at the overall semantics at the sentence level. Many words which give rise to ambiguities existed in old Bengali and are not in use anymore.