CSME 206A Natural Language & Speech Processing **Spring Semester**

Lecture 11: Parts of Speech Tagging

Lecturer: K.R. Chowdhary *: Professor of CS*

11.1 Introduction

There are eight parts of speech (POS) in English: noun, verb, pronoun, preposition, adverb, conjunction, participle, and article. The POS are also called **word classes, morphological classes**, or **lexical tags**. They are important as they give significant amount of information about word and its neighbors. It is true for nouns and verbs.

Also, when we have identified, e.g., possessive pronouns *my, your, his, her, its* and personal pronouns *I, he, you, me*, we have able to identify the vicinity words.

The POS are also used for Information retrieval, as knowing POS can help us as which morphological affixes it can have. They can also help in selecting important words, like, nouns, from the text.

Some examples of POS are as follows:

- **Prepositions:** on, under, over, near, by, at, from, to, with
- **Pronouns:** she, who, I, others
- **Wh-pronouns:** what, who, whom, why, where
- **Conjunctions:** and, but, or, as, if, whom
- **Auxiliary verbs:** can, may, should, is, are
- **Participle:** up, down, on, off, in, out, at, by

11.2 Tag-sets and Parts of Speech Tagging

There are tag-sets used for parts of speech Tagging. The table 11.1 shows some of the tag sets.

Parts of speech tagging or tagging in short is process of assigning a parts-of-speech or other lexical class marker to each works in a given text. The tagging is also called *tokenization* in terms of computer based processing for natural language text.

Input to a tagging algorithm are – a string of words and specified tag-set, and output is single best tag for each word (see fig. 11.1).

Following are the examples of tagged sentences:

Table 11.1: Tag sets for English POS tagging.

Tag	description
DT	Determiner (a,an, the, this, that, those)
IN	Preposition (of, by, in)
JJ	Adjective (yellow, other)
NN	Noun (Rajan)
NNS	Plural noun (students)
NNP	Proper noun (IBM)
PP	Personal pronoun (I, you, he)
VB	Verb (eat)
VBD	Verb past tense (ate)
VBZ	Verb 3rd pers. singular (eats)

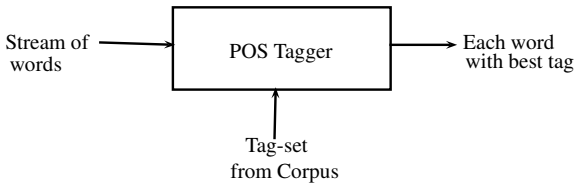

Figure 11.1: Process of Tagging.

Sentence: Book that flight.
Tag Sequence: VB DT NN.

Sentence: Does that flight serve lunch ?
Tag sequence: VBZ DT NN VB NN ?

We note the ambiguous word "book" in the above example, which makes it difficult to resolve the meaning of the sentence. The POS also resolves the ambiguity using a Corpus (like Brown corpus, or Penn Treebank tag-set). The disambiguation is carried out based on frequency of use of that words as well based on the context in that sentence.

Using the tag-set and corpus (tagged collection of sentences) it is possible to tag the words in a sentence and resolve the POS. The following example is a longer sentence with tags:

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS.

In the followings, we discuss some types of POS tagging.

11.2.1 Rule-based parts-of-speech tagging

Various languages follow the structure as SV (subject verb), SVO (subject, verb, object), or SOV (subject, object, verb). They have identify the POS based on the sentence structure.

11.2.2 Stochastic POS tagging

In this method probabilities are used in POS tagging. This method uses the algorithm known as HMM (Hidden Markov Model). Based on this approach we pick up the most likely tag for the given word. For example, the HMM tagger choose the tag sequence that maximizes the following formula:

$$P(word \mid tag) * P(tag \mid previous\ n\ tags) \tag{11.1}$$

Example 1 *HMM based POS tagger.*

Consider the following example sentences:

1. Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN.

2. People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT race/NN for/IN outer/JJ space/NN.

We shall resolve the POS of race in the first sentence. Thus we need to find out "to/TO race/??". We assume that tags for the word "race" are not given.

So wee want to find the probabilities $P(VB \mid TO) * P(race \mid VB)$ vs $P(NN \mid TO) * P(race \mid NN)$.

We can apply equation 11.1 to compute the probabilities as follows, using the probabilities given in the standard corpus as follows.

$P(NN \mid TO) = 0.021$

$P(VB \mid TO) = 0.34$

The lexical likelihoods from Brown corpus are :

$P(race \mid NN) = 0.00041$
$P(race \mid VB) = 0.00003$

If we multiply the lexical likelihoods with the tag sequence probabilities, we find the conclusions as follows:

$P(VB \mid TO) * P(race \mid VB) = 0.00001$
$P(NN \mid TO) * P(race \mid NN) = 0.000007$

Hence, in the first sentence the POS of word *race* is VB.

In the similar way we can resolve for "the/DT race/??".

□

11.2.3 Transformation-based POS tagging

It is an instance of transformation based learning (TBL) approach to machine learning. It draws the inspiration from rule based tagging as well as from Stochastic tagging.

Exercises

1. Write an algorithm of POS tagging, which uses some hypothetical Corpus lexical database of already tagged text.
2. Write an algorithm for POS tagging, that uses Stochastic method discussed in this chapter.
3. Search on the net and find out the Corpus databases, available as open source, and then write a small note about each in brief, covering mainly the technical aspects.
4. Give a comparison of three approaches discussed in this chapter for POS tagging, in respect of efficiency, complexity, and accuracy.

References

[1] D. Jurafsky and J. Martin, "Speech and Language Processing," *Pearson India*, 2002.