CS563: Natural Language Processing

Indian Institute of Technology Patna

Mid Semester Quiz Report

Name: Shashwat Mahajan
Roll Number: 1801CS46

Name: Shashwat Mahajan
Roll Number: 1801CS46

Question 1: PoS Tagging

Solution

a. Consider the following PoS tags: N (Noun), V (Verb), A (Adjective), O (Others)

With(O) Wriddhiman(N) Saha(N) gaining(V) support(N) from(O) the(O) cricket(N) fraternity(N) after(O) putting(V) out(O) a(O) WhatsApp(N) chat(N) showing(V) an(O) unnamed(A) journalist(N) using(V) strong(A)-arm(N) tactics(N) to(O) get(V) an(O) interview(N), the(O) Indian(A) cricket(N) board(N) (BCCI(N)) is(V) thinking(V) of(O) hiring(V) an(O) agency(N) to(O) deal(V) with(O) such(A) cases(N) in(O) its(O) ecosystem(N).

Assumption: strong-arm are considered separate words

b. Annotate the above sentence with the following named entity tags: Person, Location, Organization, Others

Consider the following acronyms for the above tags:
Person (Per), Location (Loc), Organization (Org), Others (Oth)

With Wriddhiman(Per) Saha(Per) gaining support from the cricket fraternity after putting out a WhatsApp(Org) chat showing an unnamed journalist using strong-arm tactics to get an interview, the Indian(Oth) cricket board(Org) (BCCI(Org)) is thinking of hiring an agency to deal with such cases in its ecosystem.

c. Explain whether the following example "I saw a man with a telescope" has only lexical ambiguity or syntactic ambiguity or both.

The sentence has only syntactic ambiguity. Lexical ambiguity arises out of ambiguity in a single word being parsed differently. Here no single word can be parsed differently. Thus no lexical ambiguity.

The syntactic ambiguity arises out of the fact that the sentence can be parsed as a whole in more than one way. Here this sentence could mean that the first person saw a man with the possession of a telescope. Or it could mean that the first person saw the second person using a telescope.

Question 2: PCFG

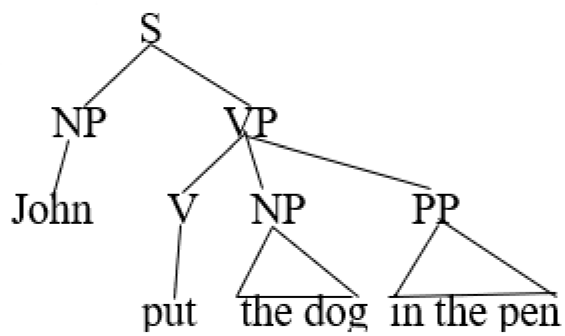
Why can't Vanila PCFG resolve all kinds of syntactic ambiguities? Provide at least one example where such PCFG fails to disambiguate, and explain how such kinds of ambiguities can be addressed using additional lexical level information.

Solution

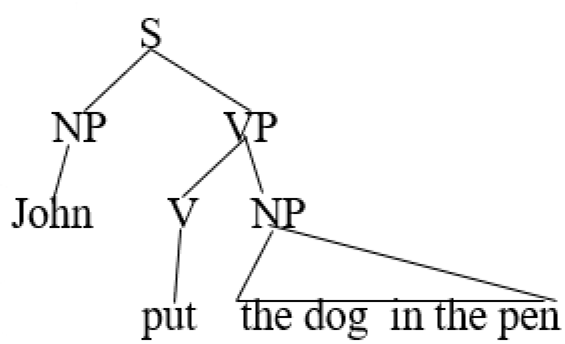
Since probabilities of productions do not rely on specific words or concepts, only general structural disambiguation is possible. Consequently, vanilla PCFGs cannot resolve syntactic ambiguities that require semantics to resolve, e.g. ate with fork vs. meatballs. In order to work well, PCFGs must be lexicalized, i.e. productions must be specialized to specific words by including their head-word in their LHS non-terminals (e.g. VP-ate).

Example:

John put the dog in the pen.



Added lexical features help the model.



Question 3:

Solution

a)

Coreference resolution is an NLP task that involves determining all referring expressions that point to the same real-world entity. A referring expression (i.e., a mention) is either a noun phrase (NP), a named entity (NE), or a pronoun, which refer to an entity in the real world known as the referent

Coreference resolution typically requires a pre-processing pipeline comprising a variety of NLP tasks (e.g., tokenization, lemmatization, named entity recognition, part-of-speech tagging). Historically, these tasks are addressed before training the coreference resolution model (in a pre-processing stage) and, consequently, errors made by pre-processing models impact coreference resolution models, which typically assume that the information provided from this pre-processing stage is correct.

Named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations or locations . Identifying the token-level boundaries of mentions (such as named entities) is a necessary step toward obtaining mention clusters from natural text; thus, NER is tightly related to coreference resolution.

Inversely, if we have details about coreference resolution then it can help us in NER, because if we know that two words are correlated then their NER result would also be same. So if we know NER result for one such word, we know its result for all words related to it.

b)

Global contextual information can be leveraged from two levels, i.e., both word and sentence.

At word-level, a document graph is constructed to model a wider range of dependencies between words, then obtain an enriched contextual representation for each word via graph neural networks (GNN).

At sentence-level, for appropriately modeling wider context beyond single sentence, we employ a cross-sentence module which encodes adjacent sentences and fuses it with the current sentence representation via attention and gating mechanisms.

In local context information, features like local content, person prefix and corporate suffix, etc. are used. However, there are contexts where the prefixes and suffixes occur in non-local places.
For eg.

1. Elon is the CEO of Spacex Inc.
2. Elon was named to be the CEO of a key player of the fast paced space exploration, SpaceEx Inc.

In both these contexts, the initcaps are the same where Elon is linked to Spacex. However, the placement of the elements makes it difficult for performing NER with local context only. In this case, modeling of non-sequential global contextual features will give better performance.