Natural Language Processing: Assignment 2

Shifei Chen

---

1 POS-Tagging

My tagger achieved an accuracy of 93.10% by using parameters `-t 3 -f 5 -s 5`. Here are some examples of errors my tagger made.

1. "From the AP comes this story."

"AP" was mistagged as a NOUN rather than a PRON. This is pretty obvious since AP indeed is the name of the news agency.

2. "The sheikh in wheel-chair has been attacked with a F-16-launched bomb."

"F" was tagged as a NOUN in the gold standard and a PRON in my result. I think both of the tags are not correct. Here "F-16-launched" should be treated as a single adjective. If we have to separate them, I would believe that "F" is a pronoun as it is a part of the model name of a plane.

3. "US troops there clashed with guerrillas in a fight that left one Iraqi dead."

My tagger believed the word "Iraqi" is a ADJ rather than a PRON. A English speaker shouldn't make that mistake since he can see clearly that the clause "left one Iraqi dead" is the consequnce of the fight and because of that, "left" is a verb and "dead" is an adjective. My tagger might not be able to trace that far away. It might simply looked at the words around "Iraqi" and believed this fit the structure of num. + adj. + noun.

4. "Xinhua alleged that 'Many of the Iraqis, who suffer ...'"

"that" should be a SCONJ instead of a DET because of the word "alleged" and the fact that everything after "that", from a semantic prospective, is the content of Xinhua agency's allegation.

5. "2015 is going to rock!"

We know "rock" can be a noun or a verb but in this sentence, a verb will make more sense semantically. Hence the word "to" should be a particle than a preposition.

In general, I think in my tagger only part of the mistakes are considered to be genuinely ambiguous, for example ADP vs. PART. There are still plenty of cases which are not ambiguous to human beings at all. The tagger is restricted to look forward or backward only N words (specified by the parameter `-t`) therefore it cannot have a better understanding of the context overall, especially in sentences consist of clauses. Another issue is the lack of semantics analysis. It made my tagger confused. There is a sentence in our corpus goes

... they hear a company who's stated goals include "Don't be evil," ...

My tagger tagged "evil" as a noun while our golden standard marked it as an adjective. Both of them make sense, although personally I believe that company is Google hence I would choose ADJ. Senteces like "I'm going to work.", "to" being a particle or preposition makes sense in either way since "work" can mean the action (verb) or the place (pronoun/noun). These two problems exist in our golden standard as well.

... the price would be too high for investors to make a real profit.

In this sentence ”for” should be a prepostion, not a subordinating conjuction. There is no clause in this sentence.

I found several tagsets for my mother language, Chinese: the Chinese Penn Treebank POS tagset[1], an SVMTool-Based Chinese POS Tagger[2], FudanNLP[3], etc. Here we take a closer look at the Chinese Penn Treebank tagset. Comparing to its English siblings, the Chinese tagset has 33 tags. Words “把” and “被”, which means ”make sth. to do” and passive voice, respectfully, are separated from other verbs and prepostions since their identities are still highly controversial. Another intresting thing I have noticed is that “的”, “地” and “得”, which are the three most common particles, are categorised into DEC, DEG, DER and DEV. This is a reasonable choice as their appearence usually gives people hints about the part-of-speech of words around them, like “得” usually indicates the word it follows is always a verb and the word after it is usually an adverb.

Therefore I hold the idea that tokenization doesn't necessarily has to be done before tagging, at least in Chinese. Of course tagging could benefit from tokenization because it gives information of word boundaries, sentence boundaries, average word length, etc. On the other hand, tagging could also help improve tokenization as different part-of-speeches will have impacts on tokenizaiton. For example distinguishing ”.” as a punctuation from a symbol of abbreviation. Another example is like in languages like Japanese, its particles usually contains rich information about the structure and semantics. “好きだ” “だ” at the end of the sentence usually means it is an auxiliary verb and we can therefore separate it from the other parts of the sentence. It also tell us that “好き” here should be a noun or a na-adjective (It actually means ”like” and in Japanese ”like” is an adjective). Tagging does not need to be after tokenization and I believe it is better if they can be carried out simultaneously.

Finally in Lab 6 we did an invesitigation on key sequences and predicted words in mobile phone inputs. By definition HMM models should always have a sequence of T observations (or signals) $O_1, O_2, O_3, ..., Q_T$, and a set of N states $Q_1, Q_2, Q_3, ..., Q_N$[4]. If we observe the sequences of number keys then our states will be letters of the words because this is what we see at the surface. Our signals will be the numbers as they are not what we can directly observe and they are hidden behind the predicted words.

References

[1] Fei Xia. "The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0)". In: *IRCS Technical Reports Series* (2000), p. 38.

[2] 王丽杰, 车万翔, and 刘挺. "基于 SVMTool 的中文词性标注". In: 中文信息学报 23.4 (2009), pp. 16–22.

[3] Xipeng Qiu, Qi Zhang, and Xuanjing Huang. "Fudannlp: A toolkit for chinese natural language processing". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2013, pp. 49–54.

[4] Daniel Jurafsky and James H. Martin. *Speech and Language Processing, 2nd Edition*. 2nd. Prentice Hall, May 2008. ISBN: 9780131873216. URL: `http://amazon.com/o/ASIN/0131873210/`.