COMP 6751 Natural Language Analysis

Project 3 Report 1 Grammar Design

Student: Yixuan Li 40079830

Table of Contents

Expectations of originality:

I, student 40079830, certify that this submission is my original work and meets the Faculty's Expectations of Originality.

Date: November 20, 2020

I. Discussion on Grammar Design

1) Sentiment attribute on Grammar Productions

To analyze the sentiment of a phrase of a sentence, I added a sentiment attribute **called SENTI** on the grammar productions, and there are in total three different sentiments: **positive, neutral, negative**, which are assigned on terminals. For example,

```
# Adjective
JJ[SENTI=positive] -> 'compelling' | 'perfect' | 'well-intentioned' | 'entertaining'
JJ[SENTI=negative] -> 'manipulative' | 'rancid' | 'ugly' | 'dull' | 'scary'
JJ[SENTI=neutral] -> 'dramatic' | 'gut-wrenching' | 'low' | 'other' | 'long'
```

The SENTI label is percolated up in a bottom-up manner from the terminals, and the SENTI label of S is the sentiment of the entire phrase or sentence.

2) Improvement on project 2 grammar productions

In addition to the SENTI attribute, I also added some additional helping attributes to analyze and limit the sentiment relations between different grammatical categories like S, NP, VP, JJ etc. Here's a table to list the additional attributes.

Attributes	Potential values	Grammatical categories which hold the attribute	Meaning
AUX	+AUX / -AUX	V	auxiliary verb
passive	+passive / -passive	V	passive mode
negation	+negation / -negation	RB	negation NOT
and	+and / -and	CC	and relation
but	+but / -but	CC	but relation
or	+or / -or	CC	or relation
of	+of / -of	IN	preposition "of"

In addition to the attributes above, there are a few additional grammatical categories introduced compared to project 2 grammar:

New Grammatical categories	Meaning	Example
ADJP	Adjective phrase	"**too long but entertaining**"
SBAR	Subordinary clause	"the people **who were there**"
WP	WH-pronoun	"the people **who** were there"
EX	Existential introducer	"the people who were **there**"

3) More explanation on sentiment percolation

- S sentiment percolation rules
  The common sentiment production for S is:

S[SENTI=?s, -INV] -> NP[NUM=?n, PERSON=?p, SENTI=?s] VP[TENSE=?t, NUM=?n, PERSON=?p, SENTI=?s]

But since sentences could be connected by conjunctions, so I added additional grammar productions for S which contains a conjunction. And for simplicity, I only present AND relation rules here, but more details can be found in the grammar file "sentianalysis_grammar_s.fcfg".

S sentiment	Grammar rule
positive	S[SENTI=positive, -INV] -> S[SENTI=positive, -INV] CC[+and] S[SENTI=positive, -INV] | S[SENTI=positive, -INV] CC[+and] S[SENTI=neutral, -INV] | S[SENTI=neutral, -INV] CC[+and] S[SENTI=positive, -INV]
negative	S[SENTI=negative, -INV] -> S[SENTI=negative, -INV] CC[+and] S[SENTI=negative, -INV] | S[SENTI=negative, -INV] CC[+and] S[SENTI=neutral, -INV] | S[SENTI=neutral, -INV] CC[+and] S[SENTI=negative, -INV]
neutral	S[SENTI=neutral, -INV] -> S[SENTI=neutral, -INV] CC[+and] S[SENTI=neutral, -INV]

- NP, VP, SBAR, PP sentiment percolation rules

The percolation for NP, VP, SBAR, PP and corresponding examples are written in sentianalysis_grammar_s.fcfg.

II. Limitations on the grammar

In my testing data, there are 9 positive sentences, 9 negative sentences and 3 neutral sentences, which are saved in "data/" directory. The grammar can label them all correctly, but there are some limitations on the grammar.

1) Limitation 1: Grammar are a little bit bloated

I took the consideration of "neutral" sentiment, so sentences such as "this is an example of movie making." will be labelled as "neutral". However, it makes the grammar consider more cases. For example, for the AND relation for S rule listed above,

S[SENTI=positive, -INV] -> *S[SENTI=positive, -INV] CC[+and] S[SENTI=positive, -INV] | S[SENTI=positive, -INV] CC[+and] S[SENTI=neutral, -INV] | S[SENTI=neutral, -INV] CC[+and] S[SENTI=positive, -INV]*

In a positive sentence which contains 2 sub-sentences conjoined by "and", there are 3 possibilities:

S sentiment	1st sub-sentence sentiment		2nd sub-sentence sentiment
positive	positive	and	positive
	positive		neutral
	neutral		positive

However, if only "positive" and "negative" sentiment labels are considered, the table above can be simplified to

S sentiment	1st sub-sentence sentiment		2nd sub-sentence sentiment
positive	positive	and	positive

So grammar will be dramatically reduced and will be clean.

I reckon that it makes more sense to label the sentence "this is an example of movie making." as "neutral" rather than "positive" or "negative", so I complemented the grammar with additional rules but a little bit bloated.

2) Limitation 2: There might be multiple sentiment outputs

For the same sentence, if there are multiple parse trees with different sentiment labels, I let the parse trees vote for their chosen sentiment, and then I will output the sentiment label with the most votes and one parse tree with that sentiment. However, in some cases, different sentiments may have equal votes, then I output them all. So if a pair of opposite sentiments have equal votes, I output them both.

Example sentence: *a perfect example of rancid , well-intentioned , but shamelessly manipulative movie making .*

The output sentiment is *[positive, negative]* meaning it is labelled as "positive" or "negative". The reason is that "positive" and "negative" got equal votes from their parse trees. This result is caused by the "of issue" in the NP rule:

NP[NUM=?n, PERSON=?p, SENTI=?s] -> *NP[NUM=?n, PERSON=?p, SENTI=?s] IN[+of] NP | NP[NUM=?n, PERSON=?p] IN[+of] NP[SENTI=?s]*

If a noun phrase contains "of", its sentiment may be determined by the part before "of" or the part after "of". I will make 2 examples to explain it.

NP sentiment	the part before "of"		the part after "of"
positive	a perfect example *[positive]*	of	movie making *[neutral]*
negative	an example *[neutral]*		manipulative movie *[negative]*

As we can see, the part before "of" and the part after "of" both have the chance to determine the sentiment of the NP. So the example sentence above "*a perfect example of rancid , well-intentioned ,*

but shamelessly manipulative movie making ." gets 1 positive vote from "a perfect example" part and 1 negative vote from "rancid , well-intentioned , but shamelessly manipulative moving making" part. And I output both sentiments and one of the parse trees of each sentiment.

And NP containing "with" has the similar issue as above, and I discussed it with examples in the Demo Report 2 under *Positive test case 5 result explanation*.

III. Discussion on semantics with the grammar

Idea 1 on semantics with feature-based grammar is to compare similarity on sentence structure given two sentences. After we refine the grammar on training sentences, we know the grammar productions that the sentences used given two sentences. Hence we could compare the sentence structure by comparing the grammar rules they used. For example

Sentence 1: The apple is green and raw.
Sentence 2: The banana is yellow but just ripe.

NP sentiment	**Used grammar rules**		
Sentence 1	S → NP VP	NP → DT NN VP → V[+AUX] ADJP	ADJP → JJ CC[+and] JJ
Sentence 2	S → NP VP	NP → DT NN VP → V[+AUX] ADJP	ADJP → JJ CC[+but] ADJP ADJP → RB[-negation] JJ

We could know that Sentence 1 and Sentence 2 are similar in terms of sentence structure.

Moreover, moving to the next step (idea 2), by leveraging and applied WordNet on idea 1, we may compare the similarity of sentence meaning given two sentences. If we take the Sentence 1 and Sentence 2 above, we can query the words in WordNet.

NP sentiment	**WordNet query results** [5]	
Sentence 1	Noun	"apple" → noun.food, noun.plant
	Adjective	"green" → adj.all "raw" → adj.all
Sentence 2	Noun	"banana" → noun.food, noun.plant
	Adjective	"yellow" → adj.all "ripe" → adj.all
	Adverb	"just" → adv.all

We know that "apple" and "banana" are from the same synset and they belong to the same NP rule. "green" and "yellow" are from the same synset but they belong to a little different ADJP rule, and similarly for "raw" and "ripe". Therefore, the additional work would possibly be

1. extend the existing pipeline of the project 3 and connect to WordNet
2. make the input to the pipeline only 2 sentences and add a module for sentence structure comparison
3. design a heuristic scoring system to compare the words in 2 sentences that belong to same/different grammatical categories

IV. References

1. References in Project 2 and Project 2 grammar productions
2. NLTK book chapter 9 (http://www.nltk.org/book/ch09.html)
3. Adjective phrases explanation (http://www.languagetools.info/grammarpedia/adjectivephrase.htm)
4. Wikipedia: Adverbial Phrase (https://en.wikipedia.org/wiki/Adverbial_phrase)
5. WordNet, Princeton University (https://wordnet.princeton.edu/)