Natural Language Processing: Assignment 4

Shifei Chen

---

1 Lexical Semantics: Error analysis

In my case, a Naive Bayes Classifier with collocational feature representation is the best performing classifier for both the word "hard" and "serve", while a same classifier with bag-of-words feature representation is the most suitable one for the word "interest". The window size is 3 for all of these classifiers.

From the lab, `HARD1`, `INTEREST1` and `SERVE10` are the most difficult sense to estimate. In the case of "HARD1", many of the errors came aloneside certain phrases. For example there were 5 "a hard time", 6 "the hard way", 3 "make a hard choice", 5 "hard lesson" and 9 "the hardest" inside these 59 failed guesses. The pattern that mis-labeling usually lives inside some fixed phrases, can also be observed from the sense `INTEREST1`. We know that by the definition, the Naive Bayes Classifier always assumes context words are independent. However from the observations sometimes words appear together to form up fixed collocation, idioms, etc. Their probabilities are not always independent. To fix this problem, I would propose that we can group these fixed collocations together and treat them as a single word to lower their probabilities amplified by calculation them independently. E.g. we can calculate $P(the + way|HARD1)$ instead of $P(the|HARD1) * P(way|HARD1)$.

Another thing I have observed is that my classifier can only categorize words into one of their senses, especially for the word "serve". There are a lot of cases where "serve" should be labeled as `SERVE10`, work for or be a servant to, while myself would argue that `SERVE6`, provide food, can also be the correct sense. The classifier is designed to make a binary choice between senses by choose the one with a higher probability, but imagine a case where both senses have very close probabilities, e.g. 0.49 and 0.51, we cannot always say that one of the senses is significantly better than the other one, just like the human annotator sometimes cannot decide which sense to assign. It might be the result that these senses are not distinct from each other enough, or it might be the result that the word itself is often ambiguous.

The script in the lab material used accuracy as a measurement of the performance of the classifier, but it also showed details where our classifier labeled correct senses and where it mislabels, as the confusion matrix of "hard" shows below.

```
         |    H    H    H  |
         |    A    A    A  |
         |    R    R    R  |
         |    D    D    D  |
         |    1    2    3  |
---------+-------------------+
HARD1    |<643>   39   20  |
HARD2    |    6 <73>    9  |
HARD3    |    5   12 <60>  |
---------+-------------------+
(row = reference; col = test)

<ConfusionMatrix: 776/867 correct>
```

Then we can plug the numbers into the formulas on the textbook[1, p. 455] to calculate the precision and the recall score for each of the sense.

$$Precision(HARD1) = \frac{643}{643 + 6 + 5} = 0.9832 \tag{1}$$

$$Precision(HARD2) = \frac{73}{39 + 73 + 12} = 0.5887 \tag{2}$$

$$Precision(HARD3) = \frac{60}{20 + 9 + 60} = 0.6742 \tag{3}$$

$$Recall(HARD1) = \frac{643}{643 + 39 + 20} = 0.9160 \tag{4}$$

$$Recall(HARD2) = \frac{73}{6 + 73 + 9} = 0.8295 \tag{5}$$

$$Recall(HARD3) = \frac{60}{5 + 12 + 60} = 0.7792 \tag{6}$$

Moreover, $F - Measure$ score provides a way to combine both the precison score and the recall score. We assume here precision and recall are euqally important, that is to assume $\beta = 1$ in the $F - Measure$ score, then the $F - Measure$ score for each of the sense of the word "hard" are listed below,

$$F_1(HARD1) = \frac{2 * 0.9832 * 0.9160}{0.9832 + 0.9160} = 0.9484 \tag{7}$$

$$F_1(HARD2) = \frac{2 * 0.5887 * 0.8295}{0.5887 + 0.8295} = 0.6887 \tag{8}$$

$$F_1(HARD3) = \frac{2 * 0.6742 * 0.7792}{0.6742 + 0.7792} = 0.7229 \tag{9}$$

Using precision and recall instead of a single accuracy score enables us to see whether we should improve the accuracy or the coverage of our algorithm, which are usually lying in two opposite directions. Like in `HARD2`, we have done better in coverage than in accuracy from the better recall score, which means that our feature vector is sufficient but not necessary enought for `HARD2` and we might need to include more collocations related to that sense while remove some of the more general collocations at the same time. On the other hand, a single accuracy score doesn't reveal the direction to improve our classifier, we could end up with working in the wrong direction and make the accuracy score even worse.

Another advantage of precision and recall over accuracy is that they give a detail view of each aspects, or senses of the same word. From the example above, if we just look at the accuracy score we might think we are very close to the human ceiling, which usually is around 90%. But if we look at the precision and the recall scores we will find out that both `HARD2` and `HARD3` still have plenty of spaces for improvements, even though `HARD1` is already better than the human ceiling.

2 Semantic Role Labeling

Labeling words with their semantic roles doesn't differ a lot from dependency parsing or part-of-speech tagging. I can imagine of labeling senses as tagging part-of-speeches and labeling arguments as drawing arcs from the head word to its dependent, despite we labeling them with different roles in the lexicon instead of drawing actual arcs from word to word. I always begin with getting a through understand of the semantic of the sentence and then decide which one of the senses the keyword fits. Since the four senses of the word "count" are quite distinct it was not confusing. What could cost me some time is labeling the core argument roles. I had to make sure whether certain roles of a specific sense does exist in the sentence or not, then pinpoint which segment of the sentence, both its the position and the length, belongs to that role.

After finishing the labeling work, people need to definie an efficient way to measure the level of agreement between different annotations. We need something more scientific than saying "There are x difference between your annotation and mine" thus people have introduced "inter-annotator agreement", a numeric value between 0 to 1 (sometimes it can be negetive) to descibe the similiarity between different annotations. An inter-annotator agreement of 1 means that annotators agree with each other completely and 0 means they don't agree with each other at all. There are some approaches to calculate inter-annotator agreement, such as Cohen's kappa[2] or Fleiss' kappa[3]. Here I have adopted Cohen's kappa to calculate my inter-annotator agreement between my classmate's annotation and mine, as there were only two annotations and Cohen's kappa is designed to handle no more than two samples.

In the original version of Cohen's kappa all of its data should be binary, which means every item should either be categorized to category k completely or not. But when annotating semantic roles I found that most of the disagreement between my classmate and I happens in the argument roles and the originial Cohen's Kappa can not describe different levels of disagreement precisely in this case. Even though there is a modified weighed Cohen's kappa available, which scores disagreements between categories[4], it is still not what I was looking for. I would like to measure the disagreement within a category by a scale thus I have decided to notate it in the

following way: for each category if both annotators agrees with the sense i and the arguments we give $cell_{[i]}$ a full score, if they agree with the sense but have controversy over the arguments we give it a half score. If they don't agree with the sense at all then we give $cell_{[i,j]}$ a full score where i and j represents the senses backed by two annotators. The number of items N is 10 since we have 10 sentences to annotate, everything other than this is the same to the original Cohen's kappa method.

The original result chart for the discussion of the annotation, and the table I used to calculate Cohen's kappa are showed below.

Table 1: Detailed Results of Semantic Role Labeling between Two Annotators

Sentence	Agree on Sense (Which Sense)	Agree on Arguments (Which Arg Disagreed)
1	Y(count.03)	N(A0)
2	Y(count.01)	Y
3	Y(count.02)	N(A2)
4	Y(count.02)	N(A2)
5	Y(count.01)	Y
6	N(count.02 & 01)	N
7	Y(count.02)	Y
8	Y(count.04)	N(A1)
9	Y(count.03)	N(A0)
10	Y(count.01)	N(A1)

Table 2: Results of Inter-Annotator Agreement Using Cohen's Kappa

	count.01	count.02	count.03	count.04	Total
count.01	2.5				2.5
count.02	1	2			3
count.03			1		1
count.04				0.5	0.5
Total	3.5	2	1	0.5	
Agreement	2.5	2	1	0.5	6
By Chance	0.875	0.6	0.1	0.025	1.6
kappa	0.5238				

Sentence 2, 5 and 7 are the ones where I agree completely with the other annotator. 2 of them are the first sense of the word "count". In fact the last sentence "They count the cans in the trash to make sure.", which was also annotated as `count.01`, was also very close to a full agreement–we only had different opinions on whether we should include the prepositional phrase "in the trash" into A1. On the other hand, `count.02` was the most difficult one to reach a completely agreement on both the sense and the arguments. For example in sentence 6, "Many illegals were not counted in the population until the mid-80s.", "counted" can be both interpreted as counting the action or being included from my point of view. It will remain ambiguous until we have the context to figure out whether they did counted those criminals one by one or not.

References

[1] 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/`.

[2] *Cohen's kappa - Wikipedia*. [Online; accessed 13. Dec. 2018]. 2018. URL: `https://en.wikipedia.org/wiki/Cohen%27s_kappa`.

[3] *Fleiss' kappa - Wikipedia*. [Online; accessed 13. Dec. 2018]. 2018. URL: `https://en.wikipedia.org/wiki/Fleiss%27_kappa`.

[4] Jacob Cohen. “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.” In: *Psychological bulletin* 70.4 (1968), p. 213.