Assignment 4: Semantics

Oreen Yousuf

December 14, 2020

1 Lexical Semantics: Error analysis

1.1

Interest1 was the hardest "sense" for the model to estimate for the word "interest". It had an 85.89% incorrect labeling by the model. Of the incorrect labels, the model guessed a meaning of Interest2 the majority of the time.

Hard3 was the hardest "sense" for the model to estimate for the word "hard". It had a 67.53% incorrect labeling by the model. Of the incorrect labels, the model overwhelming guessed a meaning of Hard1.

Serve2 was the hardest "sense" for the model to estimate for the word "serve". It had a 64.497% incorrect labeling by the model. Of the incorrect labels, the model guessed a meaning of Serve12 the majority of the time.

The Naive Bayes classifiers worked the best on all 3 keywords. I think there are patterns/systematic errors within this example, though, specifically in regards to having a less rigid rule-set for contextual information. The conceit of word sense disambiguation follows the outline of selecting a keyword in context, gaining specified features of the sentence/excerpt the keyword belongs to, like its part-of-speech (POS), neighboring word(s), and keying in on its position, computing the probabilistic approximation model, and returning the semantic meaning associated with the highest probabilistic score. If we take the word "hard" for example, because hard1 and hard3 have the same POS and therefore have, liberally, similar neighboring words, its position can be said to be the most crucial characteristic to determine it's correct semantic meaning. Which would not be surprising as there was a 67.53% incorrect labeling. In addition to this, of the total 67.53%, only 6.49% stemmed from misidentification as hard2, which has a much more different definition of "dispassionate". This drastically different definition allows for it to appear in unique contexts, such as the following scenario:

> example number: 151
>
> sentence: on monday and wednesday , multiple accidents caused miles - long morning backups before our HARD hatted friends even went to work .
>
> guess: HARD2; label: HARD3

Hard2's definition is "dispassionate", with an example of "a hard bargainer." The use of this definition of "hard" most suits nouns like people (i.e. bargainer), unlike inanimate objects/nouns that are better suited for hard1 and hard3. It is easier to see why the model guessed hard2 for example number 151 after thinking about this. The same can be said for "interest". The same can be said for "interest" and "serve". Interest2's definition of "quality of causing attention to be give to" paired with the example "they said nothing of great interest" is eerily close to overlapping instances of interest1's definition of "readiness to give attention". The difference between their applications is smallest between the six total definitions of "interest". Exemplified by the following scenario:

example number: 29

sentence: other losers included pharmaceutical and textile shares , but two issues attracted investor INTEREST because of strong earning prospects for new products .

guess: interest_2; label: interest_1

No other "interest" definition is closer to interest1's uniqueness in this, and many more, situations. Interest6's profit related definition of could've potentially made it a prospective guess by the model given it's likely neighboring words relating to money like how "invester ... earning ... products" is present in example 29. The order of most misidenifications for "serve" can be seen if you continue to follow the idea of close-meaning definitions having higher numbers of misidentification. Serve6 is overwhelming tied to things like food, serve10 slightly edged out by serve12 due to it being less tied to functionality than the latter serve's definition, which is much closer to serve2's definition of holding office or serving in a specific function.

example number: 16

sentence: she slipped through a thick tangle at the edge of the abandoned village and was attracted by the rustling of several rats that were working through a kitchen midden , when a more enticing sound caught her attention . even before she began her stalk on the rats she swung through one of the thatch-roofed pole huts and leaped easily onto a wooden platform that had once SERVED as a bed for an entire seminole family .

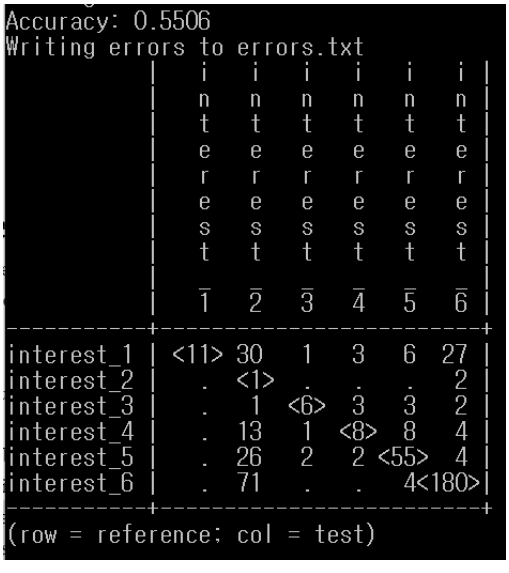
guess: SERVE12; label: SERVE2

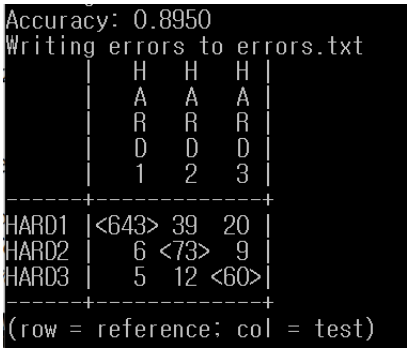
It would be hard to choose which "serve" sense would be the appropriate one between serve2 and serve12 when analyzing the excerpt of "...had once SERVED as a ..." removed from the confines of this sentence.

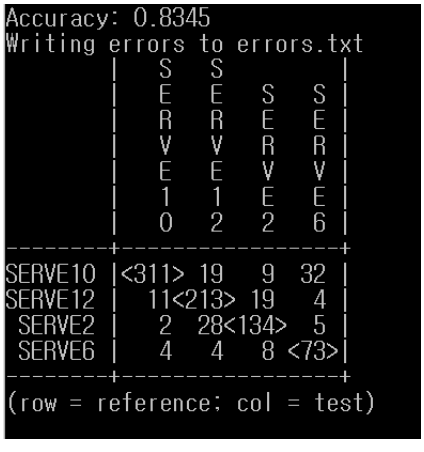
1.2

Precision(P) for Binary Classifications are: P = TruePositives/(TruePositives + FalsePositives) and for Multi-Class Classifications, like the Confusion Matrices we're working with here, the Precision is: P = TruePositivesAllClasses/(TruePositivesAllClasses + FalsePositivesAllClasses).

Below are screenshots of my confusion matrices for the three keywords we worked with:







P(hard1) = 643 / (643 + 6 +5) = 643 / 654 = 0.9831%, P(hard2) = 73 / (73 + 39 + 12) = 73 / 124 = 0.5887%, P(hard3) = 60 / (60 + 20 + 9) = 60 / 89 = 0.6741. These are for each class's precision. But for the (mean) Precision of Multi-Class Classifications/a Confusion Matrix, the precision is as follows: P(hard) = (0.9831 + 0.5887 + 0.6741) / 3 = 0.7486 .

R(hard1) = 643 / (643 + 39 + 20) = 643 / 702 = 0.91595%, R(hard2) = 73 / (73 + 6 + 9) = 73 / 88 = 0.8295, R(hard3) = 60 / (60 + 5 + 12) = 60 / 77 = 0.7792%. These are for each class's recall. But for the (mean) Recall of Multi-Class Classifications/a Confusion Matrix, the recall is as follows: R(hard) = (0.91595 + 0.8295 + 0.7792) / 3 = 0.84155 .

For multi-class classifications there is the general equation of F1 = 2 * [(P*R)/(P+R)]. This dictates that the F-Measure will be F1 = 2 * ((0.7486*0.84155))/(0.7486+0.84155)) = 0.792358 .

Precision is a helpful measure when a large number of False Positives are present because it assesses how precise, error-free, exact, etc. your model is out of predicted Positives. Spam detection is an application that can make good use of Precision. Recall can be greatly utilized when evaluating just how many True Positives were captured after your model takes in data and yields a large number of Positives. Fraud detection is an application that can make good use of Recall. F-Measure/F1 Score is simply a function of the previous two measures; a balance between them. However, when your experiment produces many True Negatives, seeing as it's an uneven class distribution, *while* still having the desire for a balance between Precision and Recall, an F-Measure/F1 Score would be a better measure to use than Accuracy.

Although Accuracy performs adequately/well with a balanced data-set or bank of information, it's somewhat limited to when your classes are balanced. Thus, it might be a good option for training set in stone data, but its costly over-attention to True Negatives and preference for similar valued False Negatives and False Positives may make you reconsider if it's always the go-to performance measure.

2 Semantic Role Labeling

I discussed my annotations with the *wonderful* Jae Eun Hong.

Sentence 1: We both had the same count sense (count.03) and arguments, an arg1 of "New Delhi" and arg2, however there was a slight difference in our arg2 strings. I had "on to render good neighborly help" whereas Jae Eun excluded "on". My initiative to include it were based on Lab 12's instructions to do so, however Jae Eun had the idea of basing the final note on the count.senses CONLL file to have count.03 to pick up on the phrase "count on" as an equivalent, which I think is a good idea.

Sentence 2: This was one of the most disputed sentences for us. I considered this as count.01 with an argo of "I" and arg1 of "like 14 explosive headlines". Jae Eun thought this would be count.02 argo "I" arg2 "like 14 explosive headlines". Even now I can see her perspective due to there potentially being a "theme" of count.02's grouping in "explosive headlines". Instinctively, I can believe both. I see this sentence as quite colloquial and somewhat of an estimation rather than concrete, numerical counting when people, including myself, use it in everyday life.

Sentence 3: I had count.03 with an argo of "I" and arg1 of "it as signed" for this sentence, however I feel as though I was confusing myself between solely analyzing the text with the sense definitions and my experience in real life. Because of this, when thinking in the former mindset I considered Jae Eun's conclusion, of count.02 with an argo of "I", arg1 of "it", and arg2 of "as signed", as correct. This is because counting this document/paper/etc. as "signed" would mean placing it in a thematic grouping of other signed documents/papers/etc. However, when thinking of how I can *hear* this sentence I believed that it could give a feeling of urgency. Quick thought process: this person has come to their conclusion,

and getting this signed is more of a formality for an already bygone conclusion. For him/her to move on, he/she must get this signed, they are depending on it being signed to move on.

Sentence 4: I had count.04 with an arg1 of "ENA's trading desk margin", and Jae Eun had count.01 with an argo of "Enron" and arg1 of "ENA's trading desk margin". I can potentially see why count.01 would be the appropriate sense here, considering the subject matter of financial gains, however I think count.04 is the true sense here. This is because I think the word "counting" can be completely substituted with "considering", and this can even further bring count.04 to the forefront as it's definition of "things mattering" works for "considering" as well as "counting".

Sentence 5: We agreed on everything for this sentence, with a count.01 and argo of "I" and arg1 of "around 25 of them". After reviewing it in on comparisons we still agree on it.

Sentence 6: We agreed on almost everything here. We had a sense of count.02 and arg1 of "many illegals", and while we both had arg2s, mine was "*in* the population", while Jae Eun's ommitted the "in". It's a small difference but I think inlcuding the "in" is noteworthy as the definition of count.02 in the count.senses CONLL file has argument 2 defined as "group", and I think being "in the population" changes what you're talking about versus simply bringing up "[a]/the population". The latter is acknowledging a population's existence, while the former shifts the perspective of the subject.

Sentence 7: We both had count.02 and an arg1 of "many immigrants", but Jae Eun didn't have any other arguments, while I had an arg2 of "social benefits". I think having the latter argument is correct because that is the thematic grouping the subject of interest/subject in question, i.e. the "many immigrants", are being excluded from.

Sentence 8: We both had an arg1 of "Marrying a Canadian", but differing senses of count.04 and count.01 for myself and Jae Eun, respectively. Jae Eun considered count.01 for the same reasons as sentence 4 stated above.

Sentence 9: We both had count.03, but I had an argo of "all my customers" and arg2 of "on it", while Jae Eun had an argo of "I" and arg2 of "it". Jae Eun's interpretation is that the speaker is informing the listener that the speaker is not "counting" on themself to tell the customers to "try it". My perspective is that the speaker is saying the customers should not "count" on this *thing* to be satisfactory after trying it on.

Sentence 10: We agreed on everything for this sentence, with a count.01 sense and an argo of "They" and arg1 of "the cans in the trash".

We used Cohen's Kappa Inter-Annotation agreement, which is the agreement between 2 annotators while taking into consideration the possibility of chance agreement. The yielding result from our calculation/code was 0.4595, or 45.95% . If you work with the Green, 1997 scale for the interpretation of Kappa, our result of 0.4595 gives a "fair/good" agreement.

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}_{1}$$

3 VG: WordNet

Polysemy is the when a single word has various meanings contained in it while those meanings are related. The word "book" can be an example, with it's meanings of a collection of pages or an action you

[1] Pr(a): actual agreement, Pr(e): expected agreement

can take to record something (i.e. booking a restaurant/hotel). Hyperonomy is a semantical relationship whereby a superordinate is part of, or represents, a superior order or category within a system of classification, and is opposite to the hyponymy. An example can be the word "look" can be the hyperonomy to words such as "view", "stare", or "peer". Conversely, hyponymy is the semantic relationship whereby a subordinate is part of, or represents, a lower order or category within a system of classification, and is opposite to the hyperonomy. Examples could be "pheasant", "falcon", or "owl" being a hyperonomy of "bird". Holonymy is the relationship between a word or phrase referring to a whole, of which a smaller part is the meronym (later defined). An example is "apple tree" being a holonym of "apple". Meronymy is a word that expresses a part or member of some whole thing. For example, an "apple" is a meronym of "apple tree". Troponomy is the existence "manner" between two lexemes. The words "gnaw" and "devour" can be each be a troponym of "eat", where "gnaw" is characterized by, and can be defined as, persistent nibbling. Whereas "devour" is understood as as eating something incredibly quickly and in almost a desperately hungry fashion.

These are all incredibly abstract words/concepts, and my own categorization of these words fall incredibly short of the WordNet entries for them. Abstraction by itself has four different direct hyponyms, one of which, "a personified abstractoin that teaches", I wouldn't have considered in my categorization of "abstraction"'s hyponyms. The same follows for "ambiguity" and "understanding". My categorization would be better for tangible concepts such as "tree" even though it contains, and stems out to, dozens upon dozens of direct hyponyms. This is because I believe "tree" and the other given words in the instructions ("paper" and "pillow") are better organized, not because the abstract words are poorly organized, but because concrete words are so well defined and "tied" to derived words people use, so it's easier to categorize such semantic relationships. However there is the potential for even these concrete words to have entries that I wouldn't have considered which are rarely used in either everyday life or even in literature.