Swarthmore College: NLP Honors Exam

Examiner: Jordan Kodner, Stony Brook University

Instructions and additional info

- You may not use a computer, calculator, notes, textbooks, or any other resources during this exam.
- Please answer with prose, equations, or calculations as necessary.

Additional information

1.1.4 Sensitivity and specificity are two common metrics for evaluating medical diagnostics.

\[
\text{Sensitivity} = \frac{\text{True Neg}}{\text{True Neg + False Pos}}
\]

\[
\text{Specificity} = \frac{\text{True Pos}}{\text{True Pos + False Pos}}
\]

2.1.1 Introduction to The Instructions of Shuruppak

In those days, in those far remote days, in those nights, in those faraway nights, in those years, in those far remote years, at that time the wise one who knew how to speak in elaborate words lived in the Land.

2.2.1 Two mathematically equivalent calculations

\[
\sum_{w \in w} \log(P(w)) = \log(\prod_{w \in w} P(w))
\]

3.2.1 An ambiguous sentence

Alice ate the egg salad with celery sticks

1 Evaluation

1.1 Metrics

1. Provide the formulae for accuracy, precision, recall, and F-measure.

2. Provide prose explanations of accuracy, precision, recall, and F-measure.

3. Precision and recall provide a trade-off. When might one want to optimize for precision even at the expense of recall, and when might one want to optimize for recall even at the expense of precision? Answer in terms of three concrete NLP tasks. For example, how would a high-precision/low-recall spam filter behave compared to a low-precision/high-recall one?

4. In medical diagnostics, it is common to report sensitivity and specificity instead of precision and recall. Their definitions are provided in the info box under 1.1.4. Compare and contrast these with precision and recall. In what way are these the same and in what ways are they different? Why do you think medicine prefers these over precision and recall? Show your reasoning.
1.2 Data Splits

1. The exact ratios 80/20 and 90/10 which are often used for training and test set splits are somewhat arbitrary, nevertheless there is good reasoning behind them. What is the trade-off as the ratio of training to testing data increases? Answer in terms of performance and generalizability.

2. What is meant by “train-on-test?” What are the consequences?

3. Data is sometimes divided into three sets instead of two: train, dev, and test. In this case, how are each of the three sets meant to be used? What problem does the three-way split solve relative to the two-way split?

4. Describe k-fold cross-validation. When might one choose this over a basic data split? What are its advantages?

2 Managing Sparsity in Language Models

2.1 Sparsity

1. Explain the difference between token frequency and type frequency. Provide the word type and token frequencies for the sentence given in the info box under 2.1.1. Enumerate the assumptions you made during tokenization.

2. Zipfian and other related distributions contribute majorly to the sparsity problem in NLP. Describe the Zipfian distribution in prose. What are some of its implications? Answer in terms of three concrete NLP tasks.

3. Does sparsity increase or decrease when n-gram width increases? If there are $|v| = 1000$ unigram types in some corpus, how many bigrams and trigrams are there? What is the formula for arbitrary $n$?

4. What is the Markov assumption? Please describe it generally in prose and provide a formula for an n-gram language model. What does it have to do with sparsity?

2.2 Accommodating sparsity

1. Why do we use log probabilities instead of regular probabilities? What invariants are upheld when converting from probabilities to log probabilities? Two mathematically equivalent expressions for converting a product of probabilities to log probabilities are provided in the info box under 2.2.1. Which one should be implemented in code, and why?

2. Describe smoothing in prose. Provide a formula for computing Laplace-smoothed (add-$\alpha$-smoothed) log probabilities. What is the minimum probability that an item can have in terms of $\alpha$?

3. Describe interpolation in prose. Provide a formula for linear interpolation of trigrams, bigrams, and unigrams. What is the purpose of the $\lambda$ coefficients, and what constraints must they obey?

4. OOV or UNK words can appear both during training and at test time. Describe why an OOV word might appear during test. Describe how OOV may be used during training. In your answer, be sure to explain how OOV is useful for managing sparsity during training.

3 Processing Sentences

3.1 Tagging

1. What would be a reasonable baseline model for a supervised part-of-speech tagger? Describe the model and make an educated guess on how well it should perform on an English news data set.

2. The Penn Treebank (PTB) tag set is about four times the size of the Universal Dependencies tag set. Why is the PTB tag set so large? What kind of extra tags does it contain, and why might these be useful?

3. Supervised part-of-speech tagging is typically approached as a sequence labeling problem. Unsupervised tagging is normally approached as a classification problem instead. Why do you think this is? Describe in general terms how one might approach unsupervised tagging.
3.2 Parsing and Translating

1. Ambiguity is one of the primary challenges in constituency parsing. The sentence in the info box under 3.2.1 has a structural ambiguity caused by prepositional phrase attachment. Describe the ambiguity and both of the sentence's meanings. Draw a tree for each meaning. The trees do not need to be detailed. The attachment of the PP is what matters.

2. Syntactic parsing is often accomplished through one of many dynamic programming algorithms. Why is dynamic programming used here? It may be helpful to contrast it with greedy and brute force approaches.

3. Most traditional approaches to machine translation are word-to-word systems. Discuss in practical terms why this might be the case. How can parsing, part-of-speech tagging, and other structural information be incorporated into MT? Why are they helpful?