NLP 2023

Study Guide: **ANSWER KEY**

1) List three characteristics of human language that make it distinct from other forms of animal communication. Give specific linguistic examples for each characteristic you list.

Compositionality: A meaning of an expression in natural language is the sum of the meanings of its parts (words) and the way they are combined (syntax). Every sentence is compositional, and allows us to create new utterances we have never heard: e.g. a child producing a new sentence like "I want a cookie" is compositional.

Creativity: Humans can constantly create new words and sometimes expressions. Normally this happens with younger generations. E.g. the word "selfie." These rules follow syntactic principles: e.g. a noun "selfie" is found in the same position as other nouns.

Arbitrariness: Form does not reflect meaning. E.g. the word "cat" has nothing to do with the concept cat. One can see this by how it is translated into other languages.

Displacement: Humans use natural language to talk about things that are not located in the (spatiotemporal) present. E.g. any expression with *I would like…, It might be…, I wish…, if only…*. Humans can talk about the past, too, using this property.

2) Which example below is NOT an example of ambiguity?

1. A word has more than one meaning.
2. **Two words have the same meaning but are used by different sociological groups.**
3. A sentence has two different semantic interpretations.
4. A sentence has two different syntactic interpretations.

Questions 3-8 refer to the following two example sentences. The POS tags have been provided automatically.

I. ***magazines often publish articles about beauty lies .***
POS: NOUN, ADV, VERB, NOUN, ADP, NOUN, NOUN, PUNCT

II. *beauty lies **in the eye of** the **beholder** .*
POS: NOUN, NOUN, ADP, DET, NOUN, ADP, DET, NOUN, PUNCT.

3) How many tokens (including punctuation) do these two sentences contain? **17**

4) How many types (including punctuation) do these two sentences contain? **13 (see bolding above)**

5) What is the maximum number of bigrams that can be obtained from the first sentence? **Total number of words – 1 = 8 - 1 = 7 (this includes punctuation; it will be clear on exam whether to include it or not).**

6) What is the maximum number of trigrams that can be obtained from the first sentence? **Total number of words – 2 = 8 - 2 = 6 (this includes punctuation; it will be clear on exam whether to include it or not).**

7) For some words, the lemma differs from the surface form. For only one of the answers, this is true for **both** words. Which one?

1. often, articles
2. **lies, magazines → lie, magazine**
3. lies, publish
4. beauty, beholder

8) One of the words has received the WRONG POS tag. Which one?

1. *lies* in sentence 1
2. ***lies* in sentence 2 → should be VERB**
3. *about* in sentence 1
4. *beauty* in sentence 2

9) What are common sources of bias in the NLP pipeline?

Generally, (1) the data (sampling and representation of genre, style, domain), (2) the annotation process (who the annotators are, how they annotate, what the annotation guidelines are), (3) the input representations (e.g. word embeddings, if they contain bias linked to physical or societal aspects of identity), (4) the models (e.g. through loss objectives in training, which may overamplify biases), and (5) the research design (or how we conceptualize our research) (e.g. focusing on English, looking at IQ measures as correlates for personality). A nice overview of this is provided in this article[1], as well as the slides from Luis' guest lecture (on Canvas).

10) List the common steps in the NLP pipeline in terms of linguisIc processing (e.g. text normalizaIon, POS tagging, syntax, etc.). Give examples of how each stage may be challenging.

There are many ways of conceptualizing this; here is one. I recommend walking through this pipeline with a task such as sentiment analysis and elaborating each step more specifically. Questions indicate challenges:

a. **Explore data and define task:** What genre is it? How many words are there? What are token counts? Is the data bias?
b. **Clean and preprocess data** with: tokenization, text normalization, POS tagging, NER, etc. How do we define words, sentences? What do we do with punctuation, contractions? Do we need to normalize characters if e.g. containing diacritics related to a specific script or dialect?
c. **Parsing**: analyze grammatical structure of a sentence and assign syntactic structure. Are the structure predicted with highest probability the correct ones? Are there ambiguities?
d. **Train/test split:** training set is designed to teach model and is where we set values for weights etc.. Is train set representative of data? Are there certain features I want to correct for in my splits?
e. **Feature engineering:** What features are important to my task, and how do I encode them into vectors?
f. **Modeling/learning algorithm:** Which learning algorithm am I using for my task and why?
g. **Evaluation:** What metric am I using to evaluate my model performance (e.g. F1, accuracy, BLEU)? Will this measure what I want it to measure for my task?

11) What is the language modeling strategy used by BERT? For which tasks is it helpful? For which tasks is it not helpful?

For this question you should review (i) bidirectional training of a Transformer model, (ii) attention, and (iii) masked language modeling. In addition to the slides, here is a helpful blog post on the topic: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

12) Describe the cognate facilitaIon effect in bilingual speakers. How does this effect translate into model training and performance?

For this question, you should be able to define "cognate" and see the slides on Multilingual NLP from Day 12.

---

1 https://compass.onlinelibrary.wiley.com/doi/10.1111/lnc3.12432#:~:text=We%20outline%20five%20sources%20where,how%20we%20conceptualize%20our%20research).

13) Give two examples of sequence labeling tasks in NLP. How are these tasks evaluated?

Two examples are POS tagging and NER, which you should be able to explain at a high level. POS tagging is evaluated with accuracy, which you should be able to calculate. NER is evaluated with F1 score, which you should understand how to calculate and what it measures.

14) Discuss the primary evaluaIon metric used for MT in terms of how it operates. Why is this used instead of accuracy or F1?

BLEU score; see slides from class on MT (Day 11 and Day 12).

15) Refer to the confusion matrix below that show model results for a *multiclass* (error in original guide) classification task on senIment analysis.

		Predicted value		
		positive	negative	neutral
Gold labels	positive	863	1343	193
	negative	585	3710	541
	neutral	26	245	7003

Calculate: (i) accuracy; (ii) precision; (iii) recall.
Given the results, explain F1 metric and discuss how it offers different insight into model performance from accuracy.

See slide 32-35 from Day 3 lecture for overview of this.

Background: True positives (TPs) are those where the values of the gold labels and the predicted values align (e.g. positive/positive, negative/negative, neutral/neutral). TPs here = 863 (positive) + 3710 (negative) + 7003 (neutral) = 11,576. False positives (FPs) are where the model predicted a value for incorrect gold labels (e.g. where model predicted *positive* for *negative and neutral*). For *positive*, FP = 585 + 26 = 611

Accuracy = # correct / # total, or (TP + TN) / (TP + TN + FP + FN). You need to calculate this for each class**.**
For *positive*,

TP = 863 **actual value and predicted value are same*
TN = 3710 + 541 + 245 + 7003 = 11,499 **sum of values of all columns and rows except positive*
FP = 585 + 26 = 611 **sum of column of predicted positive minus TP*
FN = 1343 + 193 = 1536 **sum of row of gold positive minus TP*
So **Accuracy =** (863 + 11,499) / (863 + 11,499 + 611 + 1536) = 12,362 / 14,509 = **.85**

Precision = true positives / (true positives + false positives) = 863 / (863+611**) = .59**
Recall = true positives / (true positives + false negatives) = 863 / (863+1536) **= .36**

F1 offers better insight about model performance based on its *quality*, being the harmonic mean of precision and recall . This is especially important for a problem with a class imbalance. We want to know the # and type of prediction errors made.

16) Read the tweet below with accompanying descripIon American English (AAE). Is this an example of prescripIve or descripIve linguisIcs? In general terms, how would a model trained on Standard American English (SAE) perform when tested on AAE? Give specific examples of where the model may struggle.

Descriptive linguistics. Examples may include how models will struggle to generalize to SAE when trained on AAE since finna and other variants is out of vocab.

17) What is the funcIon of weights and biases in a neural network? How do they represent how a network *learns* or *knows* informaIon related to the input and output?

Function: Weights are the real values that are attached with each input/feature and convey the relative importance of that corresponding feature in predicting the final output (this is often is what is difficult to interpret post-hoc). Weights control the signal (or strength of the connection) between two neurons. As an input enters the node or neuron, it gets multiplied by a weight value and the resulting output is either observed, or passed to the next layer in the neural network. Bias is a constant which is added to the product of input/features and weights. It is used to offset the result. It helps the models to shift the activation function towards the positive or negative side.

In terms of learning: Weights and biases are both learnable parameters inside the network that represent some kind of "knowledge" the network has. A neural network will randomize both the weight and bias values before "learning" initially begins. As training occurs, both weights and bias are adjusted toward the desired values and the correct output. The two parameters differ in the extent of their influence upon the input data. Weight affects the amount of influence a change in the input will have upon the output. A low weight value will have no change on the input, and alternatively a larger weight value will more significantly change the output. Bias represents how far off the predictions are from their intended value and make up the difference between the function's output and its intended output. A low bias suggest that the network is making more assumptions about the form of the output, whereas a high bias value makes less assumptions about the form of the output.