

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.

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

- a. A word has more than one meaning.
- b. Two words have the same meaning but are used by different sociological groups.
- c. A sentence has two different semantic interpretations.
- d. 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?

4) How many types (including punctuation) do these two sentences contain?

45) What is the maximum number of bigrams that can be obtained from the first sentence?

6) What is the maximum number of trigrams that can be obtained from the first sentence?

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. A often, articles
- 2. B lies, magazines
- 3. C lies, publish
- 4. D beauty, beholder

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

- 1. A *lies* in sentence 1
- 2. B *lies* in sentence 2
- 3. C *about* in sentence 1
- 4. D *beauty* in sentence 2

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

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

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

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

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


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

15) Refer to the confusion matrix below that show model results for a binary sentiment analysis classification task.

		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.

16) Read the tweet below with accompanying description describing a linguistic construction in African American English (AAE).¹ Is this an example of prescriptive or descriptive linguistics? 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.

 **Linguistics Girl MLS MS** @LinguisticsGirl · Mar 4, 2020
 <finna> is very much a word.

From 'This Language, A River: A History of English' by Dr. K. Aaron Smith and Dr. Susan M. Kim: amzn.to/2PK3rMO

THE MODERN PERIOD AND GLOBAL ENGLISHES 287

Another feature of AAE is the future auxiliary *finna*. *Finna* has its roots in Southern US English where it is still frequently used by most speakers in its fuller form *fixin' to*, or any number of reduced variants, *fissa*, *fitna*, etc. The form *finna*, however, appears to be more specifically the form in AAE and as such can be heard in almost every variety of AAE throughout the US. The development of *finna* from *fixing to* is phonologically analogous to the development of *gonna* from *going to*.³⁹

Fixing to originally meant arranging things in such a way to create an outcome, but as *fixing to* came to mark a future prediction, the two words fused. In many ways, *finna* has become more grammatical, and certainly more frequent than *fixin' to* and other reduced variants, suggesting that the system of future auxiliaries in AAE has become very complex, a system in which very subtle differences between *will*, *gonna* and *finna* can be expressed. Another even newer futurate auxiliary has emerged within the verb system of AAE, *tryna* (from *trying to*), that does not simply mean *attempt* and instead seems to signal yet another level of future prediction or intention.

Certain aspects of the linguistic structures we have been looking at so far show unique patterning in AAE, like the use of *ain't* as a present-tense auxiliary and the actual form *finna*. However, in another sense, the constructions that are involved in that patterning are certainly shared with other varieties of English. We have already discussed the broad usage of *ain't* in many Englishes and most non-AAE Southern US varieties use *fixing to*, or some reduced variant of it. While isogloss overlap is no more surprising in this instance than it is in any other instance of identity or geographical boundary, scholars have used facts like these to evaluate the origins of AAE.

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

¹ (Full link to thread here: https://twitter.com/Javy_wealth/status/1233894153265844224?s=20)