




LIVE SESSION 10

NPTEL NLP

(NOC24_CS39)

Shubhi Bansal
PMRF Scholar
IIT Indore



Q1

Common steps of entity linking are:

- a. Reference disambiguation->Candidate Selection
- b. Reference disambiguation->Candidate Selection-> Mention identify
- ~~c. Mention identify ->Candidate Selection-> Reference disambiguation~~
- d. All of the above

Named Entity Recognition (NER)
entity : name / place / organisation

Solution for Q1

□ Ans: C

Q2

The text span s = "Sea" occurs in ~~600~~ different Wikipedia articles.

- c1: 223
- c2: 161
- c3: 18
- c4: 11
- No Link: 187
- ⇒ □ Calculate the keyphraseness of "Sea".

A	0.232
B	0.886
C	0.688
D	0.976

acc. to above info"

Q2

- ☐ A) 0.232
- ☐ B) 0.886
- ☒ C) 0.688
- ☐ D) 0.976

(Ans.)

$$\text{Keyphraseness} = \frac{CF(w_i)}{CF(w)} = \frac{\text{No. of times } w_i \text{ is linked to other wikipedia art.}}{\text{frequency of } w \text{ (how many } w \text{ appears in wikipedia art.)}}$$

Solution for Q2

□ Solution: C)

□ $CF(si) / CF(s) = 223 + 161 + 18 + 11 / 600 = 413 / 600 = 0.688$

no. of times linked to other all wikipedia art.

$$\begin{array}{r} \text{freq(w)} \\ = 223 + 161 + 18 + 11 \\ \hline 600 \end{array}$$

Q3

$s = \text{"sea"}$
 $\text{freq}(s) = 600$

□ What is the commonness of (s,c2) in the above question?

a. 0.765

b. 0.389

c. 0.453

d. 0.910

(Ans.)

$c1 : 223$
 $c2 : 161$
 $c3 : 18$
 $c4 : 11$
No links: 187

$$\frac{|L_{w,c}|}{\sum_w L_{w,c}}$$

Total mentions

$$\text{commonness}(S, c_2) = \text{count}(\text{"sea"}, c_2)$$

$$\frac{|L_{w,c}|}{\sum_{i=1}^n |L_{w,c_i}|} = \frac{\text{No. of times } w \text{ is linked with } c_2}{\text{Total no. of times } w \text{ is linked with any } c}$$

(c₁, c₂, c₃, c₄)

$$\frac{= 161}{223 + 161 + 18 + 11} = \frac{161}{413} = 0.389$$

Solution for Q3

Ans: B)

Solution:

$$161 / (223 + 161 + 18 + 11) = 161 / 413 = 0.389$$

Q4

The text span s ='world' occurs in 764 different Wikipedia articles.

- ☐ C1: 189
- ☐ C2: 273
- ☐ C3: 87
- ☐ C4: 53
- ☐ No link: 162

Calculate keyphraseness of "world"

- a. 0.232
- b. 0.788
- c. 0.688
- d. 0.976

$$\text{Keywordseness} = \frac{CF(w_i)}{CF(w)}$$

$$CF(w)$$

No. of times 'w' is linked
to other wikipedia articles

No. of times 'w' appears in
all wikipedia articles

$$189 + 273 + 87 + 53$$

$$764$$

$$= \frac{764 - 162}{764}$$

$$= 0.7879 \approx \underline{\underline{0.788}}$$

Q5

What is the commonness of (s, c2) in the above question?

a. 0.765

b. 0.389

c. 0.453

d. 0.910

Ans

$$\frac{L_{w,c}}{\sum_{i=1}^n |L_{w,c_i}|} = \frac{L_{w,c_2}}{L_{w,c_1} + L_{w,c_2} + L_{w,c_3} + L_{w,c_4}}$$

$$= \frac{273}{189 + 273 + 87 + 53} = \frac{273}{602} = 0.45348$$

Q6

(MCO \rightarrow MSO)

Relevant feature/s for a supervised model for predicting the topics to be linked is/are

- a. Disambiguation Confidence
- b. Relatedness
- c. Link Probability
- d. All of the above

Solution for Q6

- The answer is All of the above. Here's a breakdown of why each feature is relevant for a supervised topic linking model:
- **Disambiguation Confidence:**
- Topics and entities often have ambiguous names (e.g., "Apple" could refer to the company or the fruit).
- A high disambiguation confidence score indicates the model is certain about the correct meaning/interpretation of the topic, reducing the risk of linking to the wrong concept.
- **Relatedness:**
- Measures the semantic similarity or connection between the potential topic and the surrounding text.
- A strong relatedness score suggests that the topic is contextually relevant, increasing the likelihood of a valid link.
- **Link Probability:**
- Indicates the probability that a particular topic should be linked within the given text.
- This can be learned from existing knowledge bases or training data, providing a direct measure of linking suitability.

Q7

Which of the following is an advantage of unsupervised relation extraction?

- ☒ a. Can work efficiently with small amount of hand-labeled data
- ☐ b. Not easily generalizable to different relations
- ☒ c. Need no training data
- d. Always perform better than supervised techniques

→ semi-supervised

→ supervised

Solution for Q7

- Answer: C
- Unsupervised relation extraction's core advantage: Unsupervised techniques don't rely on manually labeled training data. This is crucial when labeled data is scarce, expensive to create, or when dealing with new and emerging relations.
- Let's analyze the other options:
- Can work efficiently with a small amount of hand-labeled data: This is actually an advantage of semi-supervised techniques, not unsupervised ones. Semi-supervised methods leverage a small labeled dataset along with large amounts of unlabeled data.
- Not easily generalizable to different relations: This can be a disadvantage of both unsupervised and supervised relation extraction models. Highly specialized models might struggle to adapt to new relation types.
- Always perform better than supervised techniques: This is incorrect. Supervised techniques, when trained with sufficient high-quality data, generally outperform unsupervised methods in terms of accuracy.

Q8

gmp

Which of the following is not a Hearst's Lexico Syntactic Patterns for automatic acquisition of hyponyms?

- ☒ a. X or other Y (most valid)
- ☒ b. X and other Y
- ☒ c. Y including X
- ☒ d. X but not Y

Hyper

vehicle

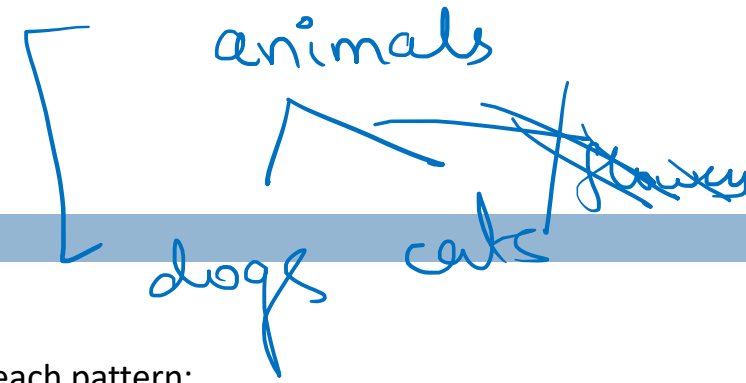
scooter / bikes

Hypon

cars trucks

(a, b, c \Rightarrow valid Hearst's lexico syntactic pattern)
d \nRightarrow valid

Solution for Q8



□ Answer: D

Hearst's Lexico-Syntactic Patterns focus on identifying hyponym-hypernym relationships. Let's break down each pattern:

□ X or other Y:

a □ Example: "fruits such as apples or other citrus"

□ Hyponym: apples, other citrus fruits

□ Hypernym: fruits

□ X and other Y:

b □ Example: "vehicles like cars and other automobiles"

□ Hyponym: cars, other automobiles

□ Hypernym: vehicles

□ Y including X:

c □ Example: "flowers including roses and tulips"

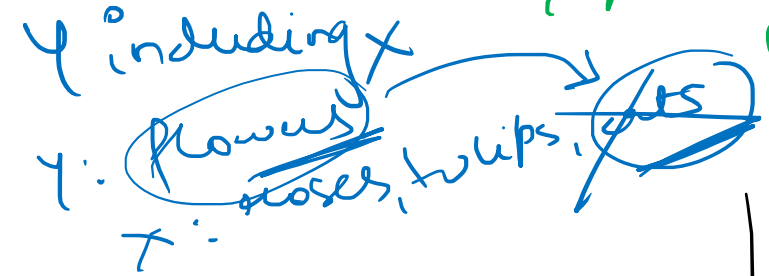
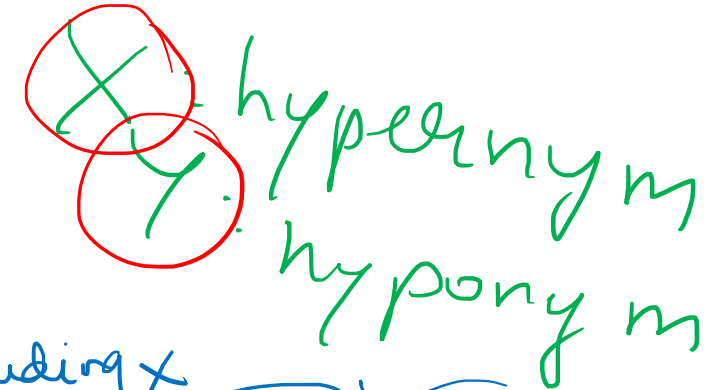
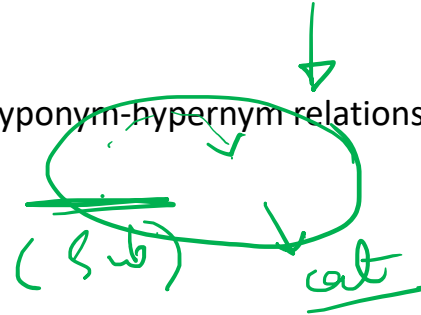
□ Hyponym: roses, tulips

□ Hypernym: flowers

□ The "X but not Y" pattern does not express a clear hyponym-hypernym relationship. Here's an example to illustrate why:

□ "I like cats but not dogs" - This sentence implies a preference rather than a hierarchical relationship between the categories of animals.

□ Remember: Hearst's patterns are powerful because they rely on specific linguistic structures that strongly signal a hyponym-hypernym relationship.



Q9

Consider a dataset with a very low number of relations - all of which are very important. For a relation extraction task on that dataset, which of the following is the most useful metric?

a. Precision

b. Recall

c. Accuracy

d. F1-Score

⇒ Harmonic mean of precision & recall
$$F1\text{score} = \frac{2 * \text{precision} * \text{recall}}{(\text{precision} + \text{recall})}$$

Solution for Q9

□ Answer: B

Solution for Q9

TP, FP, TN, FN

Recall =

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

- Ans: Recall
- Importance of Finding All Relations: Since all the relations are highly important, missing even a single one would be highly detrimental. Recall measures how well the model finds all of the true positive relations in the dataset.
- Precision vs. Recall in this scenario: While precision matters (you want the predictions to be correct), recall becomes the priority when you cannot afford to miss true relations, even if it comes at the cost of slightly lower precision.
- ✓ □ F1-Score: The F1-score is a harmonic mean between precision and recall. While useful in many scenarios, in this specific case, it doesn't put enough emphasis on finding all the important relations.
- Why Accuracy is less important: Accuracy measures the overall correct predictions. In datasets with a low number of relations, even if the model misses important relations but is correct on the majority of the non-relation cases, the accuracy could be misleadingly high.
- Key takeaway: When choosing metrics, it's essential to align them with the priorities of your specific task and dataset. In this case, recall should be prioritized.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

~~Pos~~ → true
(-ve)
False

(+ve)	(-ve)
TP	TN
FP	FN

★ ★
Evaluation metrics

Q10

What is KeyPhraseness (wikipedia)?

- a. Number of articles that mention a key phrase divided by the number of wikipedia articles containing it.
- ~~b.~~ Number of Wikipedia articles that use it as an anchor, ^{link} divided by the number of articles that mention it at all.
- c. Number of articles that mention a key phrase times by the number of wikipedia articles containing it.
- d. Number of Wikipedia articles containing the key phrases times by number of articles mentioning it.

Solution for Q10

□ Answer: B

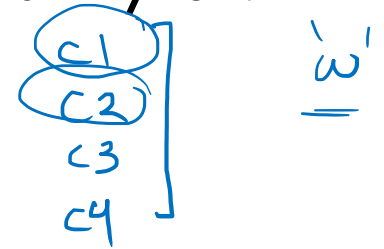
Keyphraseness = No. of times

$$\frac{CFLW_i}{CFLW}$$

Q11

Higher value of keyphraseness represents a higher probability of:

- a. ✓ An article to be selected as linkable candidate
- b. A phrase to get detected as a mention
- c. An article to be disambiguated from other candidates
- d. None of the above



Keyphraseness = $\frac{\text{No. of times } \omega \text{ is linked to other articles}}{\text{frequency of } \omega \text{ (total no. of times } \omega \text{ appears across all wikipedia articles)}}$



Q12

Which of the following problem exists in bootstrapping technique for information extraction?

- a. ✓ Sensitiveness towards the seed set
- b. High precision
- c. Less manual intervention
- d. All of the above

(Ans.)

seed data

Solution for Q12

Delhi: India
~~France: Paris~~

- Bootstrapping is a powerful technique for starting information extraction when you lack a large labeled dataset.

★ □ Sensitivity to Seed Data: The quality of the initial seed examples (a few correct patterns or examples of what to extract) heavily influences the success of bootstrapping. Poor seed data can lead the system down an incorrect path from the beginning.

□ Mitigation

How to prevent

- ★ □ Careful Seed Selection: Choosing high-quality seed data is crucial to ensure the system starts on the right track.

King: man;
Queen: woman

(correct info)
so that we can learn to identify patterns correctly

Q13

Hypernym hyponym
relationship

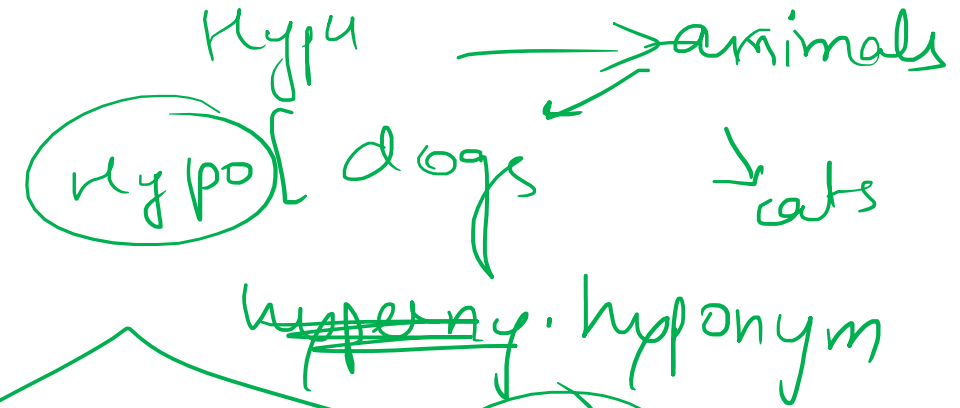
□ Which of the following is a Hearst's Lexico Syntactic Patterns for automatic acquisition of hyponyms:

a. X or other Y

b. X and other Y

c. Y including X

d. X but not Y



I like dogs but not cats

I like animals but not cats

I like cats but not animals

Solution for Q13

- Ans: X or other Y

This pattern strongly indicates a hyponym-hypernym relationship:

- X: Represents a potential hyponym (more specific term).
- other Y: Represents a set of terms belonging to the broader category Y (hypernym).

Solution for Q13

- Hearst's Lexico-Syntactic Patterns are used to identify hyponym-hypernym relationships. A hyponym is a more specific word; the hypernym is a broader category. Examples:
- X or other Y: "fruits such as apples, oranges, or other citrus" (hypernym: fruits, hyponym: apples, oranges), "vehicles like cars, trucks, or other motor vehicles" (hypernym: vehicles, hyponym: cars, trucks)
- Other common Hearst Patterns:
- Y such as X: "mammals such as cats, dogs, and elephants" (hypernym: mammals, hyponym: cats, dogs)
- X and other Y: "primates like humans, monkeys, and other apes" (hypernym: primates, hyponym: humans, monkeys)
- Y, including X: "European countries, including France, Germany, and Spain" (hypernym: European countries, hyponym: France, Germany)
- Important Note: While these patterns are a good starting point, they can sometimes lead to incorrect extractions due to the ambiguities of natural language.

Q14

Advantage of distant supervision over bootstrapping method

- a. Need more data]
- b. Less human effort]
- ☒ c. Can handle noisy data better
- d. No advantage

Solution for Q14

- Ans: Less human effort
- Distant Supervision: Leverages existing knowledge bases (like Freebase or Wikidata) to automatically label a large corpus of text. This significantly reduces the need for manual labeling of training data.
- Bootstrapping: Relies on a few correct seed examples and iteratively extracts more patterns and examples. This still requires some initial human effort to create those quality seeds, and often some intervention during the process.
- Need more data: This might be slightly true in some cases, as distant supervision often works better with larger text corpora to compensate for potential noise in the knowledge base. However, it's not the primary advantage.
- Can handle noisy data better: While distant supervision can be more robust to some noise, it's not inherently better than bootstrapping in this regard. Both methods can suffer from incorrect data.
- No advantage: This is simply incorrect. The core advantage of distant supervision is the reduced need for manual labeling effort.

Q15

- Bootstrapping can be considered as:
 - a. Supervised Approach
 - b. Unsupervised Approach
 - c. Semi-supervised Approach
 - d. All of the above
 - e. None of the above

Solution for Q15

- ❑ The correct answer is the Semi-supervised Approach. Here's why:
- ❑ Bootstrapping's nature: Bootstrapping is a hybrid approach that leverages both supervised and unsupervised learning aspects:
- ❑ Supervised aspect: It starts with some seed data (labeled examples) and iteratively refines its extraction rules in a supervised manner.
- ❑ Unsupervised aspect: Using these rules, it finds new patterns and examples in unlabeled text, expanding its knowledge in an unsupervised fashion.
- ❑ Why not the other options:
- ❑ Supervised Approach: Supervised methods rely heavily on labeled data, while bootstrapping's appeal is its ability to learn from limited labeled examples.
- ❑ Unsupervised Approach: Unsupervised methods don't use any labeled data, while bootstrapping heavily relies on those initial seed examples.
- ❑ All of the above/None of the above: These are inaccurate. Bootstrapping clearly falls into the semi-supervised category.

Q16

Distant supervision primarily addresses the challenge of:

- a. Lack of labeled data
- b. Computational complexity in NLP models
- c. Ambiguity in natural language
- d. The need for real-time model updates

Solution for Q16

- Ans: A
- Reason: Distant supervision is a core technique for creating labeled training data when manual labeling is expensive, time-consuming, or impossible at scale. It leverages existing knowledge bases or heuristics to automatically generate noisy (potentially imperfect) labels for unlabeled data.

Q17

Which of the following is a key assumption in distant supervision?

- ✓ a. Knowledge bases ~~are~~ always error-free ←
- b. Text containing mentions of known relationships likely expresses those relationships. → france: Paris ✓
- c. Manually labeled data is never required.
- d. ~~Only~~ simple grammatical structures can be reliably analyzed.

Decreasing order of suit ability

10 7^a ✓

Solution for Q17

city - country
[Delhi : India
→ [France : Paris
country - city

□ Ans: B

Text containing mentions of known relationships likely expresses those relationships.

□ Reason: Distant supervision rests on the idea that if a sentence contains two entities known to have a certain relationship (e.g., "Mumbai" and "India" have a "city-country" relationship), then that sentence is likely to express that relationship.

Q18

In a bootstrapping approach for named entity recognition, a likely initial step would be to:

- a. Train a model on a small seed set of labeled entities.
- b. Create a comprehensive knowledge base of all possible entities.
- c. Run the model on a large, unlabeled corpus and discard low-confidence outputs.
- d. Manually correct all errors produced by the model.

Solution for Q18

Answer: A) Train a model on a small seed set of labeled entities.

Reason: Bootstrapping is an iterative process. It starts with a small set of manually labeled examples (or a few patterns/rules) and uses these to find similar examples in unlabeled data. This expanded set is then used to retrain the model, and the process repeats.

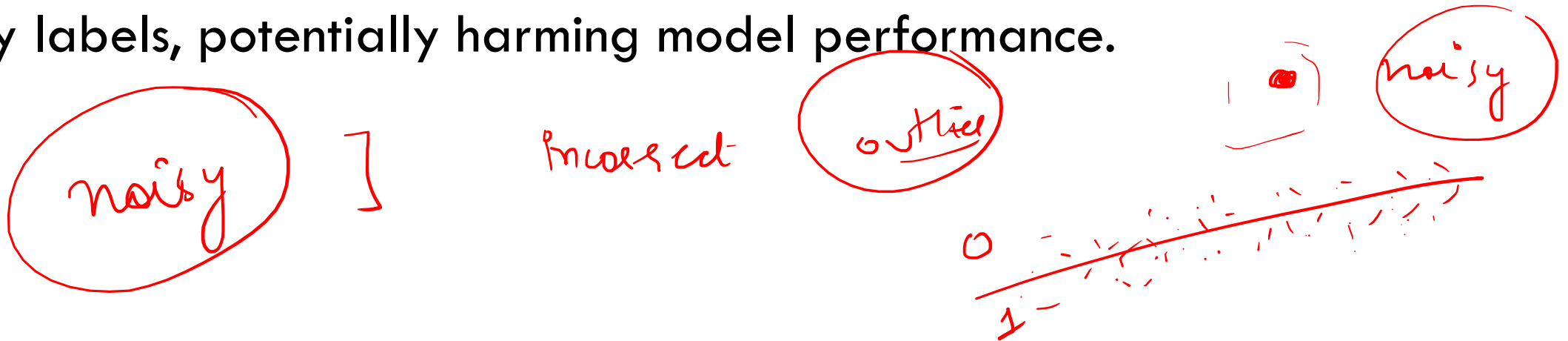
Q19

A potential issue with distant supervision is:

- a. Excessive computational time requirements
- b. Over-reliance on complex linguistic features
- c. Noisy labels introduced into the training data
- d. Inability to handle figurative language

Solution for Q19

- Answer: C) Noisy labels introduced into the training data
- Reason: While distant supervision accelerates labeling, the automatically generated labels aren't perfect. Errors from the knowledge base or incorrect assumptions about the text can lead to noisy labels, potentially harming model performance.



Q20

When choosing between bootstrapping and distant supervision, a key consideration might be:

- a. The availability of a pre-existing knowledge base.
- b. The desired precision vs. recall trade-off.
- c. Whether the NLP task involves open-domain relations.
- d. All of the above.

Solution for Q20

- Answer: D) All of the above.
- Reason:
- Knowledge Base: Distant supervision often relies on an existing knowledge base, while bootstrapping can start with just a few seed examples.
- Precision vs. Recall: Bootstrapping may offer higher precision (fewer false positives), while distant supervision can improve recall (finding more true examples).
- Open-Domain Relations: Bootstrapping is better suited for discovering new or open-domain relationships, as distant supervision is bound by the knowledge base.

Q21

A novel approach to improve the robustness of distant supervision could involve:

- a. Filtering sentences based on syntactic patterns.
- b. Incorporating human-in-the-loop feedback during training.
- c. Using multiple knowledge bases for cross-verification.
- d. All of the above.

Solution for Q21

- Answer: D) All of the above.
- Reason:
- Syntactic Patterns: Filtering sentences with specific structures can reduce noise.
- Human Feedback: Incorporating human experts to confirm or correct some labels can significantly improve label quality.
- Multiple Knowledge Bases: Using multiple resources and cross-verification can mitigate errors from any single knowledge base.

TF-IDF

Q22

term frequency

= No. of times a

Keyphraseness refers to:

a.

The frequency with which a word appears in a document

b.

The overall grammatical importance of a phrase

syntax ←

meaning / semantic →

c.

How well a phrase encapsulates a document's core topics

d.

The length of a phrase in characters

→ a) Term frequency = $\frac{\text{No. of times term 't' appears in a doc}}{\text{No. of terms in the document}}$

Solution for Q22

Answer: C) How well a phrase encapsulates a document's core topics

Reason: Keyphraseness is about identifying the phrases that best represent the essential ideas or concepts within a document.

$$\frac{CF(w_i)}{CF(w)}$$

No. of times w linked to other articles

total No. of times w appears in all articles

lovely day
pleasant weather

Q23

- Incorporating commonness measures could improve topic modeling by:
 - a. Helping identify and filter out stop words
 - b. Prioritizing highly specific and informative terms
 - c. Determining the *(alphabetical)* chronological order of topics
 - d. Detecting sarcasm or figurative language

Solution for Q23

Answer: A) Helping identify and filter out stop words

Reason: Common words ("the", "of", etc.) rarely contribute to topic identification. Commonness helps filter these out, improving topic model focus.

(stop words)

common words

1DF

Q24

For a corpus of highly technical documents, commonness measures might be less informative because:

- a. Technical terms often have low document frequency.
- b. Language models are not trained on technical jargon.
- c. Keyphraseness is irrelevant in technical domains.
- d. Technical documents tend to be very short.

Solution for Q24

Answer: A) Technical terms often have low document frequency.

Reason: Specialized terminology in technical domains is often infrequent across a general corpus. This makes commonness less informative as many relevant terms would appear uncommon.

<https://www.mygreatlearning.com/blog/nlp-interview-questions/>

- **In a corpus of N documents, one randomly chosen document contains a total of T terms and the term “hello” appears K times.**
- What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term “hello” appears in approximately one-third of the total documents?
 - a. $KT * \log(3)$
 - b. $T * \log(3) / K$
 - c. $K * \log(3) / T$
 - d. $\log(3) / KT$
- **Answer:** (c)
- formula for TF is K/T
formula for IDF is $\log(\text{total docs} / \text{no of docs containing "data"})$
 $= \log(1 / (1/3))$
 $= \log(3)$
- Hence, the correct choice is $K\log(3)/T$
- **22. In NLP, The algorithm decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents**
- - a. Term Frequency (TF)
 - b. Inverse Document Frequency (IDF)
 - c. Word2Vec
 - d. Latent Dirichlet Allocation (LDA)
- **Answer:** b)