

1. Lexikalische Zähleinheiten I (10 Punkte)

1.1 Zählen Sie nun die **Token** – ohne Berücksichtigung von Satzzeichen – für den folgenden Satz:

20

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt.

(2 Punkte)

1.2 Zählen Sie nun die **Wortformen** – ohne Berücksichtigung von Satzzeichen – für den folgenden Satz:

8 7

Fliegt eine Fliege hinter Fliegen, so fliegt eine Fliege Fliegen nach.

(2 Punkte)

1.3 Zählen Sie nun die **syntaktischen Wörter** – ohne Berücksichtigung von Satzzeichen – für den folgenden Satz:

21 22

In Hausstadt steht ein Haus, hinter dem zwei Häuser stehen und auf dem einen der beiden ist zu lesen: "Das ist unser Haus!"

(2 Punkte)

1.4 Zählen Sie nun die **Lexeme** – ohne Berücksichtigung von Satzzeichen – für den folgenden Satz:

11 13

Rechtsgelehrte wissen, dass *recht haben* und *Recht bekommen* zwei sehr unterschiedliche Dinge sind.

(2 Punkte)

1.5 Zählen Sie nun die **Lexemgruppen** – ohne Berücksichtigung von Satzzeichen – für den folgenden Satz:

9

Am Fischteich "Anglersee" bei Mittelangeln angelteten viele Hobbyangler, die ihre Angeln beim Angelverleih "Anglerbedarf Unterangeln" ausgeliehen halten.

(2 Punkte)

1. Lexikalische Zähleinheiten

Sei der folgende Satz gegeben:

Wenn hinter Fliegen eine Fliege fliegt, fliegen Fliegen einer Fliege voraus.

Zählen Sie nun die Anzahl der:

Token: 13

Wortformen: 9

syntaktischen Wörter: 9 11

Lexeme: 6

Lexemgruppen: 4

1. Lexikalische Zähleinheiten (10 Punkte)

Es sei der folgende Satz gegeben:

In Hausstadt steht ein Haus, hinter dem zwei Häuser stehen;
und auf dem einen der beiden ist zu lesen: "Das ist unser Haus!"

Zählen Sie nun – ohne Berücksichtigung von Satzzeichen – die Anzahl der:

1.1 (2 Punkte)

Token:

23

1.2 (2 Punkte)

Wortformen:

20

1.3 (2 Punkte)

syntaktischen Wörter:

23

1.4 (2 Punkte)

Lexeme:

15

1.5 (2 Punkte)

Lexemgruppen:

8

<< Zurück

Seite markieren

Weiter >>

2. Skip-Gramme I (5 Punkte)

Die aus der Vorlesung bekannte Skip-Gramm-Formel umfasst für ein bestimmtes k und ein bestimmtes n alle n -Gramme mit bis zu k Skips. Wie jedoch sieht die Formel aus, die nur genau k Skips bei der n -Gramm-Erstellung erlaubt?

2.1 Wählen Sie aus den nachfolgenden Formeln diejenige aus, die **genau k Skips** bei der Erstellung eines entsprechenden Skip- n -Gramms erlaubt. (5 Punkte)

- A : $\left\{ (w_{i_1}, w_{i_2}, \dots, w_{i_n}) \mid \sum_{j=2}^n i_{j-1} - i_j - 1 = k \right\}$
- B : $\left\{ (w_{i_1}, w_{i_2}, \dots, w_{i_n}) \mid \sum_{j=2}^n i_j - i_{j-1} = n + k \right\}$
- C : $\left\{ (w_{i_1}, w_{i_2}, \dots, w_{i_n}) \mid \sum_{j=2}^k i_j - i_{j-1} - 1 = n \right\}$

<< Zurück

Seite markieren

Weiter >>

2. Skip-Gramme

Welche 2-Skip-3-Gramme kommen in dem folgenden Text vor:

Insurgents killed in ongoing fighting.

- | | |
|---|--|
| <input checked="" type="radio"/> insurgents killed in | <input checked="" type="radio"/> insurgents killed fight |
| <input type="checkbox"/> insurgents killed | <input checked="" type="radio"/> killed in ongoing |
| <input type="checkbox"/> in fighting ongoing | <input checked="" type="radio"/> in ongoing fights |

2. Lexikalische Zähleinheiten II (5 Punkte)

2.1 Differenzieren Sie zwischen "Wortform" und "syntaktisches Wort".

Geben Sie hierzu eine (kurze) Definition der beiden Begriffe und heben Sie hervor, wie sie sich unterscheiden. (5 Punkte)

Ein syntaktisches wort ist ein einzelnes wort oder token in einem input dass durch seine grammatischen funktion bestimmt wird.
Eine Wortform ist ein Teil eines syntaktischen worts und basierend Kasus, numerus genus etc differenziert (Die Fliegen fliegen den Fliegen nach -> die 2 FLiegen instanzen sind 2 verschiedene Wortformen Nom.Plural / Dat.PL) diese beinhalten nur die Ausdruckseite des Syntaktischen worts was jedoch auch mehr inhalte wie flexionen beinhalten kann.
-> WF können keinen flexionen haben, syntaktische wörter jedoch schon

<< Zurück

Seite markieren

Weiter >>

3. Masked Language Models

Es sei ein Masked Language Model (MLM) wie BERT (Devlin et al., 2018) gegeben.

Lassen sich mit einem MLM Scores für Wörter berechnen, die nicht mit einem einzelnen Token (Sub-Word) repräsentiert werden können?

Antworten Sie zunächst mit **Ja** oder **Nein** und begründen Sie dann Ihre Antwort **kurz**.

Ja, als Beispiel nutzt BERT als Tokenisierer für sein MLM Wordpiece, welches Wörter in kleinere Subwords aufteilen kann. Das MLM kann dann für die entstandenen Subwords mithilfe von Log Likelihood Berechnungen oder Wahrscheinlichkeitsmittelung (P, average) einen Gesamtscore für die maskierten Subwords erstellen. Falls kein Whole Word masking eingestellt ist für das MLM kann es trotzdem mögliche Subwords errechnen durch die klassische Methodik der Wahrscheinlichkeitsberechnung und Ausgabe über softmax des Vokabulars

Ja. BERT verwendet als Tokenisierer das WordPiece-Verfahren, das Wörter in kleinere Subwords zerlegen kann, wenn sie nicht als Ganzes im Vokabular vorhanden sind.

Das Masked Language Model (MLM) berechnet für jedes maskierte Subword eine Wahrscheinlichkeit (z.B. über Softmax). Um einen Gesamtscore für ein Wort zu erhalten, kann man die Wahrscheinlichkeiten der einzelnen Subwords kombinieren – zum Beispiel durch Multiplikation (Log-Likelihood) oder Mittelung.

Auch ohne Whole-Word-Masking kann das MLM für mehrteilige Wörter Scores berechnen, indem es die Wahrscheinlichkeiten der maskierten Subwords ausgibt und diese entsprechend zusammenführt.

3. Skip-Gramme II (12 Punkte)

3.1 Welche 1-Skip-3-Gramme kommen in dem folgenden Text – ohne Berücksichtigung von Satzzeichen – vor:

Hinweis: Je 2 Punkte pro richtiger Antwort. (6 Punkte)

In Hausen steht ein Haus, hinter dem zwei Häuser stehen.

- In Hausen steht
- Haus hinter zwei
- Hausen ein Haus

- In steht Haus
- ein Haus zwei Häuser
- stehen zwei Häuser

3.2 Welche 2-Skip-4-Gramme kommen in dem folgenden Text – ohne Berücksichtigung von Satzzeichen – vor: 6 Punkte

Hinweis: Je 2 Punkte pro richtiger Antwort. (6 Punkte)

Die Akkulaufzeit ist gut - nur das Aufladen geht zu langsam.

- Die Laufzeit ist gut
- nur Aufladen zu langsam
- gut ist das nur

- Die gut nur das
- gut das geht langsam
- Akkulaufzeit ist das Aufladen

<< Zurück

Seite markieren

Weiter >>

3. Skip-Gramme I (6 Punkte)

3.1 Welche 2-Skip-3-Gramme kommen in dem folgenden Text – ohne Berücksichtigung von Satzzeichen – vor:

In Hausstadt steht ein Haus, hinter dem zwei Häuser stehen.

In Hausstadt Häuser

In Haus steht

stehen zwei Häuser

ein Haus zwei Häuser

In steht Haus

Haus hinter zwei

Hinweis: Multiple-Choice-Frage. (6 Punkte)

<< Zurück

Seite markieren

Weiter >>

4. Sentiment Analysis

Sei ein Sentiment-Tupel definiert als (e, a, s) , wobei gilt:

- Zielentität e
- Merkmal a von e
- Sentiment s der Meinung des Meinungsträgers bezogen auf das Merkmal a von Entität e
- s kann die folgenden Werte annehmen: negativ, neutral, positiv

In der Praxis stellen wir fest, dass Sie mit dem Universe X42 grundsätzlich gleich gute Fotos schießen wie mit dem Universe Z23 Ultra.

Bei Standardfotos stimmen Schärfe und Helligkeit, auch wenn Bilder stärker nachgeschärft werden als beim Top-Modell.

Die Farben weichen teils von der Realität ab, das Z23 Ultra zeigt natürlichere Farben an.

Die Akkulaufzeit ist gut - nur das Aufladen geht zu langsam.

Unterm Strich bekommen Sie aber sehr viel geboten für Ihr Geld.

Wählen Sie die Sentiment-Tupel aus, welche im vorangegangenen Text vertreten werden.

Hinweis: Multiple-Choice-Frage.

- (Universe Z23 Ultra, Aufladen, positiv)
- (Universe X42, Akkulaufzeit, positiv)
- (Universe X42, Bildhelligkeit, negativ)
- (Universe X42, Preis-Leistungsverhältnis, positiv)

4. Skip-Gramme II (6 Punkte)

4.1 Welche 1-Skip-4-Gramme kommen in dem folgenden Text – ohne Berücksichtigung von Satzzeichen – vor:

Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt.

Hinweis: Multiple-Choice-Frage. (6 Punkte)

- fand in seinem Bett
- Bett zu ungeheueren Ungeziefer
- aus unruhigen Träumen erwachte
- zu ungeheueren Ungeziefer verwandelt
- Gregor Samsa Morgens unruhigen
- fand er sich in seinem Bett

<< Zurück

Seite markieren

Weiter >>

4. Sprachmodelle I (4 Punkte)

4.1 Eine Forscherin/ein Forscher hat herausgefunden, dass sich die Autorenschaft eines Textes durch Muster in kurzen Sequenzen von Verben in Texten feststellen lässt. Diese Muster bestehen in Form von Sequenzen von Verben der Mindestlänge 2, allerdings muss dabei manchmal ein Verb übersprungen werden und die Gesamtlänge der Sequenz darf 3 nicht überschreiten. Die Verbsequenzen können sich über Satzgrenzen hinaus erstrecken, allerdings nicht über unterschiedliche Paragaphen.

Welches der folgenden Sprachmodelle ist geeignet, um diese Muster zu erkennen?

(4 Punkte)

- Bag-of-3-Skip-1-Grams-of-Verb-Token innerhalb des gleichen Paragraphen
- Die Vereinigung von 1-Skip-2-Grams- und 1-Skip-3-Grams-of-Verb-Token innerhalb des gleichen Paragraphen
- Bag-of-1-Skip-3-Grams-of-Verb-Token innerhalb des gesamten Textes
- Die Vereinigung von 1-Skip-2-Grams- und 2-Skip-3-Grams-of-Verb-Token innerhalb des gleichen Paragraphen
- Bag-of-1-Skip-3-Grams-of-Verb-Token innerhalb des gleichen Paragraphen

<< Zurück

Seite markieren

Weiter >>

5. Sprachmodelle II (5 Punkte)

5.1 In stark flektierenden Sprachen, wie beispielsweise Deutsch, haben Wörter in der Regel viele unterschiedliche Wortformen. So listet Wiktionary 18 flektierte Wortformen für das Verb "laufen".

Sie wollen nun Word Embeddings trainieren, dessen Repräsentationen invarianter gegenüber geringfügigen Änderungen von Wortformen des selben Wortes sind, als es das word2vec Modell von Mikolov et al. (2013a; 2013b) ist. Dabei wollen Sie aber vermeiden, dass das Modell alle Wortformen eines Wortes gleich repräsentiert.

Mit welchem der folgenden Ansätze kann Ihr Vorhaben gelingen? (5 Punkte)

- Wörter werden durch Buchstaben n -Gramme repräsentiert und Word Embeddings als die Summe der n -Gram Embeddings berechnet.
- Wörter werden durch Buchstaben n -Gramme repräsentiert und Word Embeddings als die Konkatenierung der n -Gram Embeddings berechnet.
- Sie extrahieren syntaktische Informationen und konkatenieren entsprechende Syntax-Merkmal-Embeddings mit dem Embedding der Wortform.
- Sie verwenden die Lemmata der Wörter, anstelle der Wortformen.

5. Sprachmodelle II (12 Punkte)

Eine Forscherin/ein Forscher will ein spezielles *word2vec*-Modell trainieren, bei dem ausgehend von dem *CBOW*-Modell speziell Eigennamen vorhergesagt werden sollen.

Für einen gegebenen Eingabetext bestehend aus Token t_1, \dots, t_m der Länge n und den Hyperparametern α und k hat er sich die folgende Formeln und Rechenschritte ausgedacht.

Im Folgenden dürfen Sie annehmen:

- Ein Eigename ist immer genau ein Token.
- Die Tokenisierung verlief korrekt.
- Die Menge der Eigennamen ist vollständig und endlich.
- Die lediglich mit Namen genannten Funktionen funktionieren jeweils korrekt für die gegebenen Argumente.

Hinweis: bei Kprim-Fragen (5.1, 5.2, 5.3) gibt es 4 Aussagen (hier: einzelne Formeln oder Rechenschritte des Modells), die richtig oder falsch sein können. Für 4 korrekte Antworten wird die volle Punktzahl (4 Punkte) und für 3 korrekte Antworten die halbe Punktzahl (2 Punkte) vergeben. Für weniger als 3 korrekte Antworten gibt es keine Punkte. Wählen Sie +, wenn der entsprechende Schritt korrekt ist und -, wenn der Schritt falsch ist. Bei der Beantwortung der folgenden Fragen gilt: Schritte, die aufgrund eines Fehlers in einem ungewünschten Ergebnis resultieren würden, aber ansonsten korrekt sind, sind als richtig zu werten.

5.1 Welche Schritte der Vokabularbestimmung sind korrekt? (4 Punkte)

$+$	$-$
(1)	
(2)	
(3)	
(4)	

$$V_1 = \text{Menge der Types} \quad (1)$$

$$V_2 = \text{Menge der Eigennamen} \quad (2)$$

$$d_1 = |V_1| \quad (3)$$

$$d_2 = |V_2| \quad (4)$$

5.2 Welche Schritte der Eingabevektor-Erstellung sind korrekt?

Hinweis: gibt einen d -dimensionalen Vektor zurück, der eine 1 an der dem Index des Tokens im Vokabular entsprechenden Position hat, falls vorhanden, und ansonsten mit Nullen gefüllt ist.

(4 Punkte)

- | | |
|-------------------------------------|-------------------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> |
- (5) (6) (7) (8)

$$\mathbf{x} = (\mathbf{x}_1 + \dots + \mathbf{x}_m) \quad (5)$$

wobei

$$\mathbf{x}_i = \text{One-Hot-Vector}(t_i, V_2) \quad (6)$$

$$\mathbf{x}_i \in \{0, 1\}^{d_1} \quad (7)$$

$$\sum_{j=1}^{d_1} \mathbf{x}_{i,j} \in \{0, 1\} \quad (8)$$

5.3 Welche Schritte der Gewichtsmatrix-Erstellung, der Berechnungen der Hidden-Repräsentation und des Output-Vektors sind korrekt? (4 Punkte)

- | | |
|--------------------------|-------------------------------------|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
| <input type="checkbox"/> | <input type="checkbox"/> |
- (9) (10) (11) (12)

$$\mathbf{U} \in \mathbb{R}^{d_1 \times p} \quad (9)$$

$$\mathbf{V} \in \mathbb{R}^{p \times d_1} \quad (10)$$

$$\mathbf{h} = \frac{1}{d_1} \mathbf{x}^T \mathbf{U} \quad (11)$$

$$\mathbf{y} = f(\mathbf{h}^T \mathbf{V}) \quad (12)$$

dabei gilt

$$f = \text{SoftMax} : \mathbb{R}^l \rightarrow (0, 1)^l$$

<< Zurück

Seite markieren

Weiter >>

6. Attention I (5 Punkte)

6.1 Ist die euklidische Distanz eine weitere Möglichkeit, die Funktion zu implementieren? (5 Punkte)

$$\text{Score}(t, s) = \begin{cases} \mathbf{h}_s^{(1)} \cdot \mathbf{h}_t^{(2)} = (\mathbf{h}_t^{(2)})^T \mathbf{h}_s^{(1)} & \text{Dot-Produkt} \\ (\mathbf{h}_t^{(2)})^T \mathbf{W}_a \mathbf{h}_s^{(1)} & \text{Generell: Parameter-Matrix } \mathbf{W}_a \\ v_a^T \tanh \left(\mathbf{W}_a \begin{bmatrix} \mathbf{h}_s^{(1)} \\ \mathbf{h}_t^{(2)} \end{bmatrix} \right) & \text{Concat: Parameter-Matrix } \mathbf{W}_a \text{ und Vektor } v_a \end{cases}$$

Ja

Nein

<< Zurück

Seite markieren

Weiter >>

6. Sprachmodelle III (9 Punkte)

6.1 Wenn beim Self-Attention der Attention-Mechanismus anstelle von Wörtern Part-of-Speech-Tags verwendet, wird die Zahl der verschiedenen Einbettungen für die Eingabetoken dadurch größer. Ist diese Aussage wahr? (2 Punkte) (2 Punkte)

Wahr

Falsch

6.2 Ist das Cosinus-Maß eine weitere Möglichkeit, die Funktion zu implementieren? (3 Punkte) (3 Punkte)

Ja

Nein

$$\text{Score}(t, s) = \begin{cases} \mathbf{h}_s^{(1)} \cdot \mathbf{h}_t^{(2)} = (\mathbf{h}_t^{(2)})^T \mathbf{h}_s^{(1)} & \text{Dot-Produkt} \\ (\mathbf{h}_t^{(2)})^T \mathbf{W}_a \mathbf{h}_s^{(1)} & \text{Generell: Parameter-Matrix } \mathbf{W}_a \\ v_a^T \tanh \left(\mathbf{W}_a \begin{bmatrix} \mathbf{h}_s^{(1)} \\ \mathbf{h}_t^{(2)} \end{bmatrix} \right) & \text{Concat: Parameter-Matrix } \mathbf{W}_a \text{ und Vektor } v_a \end{cases}$$

<< Zurück

Seite markieren

Weiter >>

6. Sprachmodelle III (9 Punkte)

6.1 Wenn beim Self-Attention der Attention-Mechanismus anstelle von Wörtern Part-of-Speech-Tags verwendet, wird die Zahl der verschiedenen Einbettungen für die Eingabetoken dadurch größer. Ist diese Aussage wahr? (2 Punkte) (2 Punkte)

 Wahr Falsch

6.2 Ist das Cosinus-Maß eine weitere Möglichkeit, die Funktion zu implementieren? (3 Punkte) (3 Punkte)

 Ja Nein

$$\text{Score}(t, s) = \begin{cases} \mathbf{h}_s^{(1)} \cdot \mathbf{h}_t^{(2)} = (\mathbf{h}_t^{(2)})^T \mathbf{h}_s^{(1)} & \text{Dot-Produkt} \\ (\mathbf{h}_t^{(2)})^T \mathbf{W}_a \mathbf{h}_s^{(1)} & \text{Generell: Parameter-Matrix } \mathbf{W}_a \\ v_a^T \tanh \left(\mathbf{W}_a \begin{bmatrix} \mathbf{h}_s^{(1)} \\ \mathbf{h}_t^{(2)} \end{bmatrix} \right) & \text{Concat: Parameter-Matrix } \mathbf{W}_a \text{ und Vektor } v_a \end{cases}$$

<< Zurück

Seite markieren

Weiter >>

7. Attention II (6 Punkte)

7.1 Angenommen, Sie wollen die Ausgabe von zwei Attention-Mechanismen kombinieren: einer auf der Ebene von Sub-Words, der andere auf der Ebene von Lexemgruppen.

Kann man den Output dieser Attention-Layer in jedem Fall mean-poolen?

Nein, da die attentionlayer dimensionen von den verschiedenen Mechnismen, die trainierten gewichte und auch die matrizenkompatibilität unterschiedlich sind.
Demnach ist mean pooling in diesem fall sinnfrei

Antworten Sie zunächst mit **Ja** oder **Nein** und begründen Sie dann Ihre Antwort **kurz**. (6 Punkte)

<< Zurück

Seite markieren

Weiter >>

7. Sprachmodelle IV (4 Punkte)

7.1 Transformer dienen dazu *kontextualisierte Repräsentationen* zu erstellen. Dabei ist der Kontext eines Wortes durch die anderen Wörter und ihre Position in der Eingabesequenz bestimmt.



Durch Pooling von Wörtern, die häufig in Nachbarschaft zum jeweiligen Zielwort stehen.

Durch Pooling von Input- und Output-Embeddings.

Anstelle des ursprünglichen Output-Layers wird ein Output-Layer für jede Kontextposition verwendet.

Wie kann das **Skip-Gramm basierte word2vec-Modell** angepasst werden, um Repräsentationen zu gewinnen, die beim Lernen Kontexte ähnlich zum *Transformer*-Modell nach der vorangegangenen Definition annäherungsweise abzubilden in der Lage sind?

Hinweis: annäherungsweise soll hier bedeuten "nicht gar nicht".
(4 Punkte)

<< Zurück

Seite markieren

Weiter >>

7. Sprachmodelle IV (4 Punkte)

7.1 Transformer dienen dazu *kontextualisierte Repräsentationen* zu erstellen. Dabei ist der Kontext eines Wortes durch die anderen Wörter und ihre Position in der Eingabesequenz bestimmt.

Wie kann das **Skip-Gramm basierte word2vec-Modell** angepasst werden, um Repräsentationen zu gewinnen, die beim Lernen Kontexte ähnlich zum *Transformer*-Modell nach der vorangegangenen Definition annäherungsweise abzubilden in der Lage sind?

Hinweis: annäherungsweise soll hier bedeuten "nicht gar nicht".
(4 Punkte)

- Durch Pooling von Wörtern, die häufig in Nachbarschaft zum jeweiligen Zielwort stehen.
- Durch Pooling von Input- und Output-Embeddings.
- Anstelle des ursprünglichen Output-Layers wird ein Output-Layer für jede Kontextposition verwendet.

<< Zurück

Seite markieren

Weiter >>

8. Sentiment (12 Punkte)

Sei ein Sentiment-Tupel definiert als (e, a, s) , wobei gilt:

- Zielentität e
- Merkmal a von e
- Sentiment s der Meinung des Meinungsträgers bezogen auf das Merkmal a von Entität e
- s kann die folgenden Werte annehmen: negativ, neutral, positiv

In der Praxis stellen wir fest, dass Sie mit dem Universe X42 grundsätzlich gleich gute Fotos schießen wie mit dem Universe Z23 Ultra.

Bei StandardOTOS stimmen Schärfe und Helligkeit, auch wenn Bilder stärker nachgeschärft werden als beim Top-Modell.

Die Farben weichen teils von der Realität ab, das Z23 Ultra zeigt natürlichere Farben an.

Die Akkulaufzeit ist gut - nur das Aufladen geht zu langsam.

Unterm Strich bekommen Sie aber sehr viel geboten für Ihr Geld.

8.1 Wählen Sie die Sentiment-Tupel aus, welche im vorangegangenen Text vertreten werden.

(Universe Z23 Ultra, Aufladen, positiv)

(Universe X42, Akkulaufzeit, positiv)

(Universe X42, Bildhelligkeit, negativ)

(Universe X42, Preis-Leistungsverhältnis, positiv)

Sei ein *relationales* Sentiment-Tupel definiert als: (e_x, e_y, a, s_r)

Dabei gelten die vorangegangen Definitionen für Sentiment-Tupel und:

- Relationales Sentiment s_r der Meinung des Meinungsträgers bezogen auf das Attribut a von e_x im Vergleich zum Attribut a von e_y
- s_r kann die folgenden Werte annehmen: besser, gleich, schlechter

8. Language Model Architectures (6 Punkte)

8.1 Kann BERT (Devlin et al., 2018) auch für das generative Language Modelling verwendet werden?

Antworten Sie zunächst mit **Ja** oder **Nein** und begründen Sie dann Ihre Antwort **kurz**.
(6 Punkte)

Ja, durch das Masked Language Modeling kann Bert lernen auch kontextabhängige vorhersagen zu machen und Lücken im Text zu füllen.

<< Zurück

Seite markieren

Weiter >>

9. Masked Language Modeling I (9 Punkte)

9.1 Es sei ein Masked Language Model (MLM) wie BERT (Devlin et al., 2018) gegeben.

Lassen sich mit einem MLM Scores für Wörter berechnen, die nicht mit einem einzelnen Token repräsentiert werden können?

Antworten Sie zunächst mit **Ja** oder **Nein** und begründen Sie dann Ihre Antwort **kurz**. (9 Punkte)

Ja, BERT kann Wörter oder Teile von Tokens generieren, die in den vorhandenen Kontext passen. Dadurch ist es Bert möglich, Scores zu berechnen, die nicht mit einem einzelnen Token repräsentiert werden können.

<< Zurück

Seite markieren

Weiter >>

9. Stance Detection (12 Punkte)

Bei der Stance-Detection geht es darum, die Haltung eines Autors/einer Autorin zu dem jeweiligen Aussagegehalt eines Texts zu ermitteln und ist daher nicht mit Sentiment-Analyse zu verwechseln. Ein Autor/eine Autorin kann einen bestimmten Aussagegehalt etwa für wahr oder falsch halten, und zwar auch dann, wenn seine/Ihre emotionale Haltung dabei nicht zum Ausdruck kommt.

Sind die folgenden Merkmale oder Pre-Training Aufgaben für die Stance-Detection geeignet? Beantworten Sie die Frage zunächst mit "Ja" oder "Nein" und begründen Sie dann Ihre Antwort **kurz**.

Hinweis: Je max. 4 Punkte pro Aufgabe.

9.1 Negation Scope Detection (4 Punkte)

(maximal 300 Zeichen)



Zeichen übrig für die Beantwortung dieser Frage 300

9.2 Semantic Roles (4 Punkte)

(maximal 300 Zeichen)

A large, empty rectangular text input field with a thin gray border.

Zeichen übrig für die Beantwortung dieser Frage 300

9.3 Topic Modeling (4 Punkte)

(maximal 300 Zeichen)

A large, empty rectangular text input field with a thin gray border.

Zeichen übrig für die Beantwortung dieser Frage 300

<< Zurück

Seite markieren

Weiter >>

10. Masked Language Modeling II (9 Punkte)

10.1 Es sei ein Masked Language Model (MLM) wie BERT (Devlin et al., 2018) gegeben.

Wie können Sie feststellen, ob das gegebene Modell Biases enthält, z.B. einen Gender Bias, der bestimmte Begriffe mit geschlechter-spezifischen Wörtern assoziiert und damit existierende Biases repliziert?

Skizzieren Sie zunächst Ihre Idee oder Ihren Ansatz und erläutern Sie dann **kurz** wie Sie das MLM einsetzen um die gegebene Aufgabe zu lösen. (9 Punkte)

<< Zurück

Seite markieren

Weiter >>

11. Masked Language Modelling III (9 Punkte)

11.1 Es sei ein Masked Language Model (MLM) wie BERT (Devlin et al., 2018) gegeben.

Wie können Sie ein MLM zur Text-Klassifikation (z.B. Sentiment Analysis mit drei Klassen) verwenden, ohne Änderungen an der Modellarchitektur vorzunehmen oder weitere Teile hinzuzufügen, und welche Voraussetzungen muss das Modell hierfür erfüllen?

Skizzieren sie zunächst Ihre Idee oder Ihren Ansatz und erläutern Sie dann **kurz** wie Sie das MLM einsetzen um die gegebene Aufgabe zu lösen. (9 Punkte)

Das NSP Token eines ML Models wie Bert kann mit der auswahl von Trainingsdaten und umstellung des NSP tokens so geändert werden, dass es nicht treu oder false für eine Next sentence instanz gibt sondern, zum beispiel im fall von sentiment analyse einen positiv, negativ oder neutral label. So kann ohne der Architektur veränderung in MLM model darauf trainiert werden Texte zu klassifizieren

<< Zurück

Seite markieren

Weiter >>

12. Sentiment (6 Punkte)

12.1 Sei ein Sentiment-Tupel definiert als (e, a, s) , wobei gilt:

- Zielentität e
- Merkmal a von e
- Sentiment s der Meinung des Meinungsträgers bezogen auf das Merkmal a von Entität e
- s kann die folgenden Werte annehmen: negativ, neutral, positiv

In der Praxis stellen wir fest, dass Sie mit dem Universe X42 grundsätzlich gleich gute Fotos schießen wie mit dem Universe Z23 Ultra. Bei Standardfotos stimmen Schärfe und Helligkeit, auch wenn Bilder stärker nachgeschärft werden als beim Top-Modell. Die Farben weichen teils von der Realität ab, das Z23 Ultra zeigt natürlichere Farben an. Die Akkulaufzeit ist gut - nur das Aufladen geht zu langsam. Unterm Strich bekommen Sie aber sehr viel geboten für Ihr Geld.

Wählen Sie die Sentiment-Tupel aus, welche im vorangegangenen Text vertreten werden.

Hinweis: Multiple-Choice-Frage. (6 Punkte)

- (Universe Z23 Ultra, Akkulaufzeit, neutral)
- (Universe Z23 Ultra, Bildhelligkeit, positiv)
- (Universe X42, Akkulaufzeit, positiv)
- (Universe X42, Bildhelligkeit, negativ)
- (Universe Z23 Ultra, Aufladen, positiv)
- (Universe X42, Preis-Leistungsverhältnis, positiv)

13. Sentiment II (6 Punkte)

13.1 Sei ein *relationales* Sentiment-Tupel definiert als:
 (e_x, e_y, a, s_r)

Dabei gelten die vorangegangen Definitionen für Sentiment-Tupel und:

- Relationales Sentiment s_r der Meinung des Meinungsträgers bezogen auf das Attribut a von e_x im Vergleich zum Attribut a von e_y
- s_r kann die folgenden Werte annehmen: besser, gleich, schlechter.

In der Praxis stellen wir fest, dass Sie mit dem Universe X42 grundsätzlich gleich gute Fotos schließen wie mit dem Universe Z23 Ultra.
Bei StandardOTOS stimmen Schärfe und Helligkeit, auch wenn Bilder stärker nachgeschärft werden als beim Top-Modell.
Die Farben weichen teils von der Realität ab, das Z23 Ultra zeigt natürlichere Farben an.
Die Akkulaufzeit ist gut - nur das Aufladen geht zu langsam.
Unterm Strich bekommen Sie aber sehr viel geboten für Ihr Geld.

- (Universe Z23 Ultra, Universe X42, Akkulaufzeit, besser)
- (Universe X42, Universe Z23 Ultra, Farbnatürlichkeit, schlechter)
- (Universe X42, Universe Z23 Ultra, Bildhelligkeit, besser)
- (Universe X42, Universe Z23 Ultra, Fotoqualität, gleich)

Wählen Sie die relationalen Sentiment-Tupel aus, die im vorangegangenen Text vertreten werden.

Hinweis: Multiple-Choice-Frage. (6 Punkte)

14. Entailment (12 Punkte)

Bei der Entailment-Aufgabe geht es darum, eine direktionale Beziehung zwischen zwei Texten zu bestimmen. Die beiden Texte werden als Premisse und Hypothese bezeichnet.

Positives Entailment besteht, wenn die Hypothese aus der Premisse folgt.

Negatives Entailment besteht, wenn die Hypothese genau nicht aus der Premisse folgt.

Kein/Neutrales Entailment besteht, wenn es keine Beziehung von Premisse zu Hypothese gibt, die Texte etwa vollständig unterschiedliche Themen behandeln.

Sind die folgenden Merkmale oder Pre-Training Aufgaben für das Entailment geeignet? Beantworten Sie die Frage zunächst mit Ja oder Nein und begründen Sie dann Ihre Antwort kurz.

14.1 Topic Modeling (4 Punkte)
(maximal 400 Zeichen)

Ja, Topic modeling im Pre-training, verhilft dem model ein besseres verständnis zwischen den behandelten themen zu schaffen. Dadurch können die Inhalte aus Pämisse und Hypothese deutlicher verglichen und bewertet werden, was das Entailment dieser beiden elemente verhilft

Zeichen übrig für die Beantwortung dieser Frage 106

14.2 Named Entity Recognition (4 Punkte)
(maximal 400 Zeichen)

Nein, zwar verhilft Named entity recognition token aus einem input zu kategorisieren und dadurch ein tieferes verständnis des unstruktiereten textes zu schaffen, jedoch würde dies nicht dazu verhelfen die Hypothese und Prämisse miteinander in kontext und damit ein entailment zu errechnene

Zeichen übrig für die Beantwortung dieser Frage 101

14.3 Next Sentence Prediction (4 Punkte)
(maximal 400 Zeichen)

Ja, da mit NSP ein MLM darauf trainiert werden kann die frage mit dem [NSP] so klassifiziert werden. Das NSP wird damit trainiert aus dem input [CLS]Prämisse[SEP]Hypothese entweder positiv, negativ oder neutral zu bewerten.

Zeichen übrig für die Beantwortung dieser Frage 31

<< Zurück

Seite markieren

Weiter >>