MT Übung 1

Thema: Evaluation von MÜ-Systemen

Abgabetermin: Montag, 12. März 2018, 18 Uhr

Hinweise zur Abgabe:

- **Abgabeformat: zip**
- Bitte benennt eure Datei nach folgendem Schema: olatbenutzername_mt_uebungxx.zip, z.B. `mmuster_mt_uebung01.zip`
- Für diese Übung werdet ihr ein Python-Skript sowie eine PDF-Datei abgeben. Gebt beides in derselben ZIP-Datei ab.
- Die Abgabe erfolgt über den Übungsbaustein auf OLAT. Bitte gebt pünktlich ab, denn der Baustein ist nur bis Montag, 18 Uhr offen.

1 Automatische Evaluation

In der Vorlesung habt ihr gelernt, dass BLEU (Bilingual Evaluation Understudy) eine der am häufigsten benutzten automatischen Evaluationsmetriken ist. In dieser Übung werdet ihr ein Skript schreiben, welches den BLEU-Wert eines übersetzten Textes berechnen kann. Es genügt, wenn ihr davon ausgeht, dass ihr nur eine Referenzübersetzung habt. Folgt dafür den folgenden Schritten **(je 1.0 Punkte)**:

1. Schreibt eine Funktion, die eine Datei öffnet und ihren Inhalt in einen String liest. Die Funktion soll anschliessend den Text tokenisieren und dann zurückgeben. Für die Tokenisation könnt ihr den NLTK Tokenizer benutzen. Ein Beispiel, wie das funktioniert findet ihr weiter unten.

2. Schreibt eine Funktion, die die n-Gramm-Präzision (p) von einem übersetzen Text (Hypothese) und einer Referenz für eine gegebene n-Gramm Grösse berechnen kann. Orientiert euch dafür an der Formel aus der Vorlesung und vergesst nicht das Clipping:

$$p = \frac{correct}{hyp\ length} \tag{1}$$

3. Schreibt eine Funktion, die für alle n-Gramme bis zu einer bestimmten Grösse mit der oben beschriebenen Funktion deren n-Gramm-Präzision berechnet. Berechnet das geometrische Mittel (P) aus allen n-Gramm-Präzisionen mit der Formel aus der Vorlesung (siehe unten). Die Gewichte λ könnt ihr für diese Übung auf 1 und die maximale n-Gramm-Grösse N auf 4 setzen.

$$P = (\prod_{n=1}^{N} \lambda_n\ p_n)^{\frac{1}{N}} \tag{2}$$

   Die Funktion soll weiterhin auch die Brevity Penalty (BP) berechnen nach der Formel aus der Vorlesung (siehe unten). Wie ihr die Exponentialfunktion in Python aufrufen könnt, findet ihr im Code-Beispiel weiter unten.

$$BP = \min(1.0, \exp(1 - \frac{ref\ length}{hyp\ length})) \tag{3}$$

   Zum Schluss soll die Funktion aus den beiden berechneten Werten noch den BLEU-Wert berechnen und diesen zurückgeben. Die Formel aus der Vorlesung lautet:

$$BLEU = BP * P \tag{4}$$

4. Schreibt die `main` Funktion, welche mit der ersten Funktion zwei Dateien einliest, die auf der Kommandozeile als Argument mitgegeben wurden. Gebt die tokenisierten Texte dann an die dritte Funktion, welche mit Hilfe der zweiten Funktion den BLEU-Wert berechnet. Euer Programm soll diesen Wert ausgeben.

Abgabe: Ein Python-Skript, in welchem die oben beschriebenen Funktionen implementiert sind. Ihr könnt eure Implementation testen, indem ihr die beiden Dateien `test_reference.txt` und `test_hypothesis.txt` verwendet, welche ihr im Übungsordner finden könnt. In den Dateien findet ihr das Beispiel aus der Vorlesung. Euer Skript sollte auf drei Stellen nach dem Komma gerundet den selben Wert ausgeben wie auf den Folien. Testet auch ob ihr das Clipping richtig implementiert habt. Benutzt dafür die beiden Dateien `test_clipping_reference.txt` und `test_clipping_hypothesis.txt` und lasst euch bei der zweiten Funktion die n-Gramm-Präzision ausgeben. Wie auf den Folien sollte die Präzision für eine n-Gramm-Grösse von 1 $\frac{1}{7}$ sein und für alle anderen n-Gramm-Grössen Null.

Ein Beispiel für den Gebrauch des NLTK Tokenizers:

```
1 from nltk.tokenize import word_tokenize
2
3 sentence = "This is a house."
4 tokenized = word_tokenize(sentence)
5
6 print tokenized # prints ['This', 'is', 'a', 'house', '.']
```

Ein Beispiel für den Gebrauch der Exponentialfunktion:

```
1 import numpy as np
2
3 exp_of_1 = np.exp(1)
4
5 print exp_of_1 # prints 2.71828182846
```

2 Automatische vs. Manuelle Evaluation

In der Vorlesung habt ihr auch gelernt, dass es manchmal sinnvoller ist Übersetzungen manuell zu bewerten. Darum sollt ihr in dieser Übung verschiedene maschinelle Übersetzungssysteme manuell vergleichen. Im Übungsordner findet ihr einen Artikel aus der Zeit Online, welcher in Deutsch[1] und Englisch[2] veröffentlicht wurde. Weiterhin findet ihr Übersetzungen von diesem Artikel mit Google Translate, DeepL und Microsoft Translator (Bing). Löst dazu folgende Aufgaben:

[1] `http://www.zeit.de/wissen/2017-03/march-for-science-donald-trump-forschung-wissenschaft`
[2] `http://www.zeit.de/wissen/2017-03/march-for-science-donald-trump-research-usa`

1. Benutzt euer Python-Skript aus Aufgabe 1, um die BLEU-Werte der verschiedenen Übersetzungen zu berechnen. Erstellt dann eine Tabelle nach folgendem Muster und füllt eure Werte ein **(0.25 Punkte)**:

	DE→EN	EN→DE
Google Tanslate		
DeepL		
Microsoft Translator		

2. Nun wählt zwei Systeme und eine Übersetzungsrichtung aus, die ihr genauer untersuchen wollt. Am besten vergleicht ihr die Übersetzungen mit dem interaktiven BLEU Vergleich-Werkzeug von Tilde. Dafür müsst ihr den Quelltext, die Referenzübersetzung sowie die beiden Übersetzungen der MÜ-Systeme hochladen und dann auf "Score" klicken.

3. Schaut euch die übersetzten Sätze an und vergleicht die beiden Systeme mit der Referenz. Sucht, wenn möglich, je zwei Beispiel-Sätze für die folgenden Szenarien **(je 0.25 Punkte)**:

   - Der BLEU-Wert von einem System ist höher, aber der Satz ist nach eurer Meinung schlechter übersetzt als beim anderen.
   - Die Referenzübersetzung ist nur eine sinngemässe Übersetzung, weswegen die BLEU-Werte der Systeme schlecht sind, obwohl die Übersetzungen gut sind.
   - Eine Übersetzung, die euch amüsiert.

4. Entscheidet in den folgenden Fällen, ob ihr eine automatische oder manuelle Evaluation bevorzugen würdet und begründet eure Entscheidung **(je 0.25 Punkte)**:

   - Ihr wollt überprüfen, ob euer neues MÜ-System ein spezielles linguistisches Phänomen besser übersetzt als ein Standard-MÜ-System.
   - Ihr habt ein MÜ-System für eine Firma gebaut und müsst die Projektleiter über die Qualität des resultierenden Systems informieren.
   - Ihr organisiert einen Wettbewerb für Maschinelle Übersetzung und müsst viele Systeme vergleichen, um den Gewinner festzulegen.
   - Ihr startet ein neues Projekt und möchtet zuerst einen Überblick über die Qualität verschiedener Systeme erhalten, um zu entscheiden, welches davon ihr weiter entwickeln wollt.

Abgabe: Ein PDF-Dokument mit euren Antworten zu den Aufgaben oben.

Bei Problemen oder Unklarheiten postet bitte einen Beitrag ins OLAT-Forum.
Viel Erfolg!