Lehrstuhl für Informatik 6
der RWTH Aachen
Prof. Dr.-Ing. H. Ney

Aachen, den 27. April 2021
C. Herold, J. Rosendahl

Software-Projektpraktikum Maschinelle Übersetzung 2. Übung

Thema:

Dieses Aufgabenblatt dient vor allem der Vorbereitung des ersten Übersetzungsmodells. Die Auswahl des Vokabulars stellt eine frühe und extrem wichtige Designentscheidung für das finale System dar. Die BPE Methode zur Subwort Zerlegung erlaubt eine effektive Nutzung von verhältnismäßig kleinen Vokabulars.

Anmerkungen:
Für eine effiziente Nutzung sollten alle Ihre Programme in der Lage sein an sinnvolle Modellteile (wie ein Subwort-Zerlegungsmodell oder ein Vokabular) zu speichern und zu laden. Arbeiten Sie auf den Trainingsdaten `multi30k.de.gz` und `multi30k.en.gz`, die auf der Seminarwebseite zu finden sind.

Aufgabe:

1. Der byte pair enconding (BPE) Algorithmus, welcher im Rahmen der Vorlesung vorgestellt wurde, erlaubt es Wörter in kleinere Subwort-Gruppen zu zerlegen. Diese ermöglichen es lange und seltene Wörter einer Sprache darzustellen ohne diese explizit in das Vokabular des Modells aufzunehmen.
   Schreiben Sie ein Programm, das in der Lage ist

   - anhand einer gegebenen Anzahl von Zusammenzug-Operationen sowie Trainingsdaten eine BPE Zerlegung zu lernen.
   - erlernte BPE Operationen auf Text anzuwenden.
   - die BPE Zerlegung auf einem Text rückgängig zu machen.

   Auswertung:
   Trainieren Sie diese Operationen zunächst auf den Trainingsdaten Ausgangs- und Zielsprache getrennt und anschließend auf der Vereinigung der Trainingsdaten (getrenntes und vereinte BPE). Wie verändert sich die Anzahl der verschiedenen Wörter (und damit die Größe des Vokabulars) durch das Anwenden der Subwort-Zerlegung mit 1k, 5k oder 15k Operationen? Stellen Sie Ihre Ergebnisse übersichtlich dar.

2. Implementieren Sie eine Klasse `Dictionary`. Die Klasse soll eine bidirektionale Zuweisung von Strings und Indices realisieren, um die weitere Verarbeitung zu ermöglichen. Nutzen Sie diese Klasse um ein Programm zu schreiben, das ein Vokabular aus gegebenen Trainingsdaten erzeugt. Das dabei erstellte Vokabular soll auch auf neue Datensätze angewendet werden können. Beachten Sie dabei, dass nicht jeder String der angefragt wird im `Dictionary` vorhanden sein muss.

   Unabhängig von der Methode des BPE Training sollen die Vokabulare in der Übersetzung für Eingabe- und Ausgabesprache getrennt erstellt werden.

3. Für die Weiterverarbeitung mit neuronalen Modellen empfiehlt es sich Daten in sogenannten „Batches“ zu gruppieren. Schreiben Sie eine Funktion die aus einer Reihe von Sätzen Batches, wie Sie in der Vorlesung vorgestellt wurden generiert. Ihr Programm sollte dabei ein monotones Alignment wie es in der Vorlesung vorgestellt wurde benutzen. Jeder Batch soll dabei drei arrays umfassen (in der Vorlesung als S, T und L bezeichnet).

   Konvertieren Sie die Trainingsdaten in Batches der Größe $B = 200$. Geben Sie die Elemente der Batches für die Zeilen 1100 bis 1200 des Korpus in einer Textdateien aus. Die Einträge für S, T und L sollen dabei nebeneinander stehen. Erstelle Sie eine zweite Textdatei mit gleicher Datenabfolge, nun sollen aber die die Elemente der Batches als Strings dargestellt werden.

Erinnerung:

Nächste Woche am Mittwoch den 04. Mai um 8:30h–10:00h findet eine zusätzliche Vorlesung statt. An diesem Termin wird keine neue Aufgabe ausgegeben und die Abgabe erfolgt wie gewohnt im zwei-Wochen-Rhythmus.

Abnahmetermin: Montag, 10. Mai, Uhrzeit nach Absprache

Schriftliche Ausarbeitungen werden nicht verlangt. Schicken Sie bitte die commit-ID Ihrer Abgabe (kommentierten Quelltexte und die kurzen Ausgabewerte der geforderten Ergebnisse) bis Sonntag, (09. Mai, 23:59 Uhr) an

mtprak21assi@i6.informatik.rwth-aachen.de

Am Montag erläutern Sie uns dann Ihre Lösungen und demonstrieren Ihre Programme.