Parsing mit neuronalen Netzen: Training

Nun sollen Sie das Parser-Training implementieren.

Korrigieren Sie zunächst die Fehler in Ihrem Code aus der letzten Aufgabe mit Hilfe der erhaltenen Korrekturvorschläge.

Erstellen Sie dann eine Datei **train-parser.py**, welche eine Klasse `Data` für die Datenvorverarbeitung und Ihre Klasse `Parser` für das Netzwerk importiert. Die Klasse `Data` enthält Funktionen zur Vorverarbeitung der Daten. Sie können den Code an der Adresse `https://www.cis.uni-muenchen.de/~schmid/lehre/Experimente/data/Code-Aufgabe10.zip` herunterladen.

Schreiben Sie nun eine Funktion **train**. Erzeugen Sie darin zunächst ein Objekt der Klasse `Data` mit dem Befehl

```
data = Data(path_train, path_dev)
```

Dann trainieren Sie für n Epochen (bspw. n=50) auf den Trainingsdaten `data.train_parses`, die Sie zu Beginn jeder Epoche mit `random.shuffle` umordnen. `data.train_parses` ist eine Liste von Parsebäumen. Jeder Parsebaum ist ein Paar bestehend aus der Liste der Wörter und der Liste der Konstituenten. Jede Konstituente ist ein Tripel (Label, Startposition, Endposition). Mit der Methode `data.labelID(label)` können Sie die Labels auf IDs abbilden. Mit der Methode `data.words2charIDvec(words)` können Sie eine Liste von Wörtern in eine Liste von Suffixen und eine Liste von Präfixen konvertieren, wobei jedes Suffix/Präfix aus einer Liste von Buchstaben-IDs besteht. In der Datei `Data.py` ist auch der Index der Spanklasse "keine Konstituente" in der Variablen `NoConstLabelID` verfügbar. Die Methoden `data.num_char_types()` und `data.num_label_types()` liefern Ihnen die Zahl der Zeichen und die Zahl der syntaktischen Labels.

Das Netzwerk gibt Ihnen einen 2-dimensionalen Tensor zurück, der für jeden Span eine Liste von Scores liefert. Sie müssen daher zur Berechnung des Cross Entropy Loss einen Label-Vektor erstellen, der für jeden Span die korrekte Label-ID enthält. Sie gehen dabei am besten so vor, dass Sie zunächst einen 1-dimensionalen Tensor der richtigen Länge mit `NoConstLabelIDs` füllen. Dann erzeugen Sie mit dem Split-Befehl von PyTorch eine Liste `tview` von Ansichten auf diesen Vektor, wobei jede Ansicht alle Spans mit einer bestimmten Startposition umfasst. Nun können Sie mit `tview[i][k-i-1] = m` das Label des Spans (i,k) auf die Label-ID m setzen. (Falls Sie mit Ansichten noch nicht vertraut sind, lesen Sie bitte die PyTorch-Dokumentation dazu durch.)

Als Trainingskriterium verwenden Sie das `CrossEntropyLoss` von PyTorch. Wenden Sie **Gradient Clipping** (bspw. mit einer Gradient-Norm von 1) an, um extrem große Gradienten zu vermeiden. Nach jeder Epoche berechnen Sie die Zahl der falsch gelabelten Spans in den Development-Daten `data.dev_parses`. Wenn die Zahl der Fehler die bisher kleinste war, speichern Sie das Netzwerk mit der Methode `torch.save`. Außerdem müssen Sie die Methode `data.store_parameters` aufrufen, um die Tabellen für die Abbildung von Buchstaben und Labels auf Zahlen-IDs zu speichern. Diesen Befehl müssen Sie nur einmal aufrufen. Sie können die Tabellen

und das Netzwerk in separaten Dateien mit gleichem Basisnamen und unterschiedlichen Dateiendungen speichern.

Bei einer guten Implementierung kann die Zahl der falsch gelabelten Konstituenten in den Development-Daten im Laufe des Trainings unter 6000 sinken.

Aufruf: python train-parser.py train-parses dev-parses parfile

Bitte schicken Sie mir den kompletten Code, der zur Ausführung notwendig ist, und eine Datei *num-errors.txt* mit der Anzahl der Fehler nach jeder Epoche.