Contents

1 Einleitung

Im Praktikum *Text2Sene* geht es darum, aus Textbeschreibungen Szenen zu erstellen. Dabei wurde die Arbeit unterteilt; wir beschäftigen uns mit der Thematik der *preposition sense disambiguation* (Sinneszuordnung und -erkennung von Präpositionen). Für diese Aufgabe haben wir verschiedene *State-of-the-Art-Verfahren* begutachtet. Schlussendlich haben wir uns dafür entschieden, einerseits einen *Semi-supervised* Ansatz umzusetzen und daneben KIs mit Hilfe der frameworks FlairNLP, AllenNLP und Huggingface zu programmieren.

2 FlairNLP

FlairNLP ist ein Framework, das speziell für NLP-Aufgaben konzipiert ist. Für unsere Aufgabe nutzen wir einen *text classifier*, der wie der Name sagt, eine Eingabe klassifiziert und dadurch die Präposition dem zugehörigen Sinn zuordnet.

2.1 Vorgehen

Damit einer Präposition ein Sinn zugeordnet werden kann, passieren einige Dinge. Der Hauptteil des Projekts besteht aus einer NLP-KI mit dem Flair-framework.
Damit eine Eingabe verarbeitet werden kann, muss diese zunächst vorbereitet werden. Dazu werden die Daten aus den xml-Dateien gelsen und in einer csv-Datei mit dem entsprechenden Format für Flair abgespeichert. Hierbei wird das standard-Format (__label__<label>) verwendet. Die Trainingsdaten werden in drei Dateien aufgeteilt, dabei werden 80% Training, 10% Dev und weitere 10% Test zugeschrieben.
Nachdem die Daten im csv-Format vorliegen, wird daraus ein Korpus erstellt. Dazu wird ein *Dictionary* und die entsprechenden Embeddings erstellt. Anschließend wird der *Classifier* erstellt und das Training beginnt.

2.2 Programm

Das Flair-Projekt umfasst vier Dateien:

- Flair_prepare.py - Diese Datei dient der Vorbereitung der Daten.
- Dataset_class.py - Diese Datei enthält wichtige Klassen auf Basis des FlairNLP-Frameworks zum ertsellen des Korpus.
- flair_test_classification_model.py - Diese Datei dient zum erstellen des *text classifiers* und des Trainings dessen
- predict.py - Diese Datei enthält den *predictor*, bzw. dient zum testen des Endproduktes.

2.2.1 Flair_prepare.py

Der erste Schritt befasst sich mit der Vorbereitung der Daten. Wir benutzen für unser Training die Daten des **SemEval 2007**, die Rund 16.000 Sätze beinhalten zu 34 verschiedenen Prepositionen. Diese Daten sind in xml-Format gegeben, weshalb sie in csv-Format geändert werden müssen.
Als erstes werden die xml-Dateien mit Hilfe der Python-library *xml.etree.ElementTree* ausgelesen. Dabei speichern wir die *senseid* und den Satz in einer Liste. Die *senseid* wird zeitgleich mit dem von Flair benötigten Zusatz *__label__* versehen. Fehlerhafte Einträge - solche, bei denen entweder der Sinn oder der Satz fehlt bzw. fehlerhaft ist - werden zudem aussortiert und deren *instanceid* zur Überprüfung ausgegeben.
Nachdem alle Einträge in die Liste eingefügt wurden, wird diese gemischt und anschließend in *Training*, *Dev* sowie *Test* Dateien **disjunkt** aufgeteilt (80% - 10% - 10%).

2.2.2 Dataset_class.py

Die Datei *Dataset_class.py* enthält die Klassen zum Erstellen des Korpus. Diese sind unverändert aus dem **GitHub Repository** von Flair übernommen. Für eine genaue Dokumentation dieser beiden Klassen verweisen wir auf die Dokumentation von FlairNLP. Sie sind auf csv-Dateien angepasst.

2.2.3 Flair_text_classification_model.py

Diese datei beinhaltet alle Einstellungen und Bauteile für den *Text classifier*.

```
col_name_map = {0: "label", 1: "text"}

# 1. get the corpus
corpus: Corpus = CSVClassificationCorpus('data/
    ', col_name_map)
print(Corpus)

# 2. create the label dictionary
label_dict = corpus.make_label_dictionary()
```

Als erstes wird der Korpus auf Basis der zuvor erstellten csv-Dateien und der angegebnen *col_name_map* erzeugt, sowie ein Dictionary des Korpus errichtet. Flair bietet neben einem csv-Korpus auch andere Formate, wir haben uns aber für csv entschieden, da es ein gängiges Format ist und auch für andere Teilprojekte verwendet werden kann.

```
# instantiate one-hot encoded word embeddings
    with your corpus
embeddings_1 = OneHotEmbeddings(corpus)

# instantiate word embeddings that work well
    for you
embeddings_2 = WordEmbeddings('glove')

# document pool embeddings
```

```
document_embeddings = DocumentPoolEmbeddings([
    embeddings_1, embeddings_2], fine_tune_mode='
    none')
```

Nachdem der Korpus erzeugt wurde, werden die Embeddings erzeugt. Hierbei können mehrere Embeddings durch sog. *pool-embeddings* zusammen genutzt werden. Wir haben uns für OneHotEmbeddings und WordEmbeddings (Typ *glove)* entschieden.

```
# 5. create the text classifier
#classifier = TextClassifier(
    document_embeddings, label_dictionary=
    label_dict)
classifier = TextClassifier.load('resources/
    best-model.pt')
```

Nachdem die Embeddings erstellt wurden, kann der Classifier erzeugt werden. Alternativ kann natürlicha uch ein bestehender Classifier geladen werden. Wichtig ist nur, dass dieser Classifier auch die Daten lesen kann, also dem selben Grundaufbau folgt.

```
flair.device = torch.device('cuda:0')

# 6. initialize the text classifier trainer
trainer = ModelTrainer(classifier, corpus)

# 7. start the training
trainer.train('resources/',
              learning_rate=0.1,
              mini_batch_size=32,
              anneal_factor=0.5,
              patience=5,
              max_epochs=10)
```

Bevor das Training startet, wählen wir noch, dass auf einer GPU trainiert werden soll. Da Flair über Torch läuft, kann dies einfach über troch gewählt werden.
Anschließend erstellen wir den Trainer auf Basis des zuvor erzeugten oder geladenen Classifier und Korpus und beginnen das Training. Hierbei können verschiedene Einstellungen vorgenommen werden. Zunächst wird das *output-directory* angegeben., welches wir mit *resources* angegeben haben. In diesem Verzeichnis werden dann logs und die Models abgespeichert. Die *learning rate* haben wir auf dem Sdandartwert belassen, ebenso die beiden weiteren Parameter. Die *patience* kann variabel angepasst werden. Sie verursacht, dass das Training bei zu vielen Epochen ohne Verbesserung **hintereinander** das Training abgebrochen wird. Je höher die patience, desto mehr Epochen ohne Verbesserung können vorkommen, ohne dass das Training abgebrochen wird. Der letzte Parameter ist die maximale Anzahl an Epochen. Wir haben einige Tausend Epochen trainiert. Dieses Trainign kann aber auch dauern - je Epoche bis zu 40 Sekunden oder mehr, je nach Hardware.

2.2.4 predict.py

Der Predictor dient dazu, die Fertige KI anzuwenden. Das Programm kann dafür auch aus der Konsole mit Übergabeparameter verwendet werden. Der Übergabeparameter muss dabei ein String sein.
Alternativ kann das Programm auch ohne Übergabeparameter ausgeführt werden. In diesem Fall wird eine Reihe an Testdaten predictet.

3 AllenNLP

3.1 Vorgehen

3.2 Programm

4 Blabla