CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 22.05.2017

Thema: Comparisson of Transfer Methods for Low-Ressource Morphology.

Student: Alexander Vordermaier
Betreuerin: Katharina Kann

Der Student Alexander Vordermaier schreibt ihre Bachelorarbeit im Rahmen des Projekts 2016-Task-Encoder-Decoder. Die Dateien, die hier verwendet werden, enthalten das Lemma und das flektierte Form des Wortes (*target*).

Die Motivation der Arbeit geht es um Paradigmen Komplettierung von Sprachen, sowie um die Zuordnung eines Lemmas zu seinem flektierten Formen. Das Grundprinzip des Projekts ist es, dass wenn es nicht genügend Trainingsdaten zur Verfügung stehen, kann man auch Daten von anderen Sprachen oder der gleichen Sprache mit nicht annotierten Daten benutzen. Dafür werden hier zwei unterschiedlichen Ressourcen verwendet: High Ressource (HR) in Bulgarisch und Low Ressource (LR) in Mazedonisch.

Die Herangehensweise lässt sich in drei Methoden für die Paradigmen Komplettierung gliedern:

1. Sprachübergreifende Paradigmen Komplettierung.
2. Auto Encoding
3. Kombination aus den beiden ersten Methoden.

Bei der ersten Methode sucht man zu einer LR- Sprache eine ähnliche HR-Sprache. Die Ähnlichkeit der beiden Sprachen ist besonders wichtig. Danach vermischt man die Daten aus beiden Sprachen und trainiert sie zusammen. Vordermaier erklärt, dass die große Hoffnung darin besteht, davon brauchbare Ergebnisse zu erhalten.

Größe der hier verwendeten Paketen:
LR = 50, 200
HR = 50, 200, 400, 800, 1600, 3200, 6400, 12800

Die Pakete werden hier immer Paarweise vermischt, zum Beispiel:
(LR/HR) (50/400) (200/6400)

Bei der zweiten Methode vermischt man annotierte Daten der LR- Sprache mit nicht annotierten Daten der selben LR-Sprache (hier Mazedonisch) und jedes Paketpaar wird wieder trainiert.

Größe der hier verwendeten Paketen:
LR(annotiert) = 50, 200
LR = 50, 200, 400, 800, 1600, 3200, 6400, 12800

Wie der Referent klargestellt hat, ist der Auto Encoding ein sehr simples Verfahren, bei dem die Eingabe auch gleichzeitig die Ausgabe ist. Einerseits hofft man also, dass viele Flektionen der Wörter gleich sind. Anderseits besteht ein großes Gefahr, dass dies nicht der Fall ist.

Im Falle von der dritten Methode vermischt man annotierte Daten der LR- Sprache mit annotierten Daten der HR-Sprache und mit nicht annotierten Daten der LR-Sprache.

Größe der hier verwendeten Paketen:
LR(annotiert) = 50, 200
HR = 50, 200, 400, 800, 1600, 3200, 6400, 12800
LR = 50, 200, 400, 800, 1600, 3200, 6400, 12800

Damit die Daten von dem Modell verarbeitet werden können, müssen sie eine gewisse Form besitzen. Ein trainigsbereiter Abschnitt besteht aus den Dateien der folgenden Tabelle:

Source	**Target**
Test_src	Test_trg
Dev_src	Dev_trg
Train_scr	Train_trg
Vokabulare zu Source und Target	

Wie Vordermaier präzisiert hat, in Source sind die Daten enthalten, die das Modell bekommt. In Target hingegen ist die Lösung enthalten.

Evaluierung

Genauigkeit wird einmal für Paketgröße 50 (annotierte Mazedonisch) gemessen. Man stellt fest, dass hier die Ergebnisse nicht gut sind. Die zweite Methode schneidet ziemlich schlecht. Methode 1 läuft etwas besseres und kommt fast auf die 50 % und die Kombination verbessert das ganze nur ein bisschen. Aber das Ziel sind nicht unbedingt gute Ergebnisse, sondern festzustellen ob diese Methode tatsächlich funktionieren.

Die Genauigkeit beim Parketgröße 200 sieht etwas besser. Aus. Methode 2 erreicht fast 60% Genauigkeit. Die anderen beiden Methode sind recht ok und sogar bei der Kombination aus beiden wird die 80% der Genauigkeit erreicht. Mit diesem Ergebnis, meint Vordermaier, kann man schon anfangen, damit zu arbeiten.

Fehleranalyse

Der Student hat festgestellt, dass oft die falsche Endung verwendet wird. Außerdem treten viele Fehler bei dem Auto Encoding auf Grund der Vorgehensweise auf. Für Vordermaier ist es alle noch doppelt schwer, denn er kein Mazedonisch- oder Bulgarisch- Muttersprachler ist.

Zuletzt wurden weitere Aufgaben aufgelistet, bei denen noch gearbeitet wird:

- Weitere Fehlerquellen identifizieren
- Weitere Modelle finden
- Bereits gängige Verfahren beleuchten