

## **Protokoll zur Sitzung vom 15.05.2017**

### **Kurzvorstellung von der Bachelorarbeit von Kristina Smirnov**

**BA-Betreuer: M.Sc. K. Kann**

#### *Thema: Comparison of Transfer Methods for Low-Resource Morphology*

Kristina hat in der heutigen Sitzung ihre Bachelorarbeit vorgestellt, in der es um Vergleich verschiedenen Methoden für Low-Resource-Morphologie geht. Kristina schreibt ihre Arbeit im Rahmen von Sigmorphon 2016 Shared Task, das aus 3 Tasks besteht:

- Task 1: given a lemma with its POS, generate a target inflected form
- Task 2: given an inflected form and its current tag, generate a target inflected form
- Task 3: given an inflected form without its current inflection, generate a target inflected form.

Die Hauptidee Kristinas Bachelorarbeit ist es, ein System zu trainieren, das eine richtige flektierte Form eines Wortes ausgibt. Das Wort steht als ein angegebenes Lemma im Dataset und die benötigte Form des Wortes wird als Input eingegeben. Das Experiment wurde für Russisch durchgeführt. In ihrer Bachelorarbeit verwendet Kristina zwei Methoden, und zwar eine andere Sprache zu Hilfe zu nehmen (beide sind annotiert) und die gleiche Sprache zu Hilfe zu nehmen, aber diesmal einmal annotiert und einmal nicht annotiert. Also, für den Task 1 kombiniert Kristina annotiertes Russisch und annotiertes Ukrainisch. Beim Task 2 kombiniert sie annotiertes und nicht-annotiertes Russisch. Beim Task 3 kombiniert die Studentin die Muster aus allen drei Kategorien. Training Data wurde aus CoNLL Sigmorphon 2016 Shared Task genommen im Format „lemma – target form – morphosyntactic description“. Kristina hat ein Beispiel angeführt, wie der Task aussieht:

- Das angegebene Lemma: „предмет“ („der Gegenstand“)
- Input: „родительный падеж, единственное число“ („Genetiv Singular“)
- Output: „предмета“ („des Gegenstandes“)

Um den ersten Task durchzuführen, nimmt Kristina 50 russische / 50 ukrainische Wörter bis 50 russische / 12800 ukrainische Wörter. Russisch gilt in dem Fall als Low-Resource-Sprache und Ukrainisch als High-Resource-Sprache. Also wurde Task 1 mit der ersten Methode durchgeführt, und zwar nimmt Kristina Ukrainisch zu Hilfe. Der zweiten Methode entspricht Task 2: 50 russische annotierte / 50 russische nicht-annotierte bis 50 annotierte / 12800 nicht-annotierte Wörter, also nimmt sie für das Experiment die gleiche Sprache, aber nicht-annotiert. Für Task 3 verwendet die Studentin die Muster aus den ersten zwei Tasks: sie nimmt die gleiche Anzahl von russischen annotierten / ukrainischen annotierten / russischen nicht-annotierten Wörtern wie in Tasks 1 und 2. Die Ergebnisse, die Kristina nach der Durchführung des Experiments bekommen soll, werden dann graphisch dargestellt. Für die Evaluierung stellt sich Kristina zwei Fragen. Eine davon ist eine linguistische Frage: welche Fehler treten auf. Die Informationen, die die Studentin bekommt, können verwendet werden um diese Fehler zu beheben und damit auch die Ergebnisse zu verbessern. Die zweite Frage ist, ob es irgendwelche Testmuster gefunden werden können, die verwendet werden, um das System zu verbessern.