Katja Bertholdt
CIS, LMU München
22.05.17

Protokoll zur Sitzung vom 22.05.17 – Computerlinguistik Kolloquium

3. Vortrag: Alexander Vodermaier, „Comparison of Transfer Methods for low Ressource Morpholgy“

Den letzten Vortrag des Tages hält Alexander Vordermaier. Seine Arbeit trägt den Titel „Comparison of Transfer Methods for low Ressource Morpholgy“ und wird von Katharina Kann betreut. In seiner Arbeit geht es um die Paradigmen Komplettierung von Sprachen, also die Zuordnung eines Lemmas zu seinen flektierten Formen. Es gibt bereits Systeme, die diese Zuordnung erledigen, jedoch sind dazu viele Daten benötigt. Vordermaier beschäftigt sich mit der Überlegung, diese Paradigmen Komplettierung auch mit weniger Daten lösen zu können, wenn in einer Sprache nicht genug Ressourcen zur Verfügung stehen. Hierbei bezieht er sich konkret auf Bulgarisch als sog. „High Ressource“ Sprache (HRS), in der viele Daten zur Verfügung stehen, und Mazedonisch als „Low Ressource“ Sprache (LRS), in der es nur wenige Daten gibt. Wichtig ist bei der Wahl dieser Sprachen, dass sie miteinander verwandt sind. Auf Nachfrage aus dem Publikum erklärt Vordermaier, dass sich diese Verwandtschaft vor allem auf den morphologischen Bereich beziehen muss (zb. Kasussystem und Endungen). Es soll getestet werden, ob es möglich ist, durch ein gemeinsames Trainieren von HRS und LRS, gute Ergebnisse für die Paradigmen Komplettierung der LRS zu erzielen.
Vordermaier zieht für sein Vorgehen drei Methoden heran: Die erste Methode heißt „Sprachübergreifende Paradigmen Komplettierung“. Dabei werden wenige annotierte Daten der LRS mit annotierten Daten der HRS gemischt und zusammen trainiert. Vordermaier kombiniert dafür jeweils 50 oder 200 Samples der LRS mit Sets unterschiedlicher Größe der HRS, sodass alle Kombinationen aus den verschieden großen Sets berücksichtigt werden.
Die zweite Methode ist das sogenannte „Auto Encoding“. Hierbei handelt es sich um ein Verfahren, bei dem die Eingabe gleichzeitig die Ausgabe ist. Es unterstellt, dass die Flektion der Wörter in HRS und LRS identisch ist. Auch hier werden die Daten der HRS und LRS auf gleiche Weise wie bei der ersten Methode miteinander kombiniert. Jedoch ist die HRS in dieser Methode nicht annotiert. Als dritte Methode werden die beiden ersten Methoden miteinander kombiniert. Es werden also annotierte Daten der LRS mit annotierten und nicht annotierten Daten der HRS kombiniert. Bei der Kombination muss darauf geachtet werden, dass von den HRS Daten jeweils gleichgroße Samples aus annotierten/ nicht annotierten Daten mit den 50 bzw. 200 Samples der LRS kombiniert werden.
Die Daten, mit denen Vordermaier arbeitet sind jeweils in zwei Teile aufgeteilt: in Source File und Target File. Die Source File enthält alle Lemmas zur Target File und die Target File enthält alle flektierten Formen zum Source File. Während in der Source File der Sprachübergreifenden Paradigmen Komplettierung genaue morphologische Angaben über den Input und somit auch den erwünschten Output gemacht werden, ist in der Source File des Auto Encodings dem Lemma lediglich ein „copy“ vorangestellt. In der Kombination besteht die Source File aus beiden Arten der Daten.
Seine bisherigen Ergebnisse präsentiert Vordermaier in zwei unterschiedlichen Diagrammen. Das erste Diagramm zeigt die Genauigkeit der Vorhersage bei den Kombinationen mit dem Set aus 50 Daten der LRS. Hierbei ist zu sehen, dass das Auto Encoding wie erwartet am schlechtesten abschneidet und nur knapp die 20% Marke erreicht. Bei der Sprachübergreifenden Paradigmen Komplettierung erreicht die Genauigkeit fast 50%. Die Kombination führt zu einer nochmaligen leichten Steigerung der Genauigkeit, jedoch liegt die Genauigkeit hier auch nur knapp über 50%. Im zweiten Diagramm sind die Ergebnisse der Kombinationen mit den 200 Daten der LRS dargestellt,

die bereits deutlich besser ausfallen als bei der kleineren Datenmenge. Das Auto Encoding erreicht hier knapp 60%, die Kombination erreicht fast 80% Genauigkeit. Hierbei ist erstaunlich, dass die Prozentzahl des Auto Encodings eine so positive Entwicklung genommen hat. Da Vordermaier weder die LRS noch die HRS beherrscht, fällt es ihm jedoch schwer, diesem Phänomen genau auf den Grund zu gehen.
Vordermaier will in der nächsten Zeit weitere Fehlerquellen identifizieren und andere bereits bestehende Verfahren beleuchten und mit seinem Verfahren vergleichen.