

Fortgeschrittene Methoden der statistische maschinelle Übersetzung

Übung 5

Abgabe (pdf, Zip) an kaeshammer@phil.uni-duesseldorf.de
bis einschließlich Sonntag, 25.01.2015

1

Sie haben den Auftrag ein **Übersetzungssystem Deutsch → Französisch** zu erstellen. Die Daten, die Ihnen dafür zur Verfügung stehen, sind eine Million deutsch-französische Satzpaare und zusätzlich eine Milliarde französische Sätze. Diese Daten sind bereits satzsegmentiert und die parallelen Daten sind satzaliniert. Am Ende sollen Sie nicht nur das Übersetzungssystem liefern, sondern Ihrem Auftraggeber auch berichten, *wie gut* das System funktioniert.

Nehmen wir an, Sie entscheiden sich für ein phrasenbasiertes Standard-Übersetzungssystem. Welche Tools/Software brauchen Sie für diesen Auftrag? (Z.B. phasenbasierter Decoder, Tokenisierer, Phrasenextraktion etc. Denken Sie dabei an *alle* Teilpunkte, die wir in den vergangenen Wochen behandelt haben.) Welche Eingaben/Ausgaben haben diese Tools? In welcher Reihenfolge wenden Sie die Tools an? Wie setzen Sie die Daten ein?

Fertigen Sie ein **Flussdiagramm** an, das diese Fragen beantwortet.

2

In Bezug auf Aufgabe 1: Jemand schlägt Ihnen vor, syntaxbasierte Umordnung als Vorverarbeitungsschritt anzuwenden. Erscheint Ihnen dies sinnvoll? Welche zusätzlichen Tools würden Sie brauchen? Erweitern Sie Ihr Flussdiagramm um die entsprechenden Schritte.