Language Technology

8. Übung

Prof. Tobias Scheffer
Uwe Dick

Sommer 2016

Ausgabe am: 14.06.16
Besprechung am: 20.06.16

Aufgabe 1 *Phrase-based Model*

Gegeben sei ein (kleiner) paralleler Corpus mit alignierten Sätzen auf Deutsch und Marsianisch.

$$\langle \text{ich habe das kleine Haus gekauft}, \text{gnarf hanv glub kaff} \rangle \quad (1)$$

$$\langle \text{ich habe ein Haus gekauft}, \text{gnarf hanv kaff} \rangle \quad (2)$$

$$\langle \text{das Haus wurde mir verkauft}, \text{hanv gnarf keffa} \rangle \quad (3)$$

$$\langle \text{ich habe einen Ball geklaut}, \text{gnarf gronff kaff} \rangle \quad (4)$$

Außerdem haben Sie Schätzungen von Wort-zu-Wort-Übersetzungswahrscheinlichkeiten von Marsianisch nach Deutsch gegeben. Im Folgenden ein Auschnitt. Wahrscheinlichkeiten sind beispielhaft angegeben als $P(ich|gnarf) = 0,6$.

gnarf	ich	0.6	hanv	das	0.03
gnarf	mir	0.2	hanv	des	0.01
gnarf	mich	0.2	hanv	ein	0.03
glub	kleine	0.2	kaff	gekauft	0.3
glub	winzig	0.1	kaff	kaufte	0.05
glub	klein	0.2	kaff	habe	0.05
gronff	Ball	0.4	kaff	geklaut	0.3
gronff	einen	0.1	kaff	genommen	0.1
gromp	kleine	0.3	keffa	verkauft	0.5
hanv	Hütte	0.1	keffa	wurde	0.1
hanv	Haus	0.2	keffa	gekauft	0.4
hanv	Hauses	0.1	fogro	wurde	0.05

Ihre Aufgabe ist es, ein phrasenbasiertes Übersetzungsmodell für die Übersetzung von Deutsch nach Marsianisch zu lernen. Gehen Sie dabei wie folgt vor.

1. Als erstes müssen Sie also ein Wort-zu-Wort-Alignment erstellen. Verwenden Sie dazu das IBM Model 1 (S.32,33 in den Folien) und die oben bereitgestellte Übersetzungstabelle.

2. Anschließend sollten Sie, ausgehend vom Alignment, Phrasenpaare extrahieren (S.38ff in den Folien).

3. Als letzten Schritt müssen noch Phrasenübersetzungswahrscheinlichkeiten berechnet werden (Folien S.45). Bedenken Sie, dass sie ein Übersetzungsmodell von Marsianisch nach Deutsch lernen wollen. Die Phrasenübersetzungswahrscheinlichkeiten sollten also von Deutsch nach Marsianisch sein.

Aufgabe 2 *Phrase-based Model*

Nun sollen Sie auf Grund des Übersetzungsmodells den folgenden Satz übersetzen:

$$\text{hanv glub gnarf keffa} \tag{5}$$

Ihr Dekodierer hat die Übersetzung *das kleine Haus wurde mir verkauft* geliefert, wofür die Phrasenpaare ⟨hanv glub,das kleine Haus⟩ und ⟨gnarf keffa,wurde mir verkauft⟩ verwendet wurden. Wie groß ist der Score $P(T, a|S)$ (S.18) für diese Übersetzung, wenn Sie ein deutsches 2-gram Sprachmodell annehmen, das im Folgenden ausschnittsweise gezeigt wird?

$$P(\text{das}| < START >) = 0.1, P(\text{kleine}|\text{das}) = 0.2, P(\text{Haus}|\text{kleine}) = 0.3, \tag{6}$$

$$P(\text{wurde}|\text{Haus}) = 0.1, P(\text{mir}|\text{wurde}) = 0.2, P(\text{verkauft}|\text{mir}) = 0.1 \tag{7}$$

Verwenden Sie zur Berechnung die Formeln auf den Seiten 17 und 18 aus der Vorlesung. Der distance score ist definiert als $d(x) = 0.5^{|x-1|}$.

Aufgabe 3 *BLEU Score*

Berechnen Sie den jeweiligen BLEU score für die folgenden beiden *candidate* Übersetzungen.

Cand 1:	Wir sagen sonst nicht, weil das Runde muss ins Eckige
Cand 2:	Wir reden nicht mehr, das Eckige muss ins Runde
Ref 1:	Wir haben sonst nichts zu sagen, denn am Ende muss das Runde ins Eckige
Ref 2:	Das ist alles, was wir zu sagen haben, denn am Ende muss das Runde ins Eckige
Ref 3:	Ansonsten ist Schweigen, da das Runde am Ende ins Eckige muss