Fortgeschrittene Methoden der statistische maschinelle Übersetzung

Übung 3

Abgabe (pdf, Zip) an `kaeshammer@phil.uni-duesseldorf.de`
bis einschließlich Sonntag, 07.12.2014

1

Gegeben ist folgendes Satzpaar mit Wortalinierung (Positionen beginnen bei 1):

$\vec{e}$: `she accepts the challenge`
$\vec{f}$: `sie nimmt die Herausforderung gerne an`
$A_{(\vec{e},\vec{f})} := \{(1,1),(2,2),(3,3),(4,4),(2,6)\}$

(a) Stellen Sie das Satzpaar und die Alinierung als Alinierungsmatrix dar. Listen Sie alle Phrasenpaare $(\bar{e},\bar{f})$ auf, die bezüglich dieser Alinierung wohlgeformt sind.

(b) Gehen Sie davon aus, dass Ihr Korpus $\mathfrak{K}$ nur dieses eine Satzpaar enthält. Schätzen Sie die Phrasenübersetzungswahrscheinlichkeiten $\hat{\phi}(\bar{f}|\bar{e})$ für alle Phrasenpaare $(\bar{e},\bar{f})$ im Korpus. Als Erinnerung:

$$\hat{\phi}(\bar{f}|\bar{e}) = \frac{C_{\mathfrak{K}}(\bar{e},\bar{f})}{\sum_{\bar{f}' \in V_F{}^+} C_{\mathfrak{K}}(\bar{e},\bar{f}')} = \frac{C_{\mathfrak{K}}(\bar{e},\bar{f})}{C_{\mathfrak{K}}(\bar{e})}$$

2

Mit einem parallelen Korpus "normaler Größe" ($>$ 1 Mio. Satzpaare) gibt es beim Schätzen der Übersetzungswahrscheinlichkeiten ein praktisches Problem: Es ist nicht möglich alle Phrasenpaare gleichzeitig im Arbeitsspeicher zu haben. Man kann $C_{\mathfrak{K}}(\bar{e},\bar{f})$ also nicht auf naive Art und Weise berechnen.

Stattdessen schreibt man nach und nach alle Phrasenpaare, die man aus $\mathfrak{K}$ extrahiert, in eine (große) Textdatei. (Diese befindet sich auf der Festplatte, nicht im Arbeitsspeicher, also kein Problem!) Sie sieht zum Beispiel so aus:

```
michael ||| michael
assumes ||| geht davon aus
...
michael ||| michael
...
michael ||| unser michael
...
```

Die Frage ist, wie kann man von dieser Datei ausgehend $\hat{\phi}(\bar{f}|\bar{e})$ berechnen, ohne die ganze Datei auf einmal einzulesen und zu speichern? Tipp: es existieren effiziente Sortieralgorithmen, die diese große Datei zeilenweise sortieren können (ohne sie komplett im Arbeitsspeicher zu haben, z.B. *mergesort*). Ausgehend von einer sortierten Datei (alle Phrasenpaare mit z.B. der englischen Seite `michael` stehen untereinander), wie kann man $\hat{\phi}(\bar{f}|\bar{e})$ berechnen (ohne die komplette Datei auf einmal einzulesen und zu speichern)? Skizzieren Sie den Algorithmus als Pseudocode.

3

Für die Anzahl der wohlgeformten Phrasen über einer Wortalinierung spielen die Worte selbst natürlich keine Rolle: es reicht, wenn wir die Alinierungsrelation (über den Positionen als Zahlen) gegeben haben. Ihre Aufgabe ist es, gegeben eine Alinierungsrelation, herauszufinden, *wie viele* wohlgeformte Phrasen es für das Satzpaar gibt. Die Länge der zugehörigen Sätze entspricht jeweils der höchsten Zahl, die in der entsprechenden Komponente der Relation vorkommt, also für a) haben wir Sätze der Länge 7 und 6, für b) der Länge 7 und 8, und für c) der Länge n und n.

a) $A_1 := \{(1,1),(2,2),(3,2),(4,4),(2,3),(5,4),(5,6),(6,3),(7,2)\}$

b) $A_2 := \{(7,8)\}$

c) $A_3 := \{(1,1),(2,2),(3,3),...,(n,n)\}$, für beliebige $n \in \mathbb{N}$. Die Lösung ist hier natürlich keine Zahl, sondern eine monoton steigende Folge (Funktion $\mathbb{N} \to \mathbb{N}$), sowas wie $4n + 3$. Versuchen Sie am besten, sich das Wachstum der Möglichkeiten graphisch zu veranschaulichen!