Stemmingverfahren:
Ein Porter-Stemmer für das Ungarische

Demonstration

Éva Mújdricza

Information Retrieval, HS, WS07/08
Ruprecht-Karl-Universität Heidelberg
Dozentin: PD Dr. Karin Haenelt

04.02.2008

Übersicht

- Die Implementierung
- Die Sprache
- Problematische Fälle
- Demonstration
- Evaluierung
- Fehleranalyse und Optimierungsmöglichkeiten
- Quellen

Danksagung

- Ich bedanke mich bei Carina Silberer, die mir bei der Sonderzeichenbehandlung sowie bei der Suche nach Korpora und Evaluierungsliteratur viel geholfen hat!

Éva Mújdricza

Die Implementierung

- Prinzip: light1-Stemmer von Anna Tordai (University of Amsterdam)
  - Ursprüngliche Beschreibung: http://snowball.tartarus.org/algorithms/hungarian/stemmer.html
  - Porter-Stemmer: Endungsentfernung
- Implementierung in Python – mit leichten Veränderungen
- Nur **nominale Flexionsendungen** werden abgeschnitten (Person, Numerus, Kasus, Possessivendungen)
- Schritte
  - 1. Vorbereitung: Kleinschreibung, Festlegung von R1
  - 2. Bearbeitung: 9 Schritte, nicht iterativ
  - keine Nachbereitung (bzw. Nachbereitung in den Bearbeitungsschritten eingebaut)

Die Sprache

- Das Ungarische ist eine stark **agglutinierend**e Sprache: die grammatischen Funktionen werden mit Hilfe von Affixen ausgedrückt.
  - Es gibt ausschließlich Suffixe. (Ausnahme: Superlativ)
  - **(nominale) Flexionsendungen** gibt es zweierlei Arten:
    - „jel”: es können auch mehrere hintereinander stehen:
      - Person und Numerus
      - Possessivendungen
      - Komparation (– hier nicht betrachtet)
    - „rag”: eine wortformschließende Endung mit Festlegung der grammatischen Funktion der Wortform.
      - Kasusendungen (18-24 Kasus – dafür gibt es keine Präfixe)
  - Der Stamm verändert sich in der Regel nicht (oder nach beschreibbaren Regelmäßigkeiten), aber die Endungsvielfalt ist ziemlich groß.

Problematische Fälle

- **Stammvarianten**: [Beispiele für nominale Elemente mit Akkusativendung]
  - Vokalverkürzung: *kéz* + *-et* = *k**e**zet* [Hand$_{Akk}$]
  - Vokalverlängerung im Auslaut: *körte* + *-t* = *kört**é**t* [Birne$_{Akk}$]
  - Vokaleliminierung: *hár**o**m* + *-at* = *hármat* [drei$_{Akk}$]
  - Konsonanteneinschiebung (Hiatvermeidung): *ló* + *-at* = *lo**v**at* [Pferd$_{Akk}$]

Problematische Fälle

- **Suffixvarianten**: abhängig
  - von den Vokalen des Stammes – Vokalharmonie: *ház-at* [Haus_{Akk}], *kez-et* [Hand_{Akk}]
  - von dem Auslaut des Stammes: *ház-at, körté-t*; mit Assimilation: *ház-zal* [Haus + mit], *virággal* [Blume + mit]
  - vom Stammauslaut und Stammvokal: *-n/-on/-en/-ön* [auf_{Dat}/an_{Dat}]: *körté-n, ház-on, kéz-en, föld-ön* [auf (dem) Boden]
- **falsche Ersetzung** von „langen” Konsonanten durch „kurze”: [Beispiel für Nomen + Instrumentalis (mit)]
  - *könny* + *-vel* = *könnyel* >$_{\textit{1.Schritt}}$ **köny* [Träne], weil sonst:
  - *fény* + *-vel* = *fénnyel* > *fény* [Licht], *ház* + *-val* = *házzal* > *ház*

Evaluierung

- **Korpus**: zufällig ausgewählte Texte aus der ungarischen Wikipedia
  - 11306 Tokens
  - 8533 Stämme – 75,47 % der ursprünglichen Tokenanzahl
- zum Evaluieren: zufällige Auswahl von 200 Stämmen (nicht repräsentativ)
- **Ergebnisse**:

		67,5% (135)	32,5% (65)
		Entfernung	**keine Entf**
29% (58)	**korrekt**	32	26
61% (122)	**Unterstemming**	83	39
10% (20)	**Überstemming**	20	0

- Korrektheit: 29 % – Kein gutes Ergebnis.
  - Grund: unerkannte Sonderzeichen für Vokale (*á, é, í, ó, õ[ő], ú, û[ű]*) → auch sehr wenig Überstemming.

Evaluierung: Fehleranalyse

- Positive Beispiele: ohne Sonderzeichen in der Endung:
  - *alap :['alapjai', 'alapon', 'alap', 'alapja', 'alapokon', 'alapok', 'alapul']* [Grund, Grundlage + Endungen]
  - *idõ :['idõ', 'idõknek', 'idõt', 'idõben', 'idõk', 'idõkben', 'ldõvel', 'idõkre', 'idõnek', 'idõket', 'idõre', 'idõnkig', 'idõn', 'idõvel']* [Zeit + Endungen]
- Sonderzeichen wurden nicht erkannt → Endungen mit Sonderzeichen nicht gestemmt: *kezünkbõl* > *kezünkbõl* statt *kezünk* + *-bõl* [unsere Hand + aus_{Dat}]
- nicht die ganze Endung erkannt oder Endungen falsch erkannt: *kimenni* > *kimenn* statt *kimen* + *-ni* [hinausgeh + -en]; *ismerjük* > *ismerjü* statt *ismer* + *-jük* [kenn + (wir+bestimmte Konj.)] – Diese sind verbale Endungen, die von diesem Stemmer auch nicht erkannt werden müssen.
- Stammteil als Endung erkannt: *kivált* [auslösen, loskaufen] > *kivál* [ausscheiden]*; unalom* [Langeweile] > *unal* [?]*; csendet* [$Stille_{Akk}$] > *csen* [mopsen, stehlen]

Evaluierung: Optimierung

- Sonderzeichenbehandlung korrigieren
- Reihenfolge der Endungsentfernung ändern
- zusätzliche Schritte für Entfernung von weiteren Endungen: z.B. Komparationsendungen
- auch Derivationsendungen einbeziehen
- Stoppwortliste erstellen
- Ausnahmen explizit behandeln

Quellen

- Keszler, Borbála u.a. (2000): *Magyar grammatika*. [Ungarische Grammatik] Budapest : Nemzeti Tankönyvkiadó.
- Tordai, Anna (2006): *Stem, Stemming, Stemmer. On the benefits of Stemming in Hungarian.* (Masterarbeit, University of Amsterdam) (www.cs.vu.nl/~atordai/Scriptie.pdf (Stand: 26.12.2007)

- http://snowball.tartarus.org/algorithms/hungarian/stemmer.html (Stand: 26.12.2007)
- http://snowball.tartarus.org/algorithms/hungarian/stem_ISO_8859_1.sbl (Stand: 12.12.2007)
- http://www.unine.ch/info/clef/hungarianStemmer.txt (Stand: 12.12.2007)
- http://snowball.tartarus.org/otherlangs/german_py.txt (Stand: 26.12.2007)