RO-Stemmer mit Snowball

Abschlussvortrag

05.07.2006

Doina Gligă
Erwin Glockner
Marina Stegărescu

Inhaltsübersicht

- Stemmer: Idee des Algorithmus, Porters Snowball, Stemdefinition, Problematik
- Rumänisch: Flexionsstruktur
- Implementierung: Pseudocode
- Evaluation

Was haben wir gemacht?

- Stemmer in Snowball für Rumänisch

Was sind Stemmer?

- Programme, die Wörter auf ihren gemeinsamen Kern zurückführen
- vor allem als Komponente der IR-Systeme entwickelt und benutzt
- Linguistische Analyse

Stemmer

- Lexikonbasierte
- Korpusbasierte
- Regelbasierte

Regelbasierter Stemmer

- Verfahren → Porters Algorithmus
- Sprache → Snowball

Porters Algorithmus

- Menge von Verkürzungsregeln: Bedingungen und Ableitungen für verschiedene Suffixe
- Vokal-Konsonant-Sequenzen
- Regelanwendung

Snowball

- Snowball: stringverarbeitende Sprache
- ermöglicht das einfache und exakte Repräsentieren von Stemmingalgorithmen
- entwickelt von Martin Porter

Stem

- Nicht mit dem linguistischen Stamm identisch, da Derivationssuffixe nicht immer entfernt werden
- Vielmehr ein gemeinsamer Kern von Termen

Problematik

- Entsteht dadurch, dass eine Zeichensequenz sowohl als Suffix, als auch als Teil des Kerns vorkommen kann
- Keine Formalisierung für diese semantische Unterscheidung möglich, da die Wörter von der Form her die gleiche Struktur aufweisen

  z.B. capital vs lovit, iscusitul; castană vs americană; palatalizare vs spălat

Ziel

Minimierung von:

- Overstemming: zu lange Zeichenkette wird abgeschnitten
  militar -> mil statt milit
- Understemming: zu kurze Zeichenkette wird abgeschnitten
  robotizată -> robotiza statt robot
- Misstemming: subjektiv

Stemmer fürs Rumänische

- Was ist die Idee hinten dem Code?
  - Begriffe und Definitionen
  - Was ist fürs Rumänische zu beachten?
  - Was wollen wir erreichen?
- Pseudo-Algortihmus

Begriffe

- Wurzel (Root) - die Sequenz des Wortes, die nicht mehr zerlegbar ist und in lautlicher und semantischer Hinsicht, als Ausgangsbasis entsprechender Wortfamilie angesehen wird
  - Cânt-a (singen)
- Stamm – Morphem oder Morphemkonstruktion, an die Flexionsendungen treten können
  - Descânt-a (durch Sprüche Zauber verzaubern)
- Flexionselemente – die Menge aller Elemente, die in paradigmatischer Relation mit dem Stamm eines Wortes sind, und das Flexionsparadigma dieses Wortes bildet
- Flexionsparadigma – die Menge aller Flexionsformen des Wortes

Porter's Begriffe

- a – Suffixe -> mit dem Wort zusammen gebundenen Suffixe (enclitics)
  - Italienisch, Spanisch, Portugiesisch
  - it. mandarglielo = mandare + ***gli*** + ***lo*** = to send + it + to him
  - Rumänisch – Bestimmte Artikel
- i – Suffixe (Inflektionssuffixe)
  - fit + ***ed*** -> fitted (doppel ***t***)
  - love + ***ed*** ->loved (***e*** final von love verschwindet)
- d – Suffixe (Derivationssuffixe)
  - Englisch: ***-ly*** -> greatly, kingly
  - Französisch: ***-ement*** -> rapprochement)

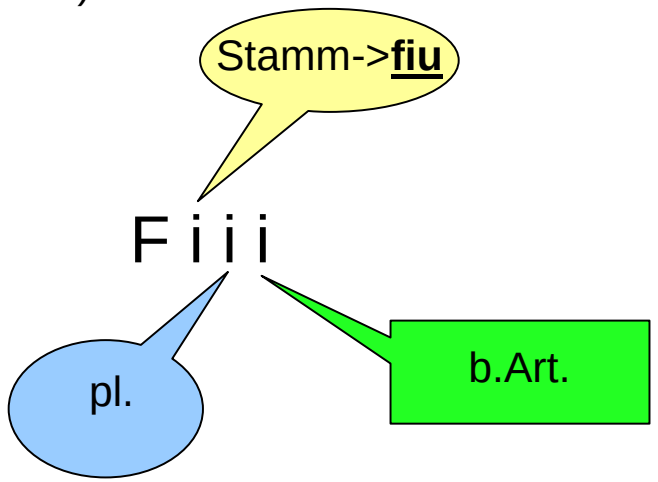
Das Rumänische

- Romanische Sprache
- 7 Vokalen <a,e,i,ă,â/î,o,u>
- 22 Konsonnanten <ş, ţ>
- Flexionsstruktur und Derivationsstruktur – umfangreich und multistratal
  - P:: Stamm +(Vok)+ (Suffix) + (Suffix) + Flektionsmarker
  - 1-3 Stämme
    - Pom <sg, o.Art> pom -i <pl., o.Art> pom-u-lui <sg. Art. G/D>
    - Fat-a fet-e
    - Om <sg, o.Art> oamen -i
    - Frumos <sg., m.> frumoş –l <pl.m.> frumoas-e <pl. f.>

Probleme

- Homonymie der Endungen:
  - -e: <+pl, N> case (Häuser)
    <+inf, V> merge (gehen)
    <+stamm> bine (gut)
  - -i <+pl,N,Adj> copaci (bäumer) frumoşi (schöne (Kinder)
    <+b.Art,N>
    <+inf, V> iubi (lieben)
    <+2.Pers,sg,V> mergi (du gehest)
    <+stamm> crai (Prinz)

- ar <Suffix> familiar
  < +stamm> marar Dill)



Ziel des Porters Stemmer vs. Rumänischen Stemmer

- „Usually we want to remove all a- and i-suffixes, and **some of** the d-suffixes.“
- Fürs Rumänische -> a-Suffixe, i-Suffixe, d – Suffixe (vielfältiger als im Englischen oder Deutschen)
  - **stabil**
  - **stabil**i
  - **stabil**it
  - **stabil**ire
  - **stabil**ibil
  - stabiliza
  - stabilizat
  - stabilizant
  - stabilizare
  - stabilizator

Pseudocode

- Diakritika integrieren
- die feste unstemmbare Grenze des Strings definieren

  /* Wörter von zwei Buchstaben nicht betrachten; r1 – ab dem ersten Konsonant, dem ein Vokal in dem String folgt; r2 – nach dem ersten in r1 eines Vokals folgender Konsonant* /
- Wenn String nicht in Exception1 vorhanden und wenn String > als 2 Buchstaben:

  /* Exc1-> stopp words und Wörter die eine a- und i-Funktion verletzen */
  - Suche nach **a** und **i** Suffixe und, wenn gefunden, entferne sie
  - Suche nach **d** Suffixe und, wenn vorhanden, entferne sie
  - Suche nach Suffixe, die zusätzlich in dem Derivationsprozess erscheinen können (munc*i vs. munc**i**tor*), oder nach Suffixe [+Partizip] (->verbale, adjektivale oder substantivale Distribution: mâncat, mâncata, mâncatul), und wenn vorhanden, entferne sie

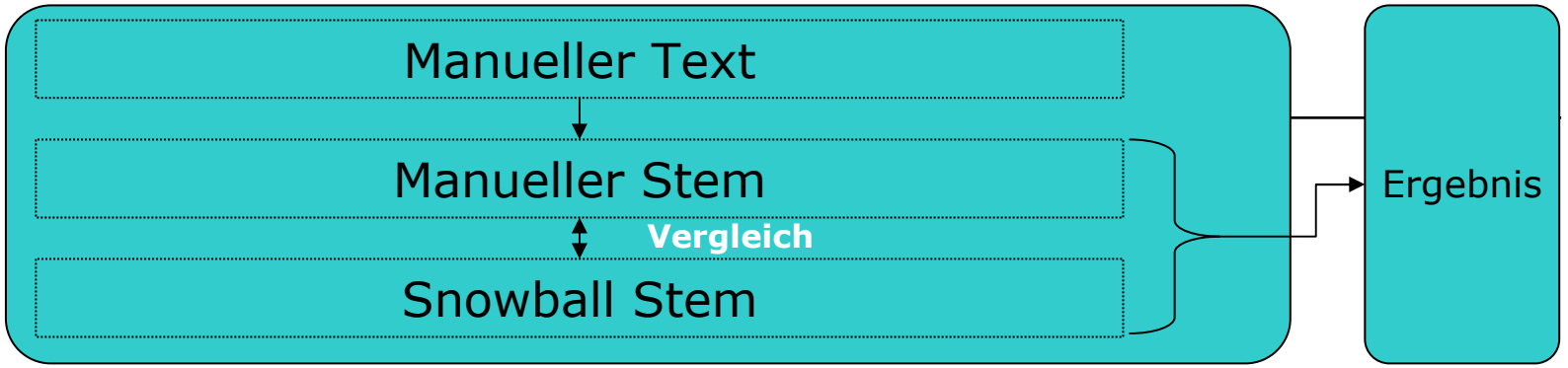
Evaluierung

- Vorbereitung
- Durchführung
- Zusammenfassung

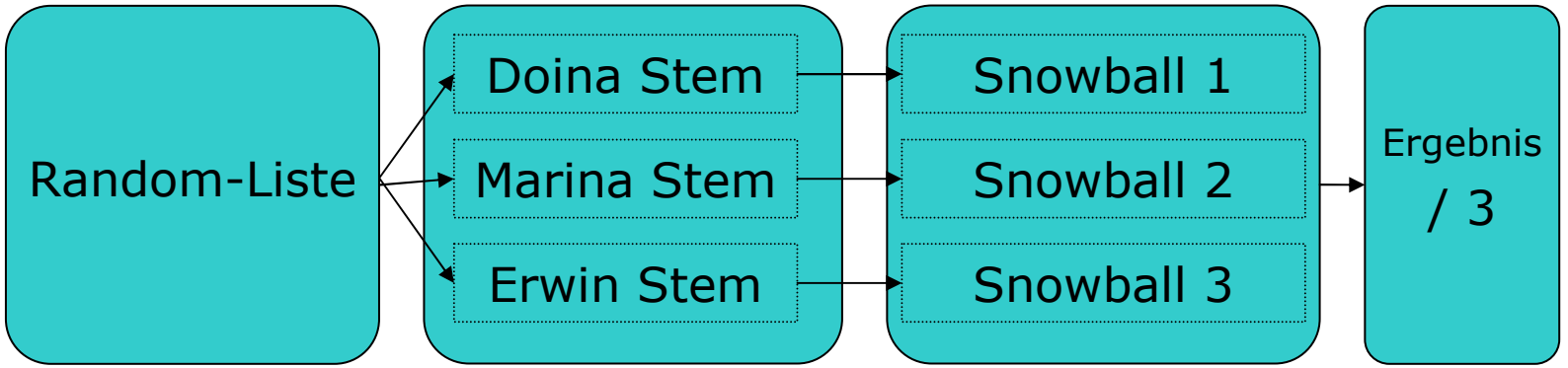
Vorbereitung/Durchführung

- Sammlung von Online-Texte
- Text in Wörter gesplittet, Wortwiederholungen (identische Wörter) und Satzzeichen entfernt
- Stopwords entfernt
- Manuell gestemmt
- Mit Snowball gestemmt
- Vergleich manueller Stem mit Snowball-Stem => Prozentsatz

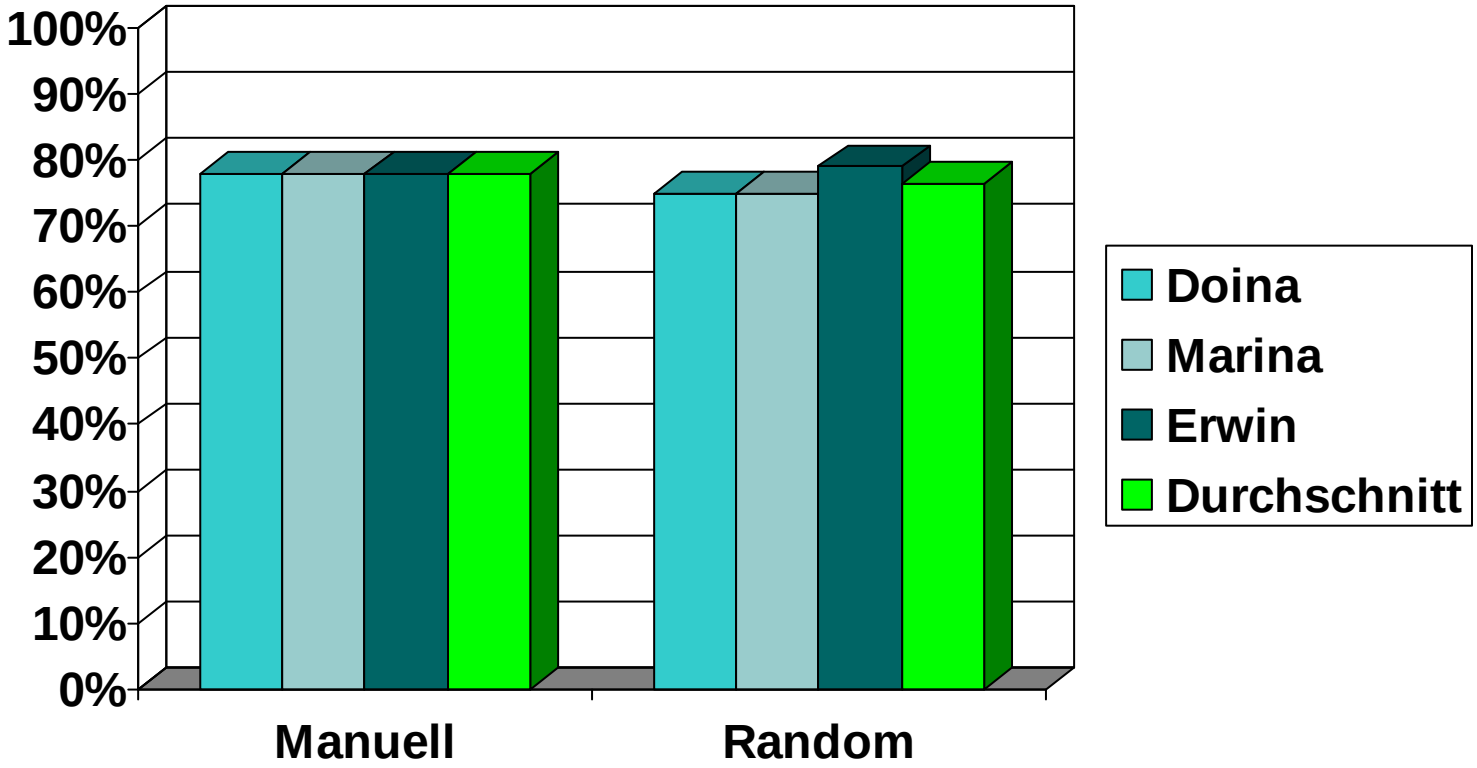
1. Methode



2. Methode



Statistik


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Manuell
Random
Doina
Marina
Erwin
Durchschnitt

Fazit

- Hohe Kompressionsrate
- schnell
- Problem wegen vielfältiger Inflektions- und Derivationsbasis
- Wenn man keine Deriovationsfälle betrachtet, sondern nur die Inflektionssuffixe => hohes Ergebnis
- Daten-Abgleich und manchmal Einigung im Team schwierig

Literaturverzeichnis

- Luciana Peev, Lidia Bibolar, Jodal, Endre, **A Formalization Model of the Romanian Morphology**
- *http://www.racai.ro/books/awde/peev.html*
- Jörg Meibauer & al. , **Einführung in die germanistische Linguistik**, Stuttgart, 2002
- I. Coteanu, **Limba română contemporană**, vol. I, Bucureşti, 1974
- *http://snowball.tartarus.org/*
- *http://kontext.fraunhofer.de/haenelt/kurs/InfoRet/index.html*