Acornym Decomposer

Spezifikationsvortrag

Programmier-Gesellstück

von

Branimira Nikolova

Seminar für Computrelinguistik
Ruprecht-Karls-Universität Heidelberg
SS 2006

Motivation

- Wieso Akronyme auflösen?
  - sehr verbreitetes Phenomen in fast jede Sprache
  - sehr produktives Teil des Lexikons
  - bereitet Schwierigkeiten bei der automatischen Textverarbeitung (NER, IE, etc.)
- Teilaufgabe der Eigennamenerkennung
- Akronyme sind ambig:

  BSE = Budapest Stock Exchange

  BSE = Bombay Stock Exchange

  BSE = Bovine Spongiforme Enzephalopathie (Rinderwahnsinn)

  BSE = Badische Stahl-Engineering GmbH

Definition

- Akronyme sind systematische Abkürzungen von (mehrteilige) Ausdrücke, meistens von Eigennamen wie z.B. Organisations- und Firmennamen oder Produktnamen.

Emerging Markets Data Plc (EMD) said on Tuesday it had agreed to an on-line partnership with British information service M.A.I.D Plc. The deal will allow M.A.I.D to globally distribute EMD's information and data via the firm's Corporate Profound and Profound for the Internet products.
"The relationship with M.A.I.D will have a considerable impact on turnover and profit," EMD said in a statement.
EMD is based in London and Dublin.

Akronyme vs. Abkürzungen

Akronyme werden explizit definiert	Abkürzungen - nicht
Akronyme werden generell großgeschrieben	Abkürzungen werden eher kleingeschrieben
Interpunktionszeichen werden bei Akronymen meistens weggelaßen	bei Ankürzungen - nicht
Akronyme werden so gelesen, wie sie im Text stehen	Abkürzungen werden beim Lesen aufgelöst
Meistens Namen	Eher selten Namen Ausnahme: Initiale von Personen

Akronymdefinition

- Eine Akronymdefinition ist das Vorkommen von einen Akronym - Vollform Paar im Text, mit der Ziel die Bedeutung des Akronyms einzuführen.

New set of rules concerning the purchase, transport and use of mercury substances has been released by the Department of Environment and Natural Resources (DENR).

Bildung von Akronyme

- Initial Matching: Übernahme der erste Buchstabe aus jeden Wort der Vollform in den Akronym

  BSE Budapest Stock Exchange

- Morpheme/Syllable Matching: Übernahme des Anfangs einer interne Morpheme/Silbe aus der Vollform in den Akronym (Komposita)

  BmLF Bundesministerium für Land- und Forstwirtschaft

  HTTP Hypertext Transfer Protocol

Bildung von Akronyme (2)

- Group Matching: Übernahme einer Gruppe von nacheinanderfolgenden Buchstaben aus einem Wort in der Vollform in dem Akronym

  ANASIN Associazione Nazionale Societa di Informatica

- Symbolic Matching:

  3M Minnesota Mining and Manufacturing Company

Teilaufagen

- Akronym Erkennung
- Vollform Erkennung
- Akronym - Vollform Matching

Bestimmung von Akronym-Vollform Paaren

- Textmarker: ()
  - Vollform (Akronym)
  - Akronym (Vollform)

- Cue words: or
  - electronic software distribution, or ESD.
  - DVD, or digital video disk

Bestimmung von Akronymkandidaten

- Die Länge des Tokens ist min 2 und max 10
- Der Token enthält mindestens eine große Buchstabe
- Der POS-Tag des Tokens ist: NNP, NNS, NN

Bestimmung von Vollformkandidaten

- Die Vollform muss in den selben Satz sein
- Die Vollform hat min(|A|+5, |A|*2) Tokens, wobei |A| die Zahl der Zeichen des Akronymkandidat ist
  - Die Vollform von kurze Akronyme(2 bis 4 Zeichen) soll nicht länger als |A|*2 sein
  - Die Vollform von lange Akronyme (5 und mehr Zeichen) soll nicht länger als |A|+5 sein
- Das erste Token der Vollform kann nicht ein Modalverb, ein Pronomen, eine Präposition oder eine Konjunktion sein
- Das erste Zeichen des ersten Tokens muss mit dem ersten Zeichen des Akronymkandidats übereinstimmen

Fall 1: Akronym in Klammern

Earlier, the Economic Intelligence and Investigation Bureau (EIIB) seized six truckloads of the cement on suspicion that there was ``an attempt to swing the imported cement to avoid the payment of duties and taxes."

- Akronymkandidat: Das Token in der Klammern (EIIB)
- SerachSpace: 4*2 = 8 Tokens ab der ersten Klammer rückwärts

Earlier, the Economic Intelligence and Investigation Bureau

Fall 2: Vollform in Klammern

Brace said high-speed digital ISDN (Integrated Services Digital Network) lines and overseas operations were growing at about 50-100 percent per year ...

- Akronymkandiat: Das Token vor der Klammer auf ISDN (
- SearchSpace: alle Tokens in den Klammern

  Integrated Services Digital Network

Fall 3: Akronym folgt cue word

Egghead joins virtual retailers such as Software.net, Strea, Internet Shopping Network and Cyberian Outpost that offer what the industry has dubbed electronic software distribution, or ESD.

- Akronymkandidat: Das Token nach „ , or “ ESD
- SearchSpace: 3*2= 6 Tokens ab den Komma rückwärts

  industry has dubbed electronic software distribution

Fall 4: Definition folgt cue word

DVD, or digital video disk, technology provides high-speed data transfer which is capable of storing seven times the amount data on a CD-ROM.

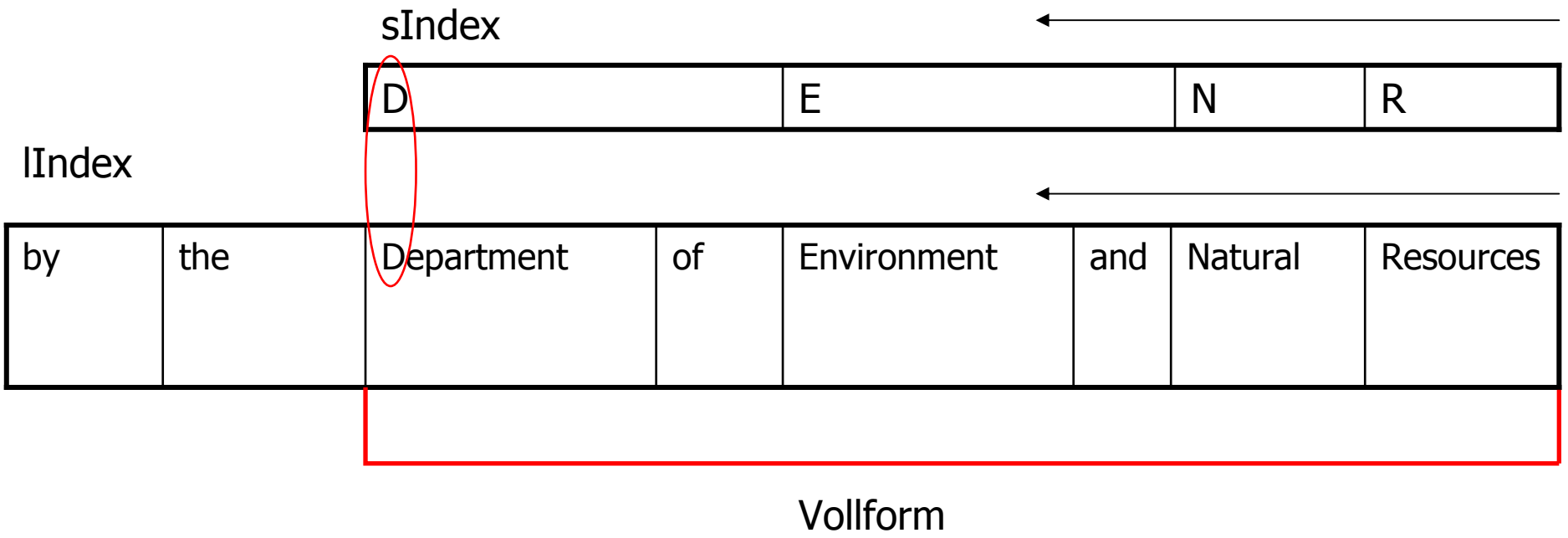
- Akronymkandidat: Das Token vor „ , or “ DVD
- SearchSpace: alle Tokens bis zu nächsten Komma oder Satzende

digital video disk

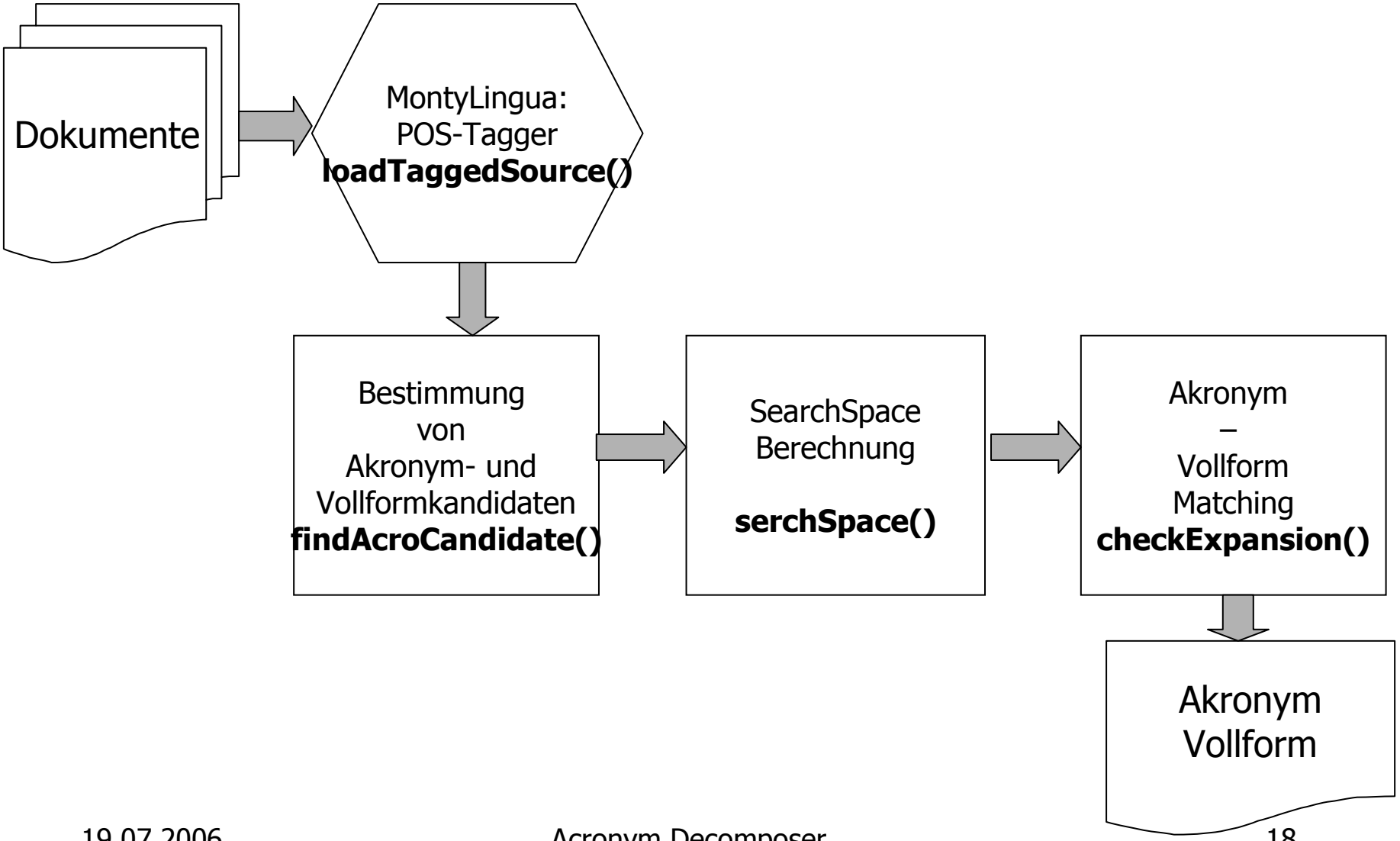
Algorithmus zur Ermittlung der korrekten Vollform

- Idee: Beginnend vom Ende des Akronyms und des Vollformkandidats, versuche die kurzeste Vollform zu finden, die mit dem Akronym übereinstimmt
- Jedes Zeichen im Akronym muss ein Zeichen in der Vollform entsprechen und in der gleiche Reihenfolge vorkommen
- Zwei Indizes: sIndex fürs Akronym und lIndex für den Vollformkandidat
- Für jede Position in sIndex wird lIndex dekrementiert, bis eine Übereinstimmung gefunden wird
- Jedes Mal wenn eine Position übereinstimmt weden beide Indizes dekrementiert
- Wenn sIndex die letzte Position erreicht hat, muss das erste Zeichen in der aktuelle lIndex String übereinstimmen

Algorithmus (2)



Ablauf


Dokumente
MontyLingua:
POS-Tagger
loadTaggedSource()
Bestimmung
von
Akronym- und
Vollformkandidaten
findAcroCandidate()
SearchSpace
Berechnung
serchSpace()
Akronym
–
Vollform
Matching
checkExpansion()
Akronym
Vollform

Externe Komponente

- Monty Lingua : Tokenisierung, Satzsegmentierung und POS-Tagging (Englisch)
  - POS-Tagger: eine Python-Implementierung von Brill‘s tbl Tagger (MontyTagger.py)
  - PENN TREEBANK tagset

- Geplant: Stemmer mit Kompositazerlegung für deutsche Texte

Status & To Do‘s

- Status:
  - loadTaggedSource():MontyTagger eingebunden
  - findAcroCandidate(): implementiert
  - searchSpace(): implementiert
  - checkExspansion(): implementier
- To Do‘s:
  - POS-Tagger und Stemmer (Deutsch)
  - Symbolic Matchig
  - Testen
  - Evaluieren ? (noch kein Evaluationskorpus gefunden)

Literatur

• Schwartz, Ariel; Hearst, Marti: A simple algorithm for identifying abbreviation definitions in biomedical text

• Park, Youngja; Byrd, Roy: Hibrid text mining for finding abbreviations and their definitions,

http://www.research.ibm.com/talent/documents/emnlp2001_48.pdf

• Zahariev, Manuel: A(Acronyms) PhD Thesis