Textzusammenfassung

07.01.2010

Einführung in die Computerlinguistik

Dominik Fischer

Gliederung

- 1. Wozu Textzusammenfassung?
- 2. Herangehensweise bei der Textzusammenfassung
- 3. statistische Herangehensweisen
- 4. linguistische Herangehensweisen
- 5. Probleme der Textzusammenfassung
- 6. Beispiel und Übung

Wozu Textzusammenfassung?

- „Informationsflut“ durch Verbreitung des Internets
- Unmöglich alle relevanten Texte zu einem Thema zu erarbeiten
- Lösung: automatische Zusammenfassung eines oder mehrerer Texte

Herangehensweise bei der Textzusammenfassung

- Aspekte der Textzusamenfassung:
  - Extract vs. Abstract (Extract: Sammlung von Sätzen des Originaltextes, Abstract: eigener Text)
  - Allgemein vs. nutzerorientiert
  - Informativ vs. Indikativ (Informativ: länger, auch Ergebnisse des Textes, Indikativ: kürzer, keine Ergebnisse)
  - Zusammenfassung eines vs. Zusammenfassung mehrerer Texte

Herangehensweise bei der Textzusammenfassung

- Grundsätzliche Vorgehensweise:
  - 1. Identifikation relevanter Textteile
  - 2. Textstellen in Textteilen isolieren
  - 3. Zusammenfassung durch Zusammenfügung der Textstellen erstellen
- Identifikation der relevanten Textteile durch statistische oder linguistische Methoden

Statistische Herangehensweisen

- Gewichtungswerte ermitteln durch:
  - Schlüsselwort- / Schlüsselphrasen-Methode (Autor gibt wichtige Fakten durch bestimmte Wörter / Phrasen zu erkennen)
  - Positionsmethode (Sätze in bestimmten Positionen wichtiger bspw. die ersten / letzten Sätze eines Textes, Überschriften)
  - Worthäufigkeit (häufige Worte abzüglich Stoppworte müssen wichtig sein)
  - Satzlänge (Lange Sätze wichtiger als kurze)

Statistische Herangehensweise

- Linguistische Erfodernisse bei den statischen Methoden:
  - Morphologie: Erkennen von Wortgrenzen und Worten
  - Lexikographie: Zuordnen von Worten zu Wortliste
  - Syntax: Erkennen von Satzgrenzen

Linguistische Herangehensweise

- Gewichtungswerte ermitteln durch:
  - Beziehungen zwischen Sätzen: Sätze die starken Bezug zu (vielen) anderen Sätzen haben sind wichtiger
- Linguistische Erfordernisse bei der linguistischen Methode:
  - Morphologie: Erkennen von Wortgrenzen und Worten
  - Lexikographie: Zuordnen von Worten zu Wortliste
  - Syntax: Erkennen von Satzgrenzen und -strukturen
  - Semantik: Erkennen von semantischen Beziehungen zwischen Sätzen

Probleme der Textzusammenfassung

- Isolation von relevanten Textteilen und –stellen: funktioniert einigermaßen
  - Probleme mit Ungenauigkeiten bei statistischen Methoden
  - Probleme mit der Erkennung semantischer und syntaktischer Strukturen bei den linguistischen Methoden

- Zusammenfassung durch Zusammenfügen relevanter Textstellen erstellen: problematisch
  - Semantische und syntaktische Probleme bei der Generierung kohärenter Fließtexte
  - Lösung: Übernahme von kompletten Sätzen des Originaltextes

Beispiel: SweSum

Übung

- Gehen Sie auf die Seite http://swesum.nada.kth.se/index-eng-adv.html
- Kopieren Sie beispielsweise den Text „Mamaia.txt“ in das Textfenster und experimentieren Sie mit den verschiedenen Einstellmöglichkeiten herum.

- Was fällt Ihnen auf?
- Welche Ihrer Meinung nach wichtigen Fakten fehlen in den Zusammenfassungen?
- Gibt es grammatikalische oder logische Probleme in den Zusammenfassungen?

Quellen

- http://archiv.tu-chemnitz.de/pub/2006/0118/data/Diplomarbeit_Endversion.pdf
- http://www.cl.uni-heidelberg.de/courses/archiv/ws06/ecl/folien/f_a22.pdf
- http://swesum.nada.kth.se/index-eng-adv.html