Korrekturprogramme

Von Emine Senol & Gihan S. El Hosami

Einleitung

Millionen von Texten werden mit dem Computern täglich erfasst

→ Fehler schleichen sich ein

Korrekturprogramme helfen diese

- zu finden
- zu korrigieren

Arten der Korrekturprogramme

Es gibt drei verscheidene Arten:

- Korrektur von „Nicht-Wörtern“
- Kontextabhängige Korrektur
- Grammatikkorrektur

Korrektur von „Nicht - Wörtern“

Anwendungsgebiet

Anwendung auf durch Tippfehler hervorgerufene Nicht-Wörter

Nicht-Wörter
= Zeichenketten ohne lexikalische Zuordnung

z.B. *Fehjler* anstatt *Fehler*

Allgemeines Verfahren

- Wörter werden mit einem Systemlexikon verglichen
- Nicht-Wörter werden aufgespürt
- Vorschläge zur Verbesserung werden gegeben
- Meisten Programme bieten auch Erweiterungsmöglichkeiten des Wörterbuches

Verfahren - Problem

Die bloße Verwendung eines Lexikons nicht ausreichend!

Grund:

Neubildung oder Änderung von Wörtern

- durch morphologische Prozesse
- durch Komposition und Derivation

Verfahren - Problem

→ Einige Programme besitzen eine Liste mit Flexions- und Derivationsaffixen

z.B. UNIX- Tool ispell

Aber:

Nur hilfreich bei Sprachen mit relativ wenigen und regelmäßigen Flexionsformen (z.B. Englisch)

Lösungsansatz

Zwei-Ebenen-Morphologie auf der Basis von endlichen Automaten

- Morphologie und Lexikon bilden einen großen endlichen Automaten
- Sehr erfolgreich zur Beschreibung vieler verschiedener Sprachen (z.B. Deutsch, Arabisch und Finnisch)

Verfahren

Korrekturverfahren der nicht im Lexikon vorhandenen Wörter:

Suche nach dem ähnlichsten String

- Basierend auf einer Funktion die den Abstand zwischen 2 Strings angibt

Andere Tippfehlerarten

Nicht immer entsteht ein Nicht-Wort:

- **Buchstabenvertauschung:**
  Die Erkennung vom Licht-Wörtern reicht nicht aus.
- **Groß- oder Kleinschreibung:**
  Die Korrekturprogramme Erkennen nicht alle Fehler.
- **Transposition:**
  Die Erkennung von Nicht-Wörtern riecht nicht aus.
- **Überschüssige Buchstaben:**
  Die Erkennung von Nicht-Wörtern reichst nicht aus.
- **Fehlende Buchstaben:**
  Die Erkennung von Nicht-Wörter_ reicht nicht aus

Kontextabhängige Korrektur

Anwendungsgebiet

Durch Tippfehler können auch andere lexikalische Wörter entstehen, die jedoch kontextfremd sind

z.B. Kennst du dir Haustür schließen?

Allgemeine Verfahren zur Korrektur solcher Fehler fehlen noch

Mögliche Verfahren

1. Zusammenfassung von ähnlich geschriebenen oder gesprochenen Wörtern zu einer „Verwechslungsmenge“

   - Beim Auftauchen einer dieser Wörter, werden zugeschnittene Heuristiken angewandt, wodurch mögliche Fehler gefunden werden sollen

   - Verwendet von IBM Critique-System und Microsoft Word

Nachteil:

- Funktioniert nur wenn genau ein Wort der Verwechslungsmenge mit einem anderen dieser Menge vertauscht wurde
- Für jedes Wort müssen eigene Heuristiken entwickelt werden

Mögliche Verfahren

2. N-Gramm-Wahrscheinlichkeit von Wörter

   = die Wahrscheinlichkeit das n benachbarte Wörter zusammen auftraten

   → Hat eine eingegebene Wortgruppe eine niedrigere N-Gramm-Wahrscheinlichkeit als eine vom Programm erzeugte, wird diese als Korrektur vorgeschlagen

Nachteil:

- Analyse von sehr großen Textsammlungen ist notwendig

Grammatikkorrektur

Anwendungsgebiet

Wird angewendet, wenn die Erkennung und Korrektur eines Fehlers nicht nur den lokalen Kontext, sondern die Analyse eines ganzen Satzes oder evtl. Textes voraussetzen.

Anwendungsgebiete - Beispiele

- **Kongruenzfehler:**
  - *Subjekt-Prädikat:*

    Die Erkennung von Nicht-Wörtern reichst nicht aus.
  - *Adjektiv-Substantiv:*

    Die Korrekturprogramme erkennen keine grammatische_ Fehler.
- **Fehlende Wörter:**

  Die Korrekturprogramme _____ **E**rkennen nicht alle Fehler.
- **Falscher Kasus:**

  Die Erkennung vo**m** Nicht-Wörter_ reicht nicht aus.

Verfahren

- Constraint Relaxation
  - Bestimmte Grammatikalitätsbedingungen werden innerhalb der Grammatik nicht als absolut feststehend berachtet
- Fehlerantizipation
  - Fehler werden durch ein Musterabgleich gefunden

→ In der Praxis werden beide Verfahren Kombiniert

Kritik

- Für den alltäglichen Gebrauch zu wenig Präzision (Precision)
  - → maximal 50% der Fehlermeldungen sind richtig
- Auch zu geringe Vollständigkeit (Recall)
  - → Viele Fehler werden nicht gefunden

Fazit

Programme arbeiten noch ungenau

Verbesserungsmöglichkeit:

→ Pflege des Wörterbuches über regelmäßiges Einfügen von neuen Wörtern

Quellen

- K.-U. Carstensen et al. (2004): Computerlinguistik und Sprachtechnologie. Eine Einführung. Spektrum, Akademischer Verlag