TreeTagger
RFTagger

Jiří Vejvoda
12.5.2011

Obsah

- Intro
- TreeTagger
- RFTagger
- Srovnání
- Publikované výsledky
- Dosažené výsledky pro Desam
- Demonstrace
- Zdroje

Intro

- Odstranění nejednoznačnosti přidělení značky určující druh slova (POS tag)
- Rozhodování podle předešlého kontextu
  - Využití pravděpodobnosti – velký objem dat
- Jednoduché × atribute vectors značky
- Jazyková závislost

TreeTagger

- Stuttgart 1994, Helmut Schmid
- Jednoduché značky
  - Spíše menší počet – cca 50 (Penn 36)
- Markovovy Modely
  - Viterbi algorithm
- Rozhodovací stromy
  - Struktura a prořezávání podle informačního zisku
- Suffix Lexicon

TreeTagger

- Třídy ekvivalence
  - Slova se stejnou množinou možných značek
- Prefix Lexicon
- Počáteční slova vět
  - Desambiguace normálních slov od vlastních jmen
- Původní vývoj a testování na angličtině, dále na němčině

RFTagger

- Stuttgart 2008, Schmid, Laws
- Značky formou vektorů atributů
  - Jemné členění slovních druhů
  - Jednotlivé atributy odděleny tečkou, pevný počet
- Skryté Markovovy Modely
- Rozhodovací stromy
  - Dekompozice na jednotlivé atributy
- Lexicon, Wordclass automaton

Srovnání

TreeTagger

- Jednoduché značky
- Rozhodovací stromy pro celou značku

RFTagger

- Attribute vectors
- Rozhodovací stromy pro jednotlivé hodnoty atributů

Publikované výsledky

TreeTagger

- Penn Treebank (2 mil.)
  - 96,81 %
- Stuttgart Zeitung (25 k)
  - 97,53 %

RFTagger

- German Tiger Treebank (886 k)
  - Baseline – 69,4 %
  - 92,2 %
- Czech Academic corpus (652 k)
  - 89,53 %

Desam

TreeTagger

- Baseline – 70,54 %
- Kontext 1 – 86,22 %
- Kontext 2 – 87,31 %
- Kontext 5 – 87,47 %
- Kontext 10 - neuspělo

RFTagger

- Kontext 1 – 90,89 %
- Kontext 2 – 92,06 %
- Kontext 10 – 92,43 %

Zdroje

- Helmut Schmid and Florian Laws: Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, COLING 2008, Manchester, Great Britain.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, pages 44–49, Manchester, UK.
- Schmid, Helmut. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In Proceedings of EACL SIGDAT workshop, Dublin, Ireland.