A Simple Rule-based Part Of Speech Tagger

Vignir Hafsteinsson

Um höfundinn

- Eric Brill
- Skrifaði pappírinn árið 1992 sem doktorsnemi í Universisty of Pennsylvania.
- Er nú hjá Microsoft þar sem að hann er með eigin rannsóknarstofu
  - Text Mining, Search and Navigation Research Group

Yfirlit

- Tölfræðilegar aðferðir hafa verið mun betri en aðferðir sem byggja á reglum við mörkun texta.
- Brill taggerinn byggir á reglum og nær tölfræðilegum mörkurum í nákvæmni með nokkrum aukakostum...

Yfirlit frh.

- Kostir Brill markarans
  - Lítið magn af upplýsingum
  - Gegnsætt
  - Einfalt að finna og útfæra viðbætur
  - Auðvelt að færa yfir í nýjan corpus, eða nýtt tungumál

Tölfræðilegir markarar

- Einfaldir
  - Hendum bara fullt af gögnum í þá.
  - Þurfum ekki einu sinni að vita neitt um mál.
- Hægt að endurbæta þá með því að
  - Framkvæma einhverja forvinnslu og/eða eftirvinnslu
  - Stilla módelið til

Brill markarinn

- Býr til reglurnar sínar sjálfkrafa með notkun corpus texta án samhengis.
- Við mörkun á orði er notað það mark sem er oftast notað í þjálfunargögnunum

- Gefur ekki góðar niðurstöður strax
- Í báðum setningum er *run* markað sem sögn
  - The ***run*** lasted thirty minutes
  - We ***run*** three miles every day
- Einfaldar endurbætur eru þó gerðar á þessu...

Einfaldar endurbætur

- Óþekkt orð sem byrja á stórum staf eru mörkuð sem nafnorð
- Óþekkt orð fá það mark sem er algengast fyrir orð með sömu þriggja stafa endingu
  - Blahblah**ous** væri lýsingarorð
- Þessi algorithmi ásamt endurbótunum nær 92,1% nákvæmni á Brown Corpus(1,1 milljón orð)

Plástrar(e. Patches)

- Markarinn þjálfar sig á 90% af gagnasafninu
- 5% notuð til að búa til plástra(patch corpus)
- 5% notuð til prófunar
- Eftir að búið er að þjálfa í fyrsta skipti er markarinn prófaður á patch corpus og villunum safnað saman
  - $<tag_a, tag_b, number>$

Plástrar frh.

- Síðan eru mismunandi plástrar prófaðir til að minnka villur
- Plástrar eru búnir til úr sniðum(e. templates)
- Fyrir hverja villuþrennu er valinn sá plástur sem minnkar villufjöldann mest.

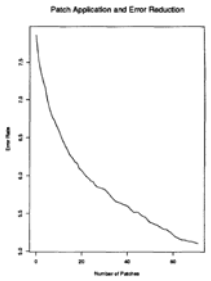
Plástrasnið(e. patch templates)

- Change tag a to tag b when:
  1. The preceding (following) word is tagged ***z***.
  2. The word two before (after) is tagged ***z***.
  3. One of the two preceding (following) words is tagged ***z***.
  4. One of the three preceding (following) words is tagged ***z***.
  5. The preceding word is tagged ***z*** and the following word is tagged ***w***.
  6. The preceding (following) word is tagged ***z*** and the word two before (after) is tagged ***w***.
  7. The current word is (is not) capitalized.
  8. The previous word is (is not) capitalized.

Fyrstu 10 plástrarnir

(1) TO IN NEXT-TAG AT
(2) VBN VBD PREV-WORD-IS-CAP YES
(3) VBD VBN PREV-1-OR-2-OR-3-TAG HVD
(4) VB NN PREV-1-OR-2-TAG AT
(5) NN VB PREV-TAG TO
(6) TO IN NEXT-WORD-IS-CAP YES
(7) NN VB PREV-TAG MD
(8) PPS PPO NEXT-TAG.
(9) VBN VBD PREV-TAG PPS
(10) NP NN CURRENT-WORD-IS-CAP NO

Niðurstöður



- Einn kostur plástrakerfisins er sá að maður getur leyft sér að búa til plástrasnið án þess að eiga hættu á að minnka gæði markarans
- Hvers vegna?

Niðurstöður

- 94,9% nákvæmni á Brown Corpus
- Aðrar tölfræðilegar aðferðir hafa náð(á sama prófunargögnum)
  - 95,5%
  - 96%
  - 96-97%
    - 3% byggðu á handskrifuðum reglum sem byggði á prófunargögnunum.
    - Sumar þessara reglna voru fundnar af markara Brills

Niðurstöður

- Þessi einfaldi markari nær svipaðri nákvæmni og flóknari tölfræðimarkarar en hefur fleiri kosti
  - Markarinn er mjög færanlegur
  - Þarf ekki mikið af gögnum
  - Miklu gegnsærri
  - Einfalt að finna og útfæra viðbætur

Endurbætur

- Some Advances in Transformation-Based Part of Speech Tagging (Brill, 1994)
- Nýjum plástrasniðum bætt við sem innihélt lexical gögn
- Change tag a to tag b when:
  - The preceding (following) word **is** *w*.
  - The word two before (after) **is** *w*.

Some Advances in Transformation-Based Part of Speech Tagging frh.

- Betur tekið á óþekktum orðum
- Ný snið búin til
- Change the tag of an unknown word (from X) to Y if:
  - Deleting the prefix x, |x| <= 4, results in a word (x is any string of length 1 to 4).
  - The first (1,2,3,4) characters of the word are x.
- Change tag:
  - From common noun to plural common noun ifthe word has suffix –s
  - from plural common noun to singular common noun if the word has suffix ss.

Spurningar

?