The Penn Treebank

präsentiert
von

Galja Georgieva

Penn Treebank

- Introduction
- Part – of – Speech Tagging
- Bracketing
- Tagging + Bracketing

Introduction

In diesem Papier besprechen wir die Erfahrung, einen großen mit Anmerkungen versehenen Korpus - The Penn Treebank. Dieser Korpus bestehet aus mehr als 4,5 Millionen Wörter aus den amerikanischen Englisch. Im lange der ersten drei Jahren dieses Projekts (1989-1992) war der Korpus mit part – of – speech(POS) Informationen vermerkt.

Part – of – Speech Tagging

Battle-tested/NNP Japanese/NNP industrial/JJ managers/NNS here/RB always/RB buck/VB up/IN nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.

Part – of – Speech Tagging

- The Brown Corpus
- 87 simple tags
- Francis 1964
- Francis und Kucera 1982

- The Penn Treebank ist basiert auf The Brown Corpus
- reduzieren die Wortklassen und beachten lexikalische und syntaktische Informationen
- 48 simple tags

Part – of – Speech Tagging

- The Brown Corpus
- teilen die drei Wortklassen pre-qualifiers(quite, rather, such), pre-quantifiers(all, half, many, nary) and both.
- haben die zwei Kategorien PPL(singular reflexive pronoun) and PPLS(plural reflexive pronoun).

- The Penn Treebank
- haben dagegen nur die Kategorie PDT(prederminer)
- fast die beiden Kategorien zusammen in PRP(personal pronoun)

The Penn Treebank POS tagset

1. CC Coordinating conjunction
2. CD Cardinal number
3. DT Determiner
4. EX Existential there
5. FW Foreign word
6. IN Preposition/subord. conjunction
7. JJ Adjective
8. JJR Adjective, comparative
9. JJS Adjective, superlative
10. LS List item marker
11. MD Modal
12. NN Noun, singular or mass
13. NNS Noun, plural
14. NNP Proper noun, singular
15. NNPS Proper noun, plural
16. PDT Predeterminer
17. POS Possessive ending
18. PRP Personal pronoun
19. PP$ Possessive pronoun
20. RB Adverb
21. RBR Adverb, comparative
22. RBS Adverb, superlative
23. RP Particle
24. SYM Symbol (mathematical or scienti c)
25. TO to
26. UH Interjection
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund/present participle
30. VBN Verb, past participle
31. VBP Verb, non-3rd ps. sing. present
32. VBZ Verb, 3rd ps. sing. present
33. WDT wh-determiner
34. WP wh-pronoun
35. WP$ Possessive wh-pronoun
36. WRB wh-adverb
37. # Pound sign
38. $ Dollar sign
39. . Sentence- nal punctuation
40. , Comma
41. : Colon, semi-colon
42. (Left bracket character
43.) Right bracket character
44. Straight double quote
45. ` Left open single quote
46. \ Left open double quote
47. ' Right close single quote
48. " Right close double quote

The POS Tagging Prozess

- Automated Stage – die früheren Etappen vom Projekt Penn Treebank, die erste automatische POS, war angeführt von PARTS(Church 1988), ein stochastischer Algirithmus entdeckt in AT&T Bell Labs. PARTS benutzt geänderte Version von The Brown Corpus tagset, ähnlich wie the Penn Treebank tagset und zeigt POS tags mit einer Fehler in Höhe von 3-5%. Das Ergebnis von PARTS war automatisch tokenized und die Wortklassen gezeigt von PARTS waren automatisch abgebildet auf The Penn Treebank tagset. Diese Abbildung zeigt ungefähr 4% Fehler.

Automated Stage

Battle-tested/NNP Japanese/NNP industrial/JJ managers/NNS here/RB always/RB buck/VB up/IN nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.

The POS Tagging Prozess

Manual Correction Stage – Bei dieser Etappe wird das Resultat von den automatischen Etappe zur handliche Korrektur gegeben. Die Überprüfer benutzen eine maus-basierte Ansammlung geschrieben in GNU Emacs Lisp. Diese Ansammlung ist eindebaut in der GNU Emacs Editor (Lewis 1990). Mit dieser Ansammlung kann man die Fehler korrigieren, wobei die Maus auf die falsche Wortklasse positioniert wird und dann kann mann der Fehler gleich korrigieren.

Beispiel: vor und nach der Korrektur

Battle-tested/NNP Japanese/NNP industrial/JJ managers/NNS here/RB always/RB buck/VB up/IN nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.

Battle-tested/NNP*/JJ Japanese/NNP*/JJ industrial/JJ managers/NNS here/RB always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.

Das Experiment

Um zu entscheiden, wie das ganze Prozess schneller laufen kann, haben die Leute von Penn Treebank ein Experiment durchgeführt. Es geht um Vergleich von zwei Arten der Notierung:

- „tagging“ - wo der Text nur mit Hand notiert wird
- „correcting“- hier benutzen sie das Ergebnis von PARTS und korrigieren es danach.

Das Experiment zeigt, dass bei „tagging“ man doppelt mehr Zeit braucht, als bei „correcting“.

Bracketing

Die Methodologie der Ausklammerung läuft absolut paralell mit der Bestimmung der Wortklassen

Beispiel

```
( (S
   (NP Battle-tested industrial managers
      here)
    always
   (VP buck
      up
      (NP nervous newcomers)
      (PP with
         (NP the tale
            (PP of
               (NP (NP the
                     (ADJP first
                        (PP of
                           (NP their countrymen)))
                     (S (NP *)
                        to
                        (VP visit
                           (NP Mexico))))
                  ,
                  (NP (NP a boatload
                        (PP of
                           (NP (NP warriors)
                              (VP-1 blown
                                 ashore
                                 (ADVP (NP 375 years)
                                    ago)))))
                     (VP-1 *pseudo-attach*)))))))
    .)
```

Bracketing

Der Parser **Fidditch** ist entfaltet von Donald Hindle erstmal in der Universität in Pennsylvania und ausschließend in AT&T Bell Labs(Hindle 1983, 1989).

Fidditch hat drei Eigenschaften, die ihn ideal machen als Vorbereiter für Hand-Korrektur:

- Fiddich liefert immer genau eine Analyse für jeden gegeben Satz, so dass die Notierer nicht durch mehrere Analyse aussuchen sollen.
- Fidditch bestimmt keinen Komponent, dessen Rolle nicht mit Sicherheit deffinieren kann. Im Fall von Unsicherheit Fidditch zerkleinert die Eingabe in eine Folge von Bäume, bietet nur eine Teilstruktur für jeden Satz.
- Fidditch hat ziemlich gute grammatische Umfang, so dass das, was er zeigt normalerweise korrekt ist.

The Penn Treebank syntactic tagset

1. ADJP Adjective phrase
2. ADVP Adverb phrase
3. NP Noun phrase
4. PP Prepositional phrase
5. S Simple declarative clause
6. SBAR Clause introduced by subordinating conjunction or 0 (see below)
7. SBARQ Direct question introduced by wh-word or wh-phrase
8. SINV Declarative sentence with subject-aux inversion
9. SQ Subconstituent of SBARQ excluding wh-word or wh-phrase
10. VP Verb phrase
11. WHADVP W h-adverb phrase
12. WHNP W h-noun phrase
13. WHPP W h-prepositional phrase
14. X Constituent of unknown or uncertain category

Null elements

1. * \Understood" subject of in nitive or imperative
2. 0 Zero variant of that in subordinate clauses
3. T Trace|marks position where moved wh-constituent is interpreted
4. NIL Marks position where preposition is interpreted in pied-piping contexts

Das Ergebnis von Fidditch

```
( (S
   (NP (NBAR (ADJP (ADJ "Battle-
tested/JJ")
                   (ADJ "industrial/JJ"))
             (NPL "managers/NNS")))
   (? (ADV "here/RB"))
   (? (ADV "always/RB"))
   (AUX (TNS *))
   (VP (VPRES "buck/VBP"))
   (? (PP (PREP "up/RP")
         (NP (NBAR (ADJ "nervous/JJ")
                   (NPL "newcomers/NNS")))))
   (? (PP (PREP "with/IN")
         (NP (DART "the/DT")
            (NBAR (N "tale/NN"))
                  (PP of/PREP
                     (NP (DART "the/DT")
                        (NBAR (ADJP
                              (ADJ "first/JJ")))))))))
(? (PP of/PREP
      (NP (PROS "their/PP$")
         (NBAR (NPL
"countrymen/NNS")))))
(? (S (NP (PRO *))
         (AUX to/TNS)
         (VP (V "visit/VB")
            (NP (PNP "Mexico/NNP")))))
(? (MID ",/,"))
(? (NP (IART "a/DT")
      (NBAR (N "boatload/NN"))
            (PP of/PREP
               (NP (NBAR
                  (NPL "warriors/NNS"))))
            (VP (VPPRT "blown/VBN")
               (? (ADV "ashore/RB"))
               (NP (NBAR (CARD "375/CD")
                         (NPL "years/NNS"))))))
(? (ADV "ago/RB"))
(? (FIN "./.")))
```

Bracketing

Das originale Design der Baumstruktur für syntaktische Analyse ist vergleichbar mit der skeletische Analyse benutzt von Lancaster Treebank, aber ein Experiment war früher in den Projekt vorgeführt, um aufzuforschen, ob es machbar ist, eine größere Niveau der strukturellen Einzelheiten zu beschaffen. Da aber die Ergebnisse nicht ganz klar waren, war da ein Beweis, dass die Notierer behaupten können, eine viel schnellere handliche Korrektur, wenn das Ergebnis des Parsers in verschiedenen Weisen vereinfacht wird. Die wichtigsten Ergebnisse dieses Experiment zeigen dass:

- Die Notierer brauchen längere Zeit Bracketing zu lernen als POS tagging
- Die Notierer Können das Ergebnis von Fidditch mit eine Schnelligkeit von 375 Wörter pro Stunde korrigieren. Das passiert nach 3 Wochen tainieren und nach 6 Wochen überprüft man schon 475 Wörter per Stunde.

Beispiel: nach der Vereinfachung

```
( (S
    (NP (ADJP Battle-tested
industrial)
        managers)
    (? here)
    (? always)
    (VP buck))
    (? (PP up
        (NP nervous
newcomers)))
    (? (PP with
        (NP the tale
            (PP of
                (NP the
                    (ADJP
first))))))
```

```
 (? (PP of
      (    (NP their
countrymen)))
    (? (S (NP *)
          to
         (VP visit
            (NP Mexico))))
    (? ,)
    (? (NP a boatload
         (PP of
             (NP warriors))
         (VP blown
             (? ashore)
             (NP 375 years))))
    (? ago)
    (? .))
```

Bracketing: nach der Korrektur

```
( (S
  (NP Battle-tested industrial
managers
      here)
   always
  (VP buck
      up
      (NP nervous newcomers)
      (PP with
         (NP the tale
            (PP of
               (NP (NP the
                     (ADJP first
                        (PP of
                           (NP their
countrymen)))

 .)
```

Bracketing: nach der Korrektur

```
(S (NP *)
                to
                (VP visit
                   (NP Mexico))))
             ,
             (NP (NP a boatload
                  (PP of
                     (NP (NP warriors)
                        (VP-1 blown
                           ashore
                           (ADVP (NP 375
years)
                                    ago)))))
                (VP-1 *pseudo-attach*)))))))
```

Tagging + Bracketing

```
(S
  (NP-SBJ
    (NP (JJ Battle-tested) (JJ Japanese) (JJ industrial) (NNS managers) )
    (ADVP-LOC (RB here) ))
  (ADVP-TMP (RB always) )
  (VP (VBP buck)
    (PRT (RP up) )
    (NP (JJ nervous) (NNS newcomers) )
    (PP-CLR (IN with)
      (NP
        (NP (DT the) (NN tale) )
        (PP (IN of)
          (NP
            (NP
              (NP (DT the) (JJ first) )
              (PP (IN of)
                (NP (PRP$ their) (NNS countrymen) ))
              (SBAR
                (WHNP-1 (-NONE- 0) )
                (S
                  (NP-SBJ (-NONE- *T*-1) )
                  (VP (TO to)
                    (VP (VB visit)
                      (NP (NNP Mexico) ))))))
            (, ,)
            (NP
              (NP (DT a) (NN boatload) )
              (PP (IN of)
                (NP (FW samurai) (NNS warriors) ))
              (VP (VBN blown)
                (NP (-NONE- *) )
                (ADVP-CLR (RB ashore) )
                (NP-TMP (CD 375) (NNS years) (RB ago) ))))))))
  (. .) ))
```