ΜΕΤΑΣΧΗΜΑΤΙΣΤΙΚΟΙ ΚΑΝΟΝΕΣ

Ο Brill αντιμετωπίζει το πρόβλημα της μορφοσυντακτικής ανάλυσης, όπως και πολλά άλλα προβλήματα γλωσσικής επεξεργασίας, ως μια ακολουθία μετασχηματισμών από λιγότερο ορθές αναλύσεις σε περισσότερο ορθές αναλύσεις (transformation-based approach).

Αρχικά, η κάθε λέξη λαμβάνει από το λεξικό τη μορφοσυντακτική ετικέτα με την οποία εμφανίζεται πιο συχνά μέσα σε κείμενα. Οι λέξεις που δεν περιέχονται στο λεξικό λαμβάνουν την ετικέτα "ουσιαστικό" (η πιο πιθανή κατηγορία για μια άγνωστη λέξη). Στη συνέχεια, εφαρμόζεται μια ακολουθία κανόνων, στόχος των οποίων είναι να διορθώσουν ετικέτες λέξεων που δε συμβιβάζονται με τα συμφραζόμενα. Τυπικό παράδειγμα κανόνα είναι: «ΑΛΛΑΞΕ την ετικέτα της τρέχουσας λέξης ΑΠΟ "ρήμα" ΣΕ "ουσιαστικό" ΑΝ η προηγούμενη λέξη έχει την ετικέτα "άρθρο"».

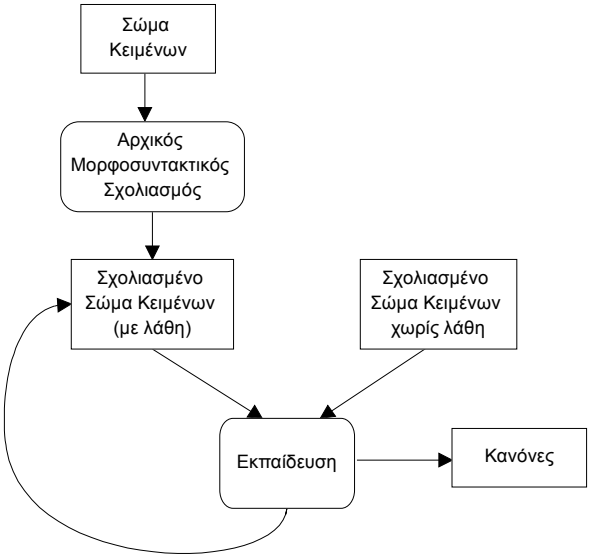
Οι μετασχηματιστικοί κανόνες (τους οποίους ο Brill ονομάζει *patches – μπαλώματα*) κάνουν όλοι την ίδια απλή λειτουργία: αλλάζουν τη μορφοσυντακτική ετικέτα της τρέχουσας λέξης με κάποια άλλη, όταν τα συμφραζόμενα ικανοποιούν συγκεκριμένες συνθήκες. Η δομή των κανόνων καθορίζεται από τα παρακάτω εκμαγεία (templates):

ΑΛΛΑΞΕ την ετικέτα της τρέχουσας λέξης ΑΠΟ ***a*** ΣΕ ***b*** ΑΝ:

1. Η προηγούμενη/επόμενη λέξη έχει την ετικέτα ***t***
2. Η προ-προηγούμενη/μεθεπόμενη λέξη έχει την ετικέτα ***t***
3. Μία από τις δύο προηγούμενες/επόμενες λέξεις έχει την ετικέτα ***t***
4. Μία από τις τρεις προηγούμενες/επόμενες λέξεις έχει την ετικέτα ***t***
5. Η προηγούμενη λέξη έχει την ετικέτα t_1 και η επόμενη την ετικέτα t_2
6. Η προηγούμενη/επόμενη λέξη έχει την ετικέτα t_1 και η προ-προηγούμενη την ετικέτα t_2
7. Η προηγούμενη/επόμενη λέξη είναι η ***w***
8. Η προ-προηγούμενη/μεθεπόμενη λέξη είναι η ***w***
9. Μία από τις δύο προηγούμενες/επόμενες λέξεις είναι η ***w***
10. Η τρέχουσα λέξη είναι η w_1 και η προηγούμενη/επόμενη λέξη είναι η w_2
11. Η τρέχουσα λέξη είναι η ***w*** και η προηγούμενη/επόμενη λέξη έχει την ετικέτα ***t***
12. Η τρέχουσα λέξη είναι η ***w***
13. Η προηγούμενη/επόμενη λέξη είναι η ***w*** και η προηγούμενη/επόμενη λέξη έχει την ετικέτα ***t***
14. Η τρέχουσα λέξη είναι η w_1, η προηγούμενη/επόμενη λέξη είναι η w_2 και η προηγούμενη/επόμενη λέξη έχει την ετικέτα ***t***

Τα εκμαγεία 1-6 αναφέρονται σε μορφοσυντακτικές ετικέτες από τα συμφραζόμενα (κανόνες συμφραζομένων – contextual rules) ενώ τα εκμαγεία 7-14 αναφέρονται σε συγκεκριμένες λέξεις (λεξικοί κανόνες – lexical rules). Κάθε μετασχηματιστικός κανόνας παράγεται από κάποιο εκμαγείο, δίνοντας συγκεκριμένες τιμές στις μεταβλητές ***a***, ***b***, ***t*** και ***w***. Για παράδειγμα, ο κανόνας «ΑΛΛΑΞΕ την ετικέτα της τρέχουσας λέξης ΑΠΟ "ρήμα" ΣΕ "ουσιαστικό" ΑΝ η προηγούμενη λέξη έχει την ετικέτα "άρθρο"» έχει παραχθεί από το εκμαγείο 1, θέτοντας ***a*** = "ρήμα", ***b*** = "ουσιαστικό" και ***t*** = "άρθρο".

Η διαδικασία για τη μηχανική μάθηση των κανόνων (την οποία ο Brill ονομάζει transformation-based error-driven learning) παρουσιάζεται στο Σχήμα 1:



Σχήμα 1. Transformation-based error-driven learning

Δεδομένου ενός σώματος κειμένων με μορφοσυντακτικό σχολιασμό και ενός συνόλου εκμαγείων, η μηχανική μάθηση των μετασχηματιστικών κανόνων γίνεται ως εξής:

1. Το σχολιασμένο σώμα κειμένων χωρίζεται σε δύο μέρη, το σώμα ***C1*** και το σώμα ***C2*** (τυπική αναλογία 9:1).
2. Από το σώμα ***C1*** εξάγονται όλες οι λέξεις μαζί με την πιο συχνή μορφοσυντακτική τους ετικέτα. Έτσι συγκροτείται το λεξικό του μορφοσυντακτικού σχολιαστή.
3. Δημιουργείται ένα αντίγραφο του σώματος ***C2***, το ***C2a***. Το λεξικό χρησιμοποιείται για να χαρακτηρίσει εκ νέου τις λέξεις του σώματος ***C2a***. Όσες λέξεις δεν υπάρχουν στο λεξικό λαμβάνουν την ετικέτα "ουσιαστικό". Έτσι προκύπτει ένας πρώτος αυτόματος μορφοσυντακτικός σχολιασμός του σώματος ***C2a***.
4. Συγκρίνονται οι ετικέτες που αποδόθηκαν στις λέξεις του σώματος ***C2a*** με αυτές που έχουν οι ίδιες λέξεις στο σώμα ***C2*** και δημιουργείται μια λίστα με λάθη. Το κάθε λάθος είναι μια τριάδα <***a***, ***b***, ***n***>, η οποία δηλώνει ότι ***n*** λέξεις έλαβαν την ετικέτα ***a***, ενώ θα έπρεπε να λάβουν την ετικέτα ***b***.
5. Για το κάθε λάθος $e_i = <a_i, b_i, n_i>$ επιλέγονται διαδοχικά ένα-ένα τα εκμαγεία. Το κάθε εκμαγείο παράγει μια σειρά μετασχηματιστικών κανόνων, π.χ. το εκμαγείο «ΑΛΛΑΞΕ την ετικέτα της τρέχουσας λέξης ΑΠΟ a_i ΣΕ b_i ΑΝ η προηγούμενη/επόμενη λέξη έχει την ετικέτα ***t***» παράγει τόσους κανόνες όσες τιμές πάρει η μεταβλητή ***t***. Καθένας από αυτούς τους κανόνες εφαρμόζεται στο ***C2a*** και βαθμολογείται σύμφωνα με τη μείωση (ή αύξηση) που προκαλεί στο λάθος e_i. Αν ***R*** είναι ένας κανόνας, τότε $\textbf{βαθμός}_R = n_i - (\text{πλήθος σωστών αλλαγών}_R - \text{πλήθος λανθασμένων αλλαγών}_R)$. Σωστή αλλαγή_R είναι μια αλλαγή που έκανε ο ***R*** σε ετικέτα λέξης του ***C2a*** από a_i σε b_i και η ίδια λέξη έχει στο ***C2*** ετικέτα b_i. Λανθασμένη αλλαγή_R είναι μια αλλαγή που έκανε ο ***R*** σε ετικέτα λέξης του ***C2a*** από a_i σε b_i και η ίδια λέξη δεν έχει στο ***C2*** ετικέτα b_i. Υπολογίζεται η βαθμολογία όλων των κανόνων που παράγονται από όλα τα

εκμαγεία για το λάθος e_i και ο κανόνας με την υψηλότερη βαθμολογία προστίθεται σε μια λίστα κανόνων. Έτσι προκύπτουν οι καλύτεροι κανόνες για όλα τα λάθη e_i, οι οποίοι εφαρμόζονται στο ***C2a*** και η διαδικασία επαναλαμβάνεται από το βήμα 4, μέχρις ότου νέοι κανόνες να μην ελαττώνουν πλέον το συνολικό πλήθος των λαθών.

Με την παραπάνω διαδικασία κατασκευάζεται αυτόματα ένας μορφοσυντακτικός σχολιαστής, ο οποίος αποτελείται από ένα λεξικό για τον αρχικό σχολιασμό οποιουδήποτε άγνωστου κειμένου και μια διατεταγμένη λίστα από μετασχηματιστικούς κανόνες, η διαδοχική εφαρμογή των οποίων διορθώνει τα λάθη του αρχικού σχολιασμού. Το συγκεκριμένο μοντέλο φτάνει για την Αγγλική σε ποσοστό ακρίβειας >95%, ακρίβεια