Προπαρασκευαστική 1ης Εργαστηριακής Άσκησης Επεξεργασίας Φωνής και Φυσικής Γλώσσας

Γιώργος Ξυπόλιτος 03113629

5-11-2017

Σκοπός της Άσκησης

Σκοπός της προπαρασκευαστικής είναι η κατασκευή ενός μετατροπέα greeklish σε greek μέσω της χρήσης training samples για την αντιστοίχιση των λατινικών χαρακτήρων σε αντίστοιχους χαρακτήρες ελληνικών. Για τον σκοπό αυτό δημιουργούμε έναν transducer.
Επίσης, δημιουργήθηκαν και acceptors ελληνικών και αγγλικών λέξεων, μέσω αντίστοιχων λεξιλόγιων.
Για την υλοποίηση των βημάτων χρησιμοποιήθηκε η γλώσσα σεναρίων Python και η βιβλιοθήκη OpenFst1.6.5 για το shell.

Βήμα 1, 2

Αρχικά διαβάζουμε τα δεδομένα από τα αρχεία *train_gr.txt* και *train_greng.txt* και τα χωρίζουμε σε λέξεις, οι οποίες αντιστοιχούν 1-1.
Έπειτα, αφαιρούμε τις λέξεις που είναι ακριβώς ίδιες μεταξύ των δύο κειμένω, αφού αποτελούν αγγλικές λέξεις, δηλαδή δεν χρειάζονται μετατροπή.
Για την εύρεση των κανόνων ελέγχουμε μία λέξη από κάθε κείμενο την φορά. Το μήκος των λέξεων που αντιστοιχούν μεταξύ τους δεν είναι απαραίτητα ίσο, επομένως απαιτούνται κάποιοι ειδικοί κανόνες μετατροπής των λέξεων που εμφανίζονται στα λεξικά sp_ch_gr και sp_ch_en. Αυτοί οι κανόνες βρέθηκαν εμπειρικά, παρατηρώντας τα δεδομένα.
Δεδομένου ότι ένα μέρος των κανόνων βρέθηκαν εμπειρικά, είναι πιθανό να υπολείπονται κάποιοι και κάποιοι να μην είναι πλήρεις ή σωστοί. Αυτό βέβαια διορθώνεται, διότι η πιθανότητα εμφάνισης των λανθασμένων κανόνων είναι μικροί, επομένως θα έχουν μεγάλο βάρος στο τελικό fst και δεν θα επιλέγονται συχνά.
Το αρχείο *convertions.txt* δείχνει όλους τους κανόνες μετατροπής με την αντίστοιχη συχνότητα εμφάνισής τους και την κανονικοποιημένη, ως προς το συνολικό αριθμό εφαρμογής των κανονών, συχνότητά τους, καθώς και τα αντίστοιχα κόστη τους(αρνητικός λογάριθμος της πιθανότητας εμφάνισής τους). Στο αρχείο *rules.txt* περιέχονται όλες οι αντιστοιχίσεις κανόνων που έγιναν για κάθε γράμμα.

Βήματα 3, 4

Με την εκτέλεση των σεναρίων scripts $fstG.py$ και $fstI.py$ δημιουργούνται τα FSTs G και I. Δηλαδή γράφουμε σε ένα αρχείο τους κανόνες που βρήκαμε προηγουμένως στην μορφή που απαιτεί η βιβλιοθήκη OpenFst. Το FST I χρειάζεται για να είναι δυνατή η διατήρηση των αγγλικών λέξεων, σε περίπτωση ύπαρξής τους στο κείμενο που θα εισάγουμε στο FST.

Βήματα 5, 6, 7

Για την δημιουργία των FSA A1 A2, τρέχουμε το script *A.py* με τα κατάλληλα ορίσματα όπως αναφέρεται στο *README* και διαβάζουμε τις λέξεις από το αντίστοιχο λεξικό, είτε αγγλικό για το A1 είτε ελληνικό για το A2, φτιάχνοντας για κάθε μια, μία αλληλουχία καταστάσεων που την αποδέχεται ξεκινόντας από μία αρχική κατάσταση, κοινή για όλες τις λέξεις, και με ϵ μετάβαση μεταφερόμαστε

στην αρχή κάθε αλληλουχίας. Αφού διαβαστεί μία λέξη που υπάρχει στο λεξικό με ϵ μετάβαση μεταφερόμαστε στην αποδεκτή κατάσταση. Αυτό επαναλαμβάνεται για όλες τις λέξεις στο λεξικό, ώσπου τελικά έχουμε ένα μη ντετερμινιστικό αποδοχέα λέξεων.
Για την μετατροπή τους σε μη ντετερμινιστικά και έπειτα ελάχιστα, χρησιμοποιούνται οι εντολές της OpenFst fstrmepsilon, fstdeterminize, fstminimize με την σειρά που αναφέρονται.

Παρατηρήσεις

Λόγω μεγάλου όγκου δεδομένων, έγινε γραφική απεικόνιση των αποδοχέων λέξεων, για ένα δείγμα 80 λέξεων από κάθε λεξικό.
Όλα τα δεδομένα που παράγονται από τα scripts για τα εκάστοστε FS βρίσκονται στους αντίστοιχους φακέλους.