1η Εργαστηριακή Άσκηση
Επεξεργασίας Φωνής και Φυσικής Γλώσσας

Γιώργος Ξυπόλιτος 03113629

20-11-2017

Σκοπός της Άσκησης

Σκοπός του υπόλοιπου μέρους της εργαστηριακής άσκησης είναι η χρήση των transducers και acceptors που κατασκευάστηκαν στο προπαρασκευαστικό μέρος, για την κατασκευή ενός μετατροπέα greeklish σε greek και η επέκτασή του με την χρήση ενός ορθογράφου για την διόρθωση τυχόν λαθών.

Βήματα 8, 9, 10, 11

Αρχικά ενώνουμε τους δύο acceptors των λεξικών, για την δημιουργία ενός acceptor που δέχεται τις λέξεις που υπάρχουν στα αντίστοιχα λεξικά. Ενώνουμε τον αποδοχέα με την κλειστότητα της ένωσης των transducers G και I, ώστε να δημιουργηθεί ένας transducer T ο οποίος μετατρέπει λέξεις σε greeklish σε ελληνικές, ενώ αφήνει αμετάβλητες τις αγγλικές.

Έπειτα, για κάθε λέξη που βρίσκεται στο αρχείο text_greng.txt δημιουργούμε έναν αποδοχέα της και τον συνθέτουμε με τον transducer T και βρίσκουμε το καλύτερο μονοπάτι πάνω του. Έτσι βρίσκουμε την πιο πιθανή μετατροπή σε ελληνικά της εκάστοτε λέξης, την οποία και συγκρίνουμε με την αντίστοιχη της στο αρχείο test_gr.txt.

Τα αποτελέσματα του παραπάνω βήματος βρίσκονται στον φάκελο final_output στο αρχείο out_v1.txt και η ακρίβεια του συστήματος ήταν η εξής:

Conversions	*Accuracy*
Unmatched	0.99%
English to Greek	0.00%
False	3.37%
Correct	**95.64%**

Βήμα 12

Παίρνοντας τα δεδομένα που δίνονται στα αρχεία train_wr.txt, train_co.txt και συγκρίνοντάς μέσω του script C.py υπολογίζονται τα πλήθη λαθών εισαγωγής, διαγραφής, αντικατάστασης και αναδιάταξης χαρακτήρων. Μέσω αυτών, υπολογίζονται οι πιθανότητες και τα αντίστοιχα βάρη που θα τοποθετηθούν στα επόμενα transducers.

Ο τρόπος σύγκρισης των λαθών των δύο κειμένων φαίνεται στο αρχείο convertions.txt του φακέλου final_output. Για τα λάθη ισχύει ότι όταν γράμματα της λανθασμένης και της σωστής λέξης διαφέρουν, τότε υπάρχει λάθος. Άμα δύο γράμματα της λανθασμένης λέξης αντιστοιχούν σε ένα γράμμα της σωστής λέξης, ενώ και τα δύο γράμμα της λανθασμένης λέξης διαφέρουν από το αντίστοιχο της σωστής, τότε θεωρείται λάθος διαγραφής και αντικατάστασης, αλλιώς απλώς αντικατάστασης. Ανάποδα, όταν ένα γράμμα της λανθασμένης λέξης αντιστοιχεί σε δύο γράμματα της σωστής, ενώ και τα δύο γράμματα της σωστής διαφέρουν από της λανθασμένης, τότε θεωρείται λάθος εισαγωγής και αντικατάστασης, αλλιώς απλώς αντικατάστασης. Αν απλώς ένα γράμμα λείπει από την λανθασμένη,

τότε θεωρείται λάθος διαγραφής. Ομοίως, αν υπάρχει παραπάνω γράμμα στην λανθασμένη λέξη, τότε θεωρείται λάθος εισαγωγής. Επίσης, άμα δύο γράμματα της λανθασμένης είναι όμοια εναλλάξ με τα αντίστοιχα δύο γράμματα της σωστής, τότε θεωρείται λάθος αναδιάταξης. Τέλος, αν μόνο ένα γράμμα διαφέρει και δεν ισχύει κάποιος από τους παραπάνω κανόνες, τότε θεωρείται λάθος αντικατάστασης, αλλιώς δεν υπάρχει λάθος.

Επειδή οι κανόνες που δημιουργούνται δεν υπάρχουν για κάθε σύμβολο των transducers, θεωρήθηκε ένα βάρος για κάθε μία από τις παραπάνω κατηγορίες λαθών. Οι πιθανότητες και τα αντίστοιχα βάρη ήταν:

Error Type	*Probability*	*Negative Log (Weight)*
Insertion	0.05%	2.9281
Deletion	0.03%	3.4136
Substitution	0.11%	2.1880
Rearrangement	0.01%	4.4815

Για τον μετατροπέα I, δημιουργήθηκαν δύο καταστάσεις, μία αρχική και μια τελική και συνδέθηκαν με τόξα, ένα για κάθε γράμμα της αγγλικής και της ελληνικής αλφαβήτου, που ξεκινούσαν από την αρχική και κατέληγαν στην τελική, όπου ως ετικέτες είχαν τα γράμματα και μηδενικά βάρη.

Για τον μετατροπέα E, δημιουργήθηκαν τα ίδια τόξα με τον I και επιπλέον δημιουργήθηκαν ετικέτες και καταστάσεις για όλους τους τύπους λαθών. Συγκεκριμένα, για τα λάθη εισαγωγης δημιουργήθηκαν τόξα μεταξύ της αρχικής και της τελικής κατάστασης, με ετικέτες όπου μετέτρεπαν την έψιλον κατάσταση σε όλα τα γράμματα της αλφαβήτου. Για τα λάθη διαγραφής δημιουργήθηκαν τόξα μεταξύ της αρχικής και της τελικής κατάστασης, με ετικέτες που μετέτρεπαν όλα τα γράμμματα της αλφαβήτου σε έψιλον καταστάσης. Για τα λάθης αντικατάστασης, δημιουργήθηκαν τόξα μεταξύ της αρχικής και της τελικής κατάστασης, με ετικέτες που μετέτρεπαν όλα τα γράμματα της αλφαβήτου σε όλα τα υπόλοιπα γράμματα. Τέλος, για τα λάθη αναδιάταξης χρειάστηκε να δημιουργηθούν επιπλέον βοηθητικές καταστάσεις, μία για κάθε αναδιάταξη γραμμάτων. Δηλαδή, υπήρχε μετάβαση από την αρχική κατάσταση στην εκάστοτε βοηθητική με ετικέτα που μετέτρεπε ένα γράμμα σε ένα άλλο(διαφορετικά μεταξύ τους) και μία μετάβαση από την βοηθητική στην τελική όπου μετέτρεπε το γράμμα που είχαμε ως έξοδο στο γράμμα που είχαμε ως είσοδο στην προηγούμενη μετάβαση που μας έφερε στην βοηθητική.

Για την δημιουργία του S1 δημιουργήθηκε το closure του I, το οποίο έγινε concatenated με το E και έπειτα με τον εαυτό του πάλι.

Για την δημιουργία του S2 έγινε concatenation του S1 με το E και έπειτα με το closure του I.

Ο μετατροπέας L αντιστοιχεί στην ένωση των λεξικών A1, A2 που δημιουργήθηκαν στο προπαρασκευαστικό μέρος της εργασίας.

Η σύγκριση του test_co.txt με το test_wr.txt έγινε όμοια με το βήμα 11, με την διαφορά ότι αντί του T γινόταν σύνθεση του W με το S1 ή το S2 και έπειτα με το L.

Τα αποτελέσματα για το S1 και το S2 βρίσκονται στα αρχεία out_v21.txt, out_v22.txt του φακέλου final_output και τα αντίστοιχα στατιστικά είναι:

Results S1

Conversions	*Accuracy*
Unmatched	3.88%
False	10.19%
Correct	**85.92%**

Results S2

Conversions	*Accuracy*
Unmatched	0.49%
False	16.02%
Correct	**83.50%**

Με την χρήση των S1, S2 παρατηρείται ότι παρά την δυνατότητα διόρθωσης μιας λέξεις, ακόμη υπάρχουν μη αναγνωρισμένες λέξεις και καθώς δεν συνυπολογίζονται τα συμφραζόμενα, το ποσοστό επιτυχίας είναι σχετικά μικρό.

Για το τελευταίο βήμα πήραμε την σύνθεση του αποδοχέα της εκάστοτε λέξης με το closure της ένωσης του G και του I, έπειτα το συνθέσαμε με είτε το S1 είτε το S2 και έπειτα με το L και παίρνουμε το ελάχιστο μονοπάτι. Έτσι πήραμε την καλύτερη διορθωμένη μετατροπή μιας λέξης από greeklish σε ελληνικά ή αγγλικά.

Τα αποτελέσματα για το S1 και το S2 βρίσκονται στα αρχεία out_v31.txt, out_v32.txt του φακέλου final_output και τα αντίστοιχα στατιστικά είναι:

Results S1

Conversions	*Accuracy*
Unmatched	0.00%
English to Greek	1.78%
False	7.52%
Correct	**90.69%**

Results S2

Conversions	*Accuracy*
Unmatched	0.00%
English to Greek	2.18%
False	9.90%
Correct	**87.92%**

Όπως φαίνεται από τα αποτελέσματα η προσθήκη διορθωτή μηδένισε τις λέξεις που δεν αναγνωρίζονταν λόγω ορθογραφικών λαθών, αλλά μείωσε το ποσοστό των σωστών μετατροπών. Αυτό οφείλεται αρχικά στην ανακρίβεια του μοντέλου μας, καθώς είναι ισοπίθανοι όλοι οι συνδυασμοί που αφορούν ένα τύπο λάθους, κάτι που φυσικά δεν ανταποκρίνεται στην πραγματικότητα. Δεύτερον, όπως αναφέρθηκε και προηγουμένως, δεν επηρεάζεται η μετατροπή από τα συμφραζόμενα επομένως η λέξη EYXARISTW μπορεί να μετατραπεί στην λέξη ΕΥΧΑΡΙΣΤΟ αντί για ΕΥΧΑΡΙΣΤΩ καθώς έχει μικρότερο βάρος, χωρίς να αναγνωρίσει ο μετατροπέας την σημασιολογική της χρήση. Τρίτον, το κείμενο δοκιμής δεν περιέχει ορθογραφικά σφάλματα και τέλος οι αγγλικές λέξεις λόγω του μεγάλου βάρους τους δεν επιλέγονται από τον μετατροπέα, αντιθέτως επιλέγει να τις μετατρέψει σε μια κοντινή ελληνική με μικρότερο βάρος.