Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής & Υπολογιστών

Μηχανική Μετάφραση της Αγγλικής Γλώσσας στην Ελληνική με χρήση Βαθιών Νευρωνικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΥΡΙΑΚΗ Γ. ΠΕΤΡΟΥ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Νοέμβριος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας, Πληροφορικής & Υπολογιστών

Μηχανική Μετάφραση της Αγγλικής Γλώσσας στην Ελληνική με χρήση Βαθιών Νευρωνικών Δικτύων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΥΡΙΑΚΗ Γ. ΠΕΤΡΟΥ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 24η Νοεμβρίου 2020.

Ανδρέας-Γεώργιος Σταφυλοπάτης	Στέφανος Κόλλιας	Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.

Αθήνα, Νοέμβριος 2020

Κυριακή Γ. Πέτρου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η αυτοματοποιημένη μηχανική μετάφραση της αγγλικής γλώσσας στην ελληνική, με χρήση μοντέλων βαθιάς μηχανικής μάθησης. Το πρόβλημα της μηχανικής μετάφρασης ανήκει στον τομέα της υπολογιστικής γλωσσολογίας και επικεντρώνεται στη μετάφραση κειμένου ή λόγου από μια γλώσσα σε μια άλλη αξιοποιώντας την τεχνητή νοημοσύνη, τα μαθηματικά, τη λογική, τη γλωσσολογία και άλλες επιστήμες, ώστε να αυτοματοποιήσει τη διαδικασία με όσο το δυνατόν καλύτερο αποτέλεσμα.

Στην εργασία αυτή θα εξετάσουμε αρχιτεκτονικές νευρωνικών δικτύων, οι οποίες δίνουν πολύ ενθαρρυντικά αποτελέσματα σε σύγκριση με τις κλασικές μεθόδους της στατιστικής μηχανικής μετάφρασης. Οι μεθοδολογίες που αναπτύσσονται αφορούν μοντέλα κωδικοποίησης της πηγαίας γλώσσας και αποκωδικοποίησης της στη γλώσσα-στόχο. Πιο συγκεκριμένα, θα χρησιμοποιήσουμε ακολουθιακά μοντέλα τα οποία περιλαμβάνουν δίκτυα μακράς βραχυπρόθεσμης μνήμης για την κωδικοποίηση και αποκωδικοποίηση των δεδομένων. Θα πειραματιστούμε με τη χρήση αμφίδρομων και μη σχετικών δικτύων στον κωδικοποιητή και στη συνέχεια θα επιχειρήσουμε να βελτιώσουμε τη διαδικασία της κωδικοποίησης προσθέτοντας ένα μηχανισμό προσοχής. Ακόμα, θα προσπαθήσουμε να αντιμετωπίσουμε προβλήματα που ανακύπτουν κατά τη μετάφραση με χρήση του αλγορίθμου ακτινικής αναζήτησης. Τέλος, παραθέτουμε αδυναμίες στην αντιστοίχηση των δύο γλωσσών καθώς και περιορισμούς λόγω έλλειψης επαρκούς ποσότητας δεδομένων.

Λέξεις κλειδιά

Μηχανική Μετάφραση, Επεξεργασία Φυσικής Γλώσσας, Ακολουθιακά Δίκτυα, Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης, Επίπεδα προσοχής, Βαθιά Μηχανική Μάθηση, Αλγόριθμος Ακτινικής Αναζήτησης

Abstract

The objective of this dissertation is to study the problem of neural machine translation from English to Greek using deep learning techniques. Machine translation is a sub-field of computational linguistics and natural language processing which enables an automatic translation of sentences or speech from one language to another by leveraging artificial intelligence, mathematics, logic linguistics and other scientific areas, in order to automate the procedure, obtaining the best possible outcome.

In the current dissertation, we are going to examine deep neural architectures that have recently achieved promising results, compared to statistical machine translation models. The methodologies outlined in this work are related to models encoding the source language and decoding it to the target language. More specifically, we are going to use sequential models that include long short-term memory networks for the encoding and decoding of data. We are going to experiment with the usage of single and bi-directional related networks in the encoder and following, we will try to improve the encoding procedure, adding an attention mechanism. Moreover, we will try to deal with problems arising during translation, using the beam search algorithm. Finally, we discuss weaknesses in matching concepts among the two languages as well as limitations due to the insufficient volume of data.

Key words

Machine Translation, Natural Language Processing, Sequence-to-Sequence Networks, Long Short Term Memory Networks, Attention Layers, Deep Learning, Beam Search

Ευχαριστίες

Η παρούσα διπλωματική εκπονήθηκε στο Εργαστήριο Ευφυών Συστημάτων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα Καθηγητή Ε.Μ.Π. κ. Ανδρέα-Γεώργιο Σταφυλοπάτη για την εμπιστοσύνη που μου έδειξε με την ανάθεση της συγκεκριμένης διπλωματικής εργασίας, καθώς επίσης και τον συνεπιβλέποντα κ. Γεώργιο Αλεξανδρίδη, Ε.ΔΙ.Π. Ε.Μ.Π., για τη συνεργασία και καθοδήγηση που μου παρείχε κατά την εκπόνηση της. Επίσης θα ήθελα να ευχαριστήσω τους κ.κ. Στέφανο Κόλλια , Καθηγητή Ε.Μ.Π. και Γεώργιο Στάμου, Αναπληρωτή Καθηγητή Ε.Μ.Π., για την τιμή που μου έκαναν να συμμετέχουν στην επιτροπή εξέτασης.

Ακόμα, θα ήθελα να εκφράσω από καρδιάς τις ευχαριστίες μου σε όλους όσους συνέβαλαν στην ολοκλήρωση αυτού του ταξιδιού και με βοήθησαν να φτάσω ως εδώ, ιδιαίτερα στους συμφοιτητές και, πλέον φίλους μου, που όλα αυτά τα χρόνια συνεργαστήκαμε και προσπαθήσαμε μαζί για το καλύτερο. Σε εκείνους που υπέμειναν τις ιδιοτροπίες μου και τις ανησυχίες μου για το μέλλον, οφείλω ένα μεγάλο ευχαριστώ.

Τέλος, στις αδερφές μου και στους γονείς μου Γιώργο και Μαρία, που με στηρίζουν και επικροτούν τις επιλογές μου από πολύ μικρή ηλικία.

Κυριακή Γ. Πέτρου,

Αθήνα, 24η Νοεμβρίου 2020

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

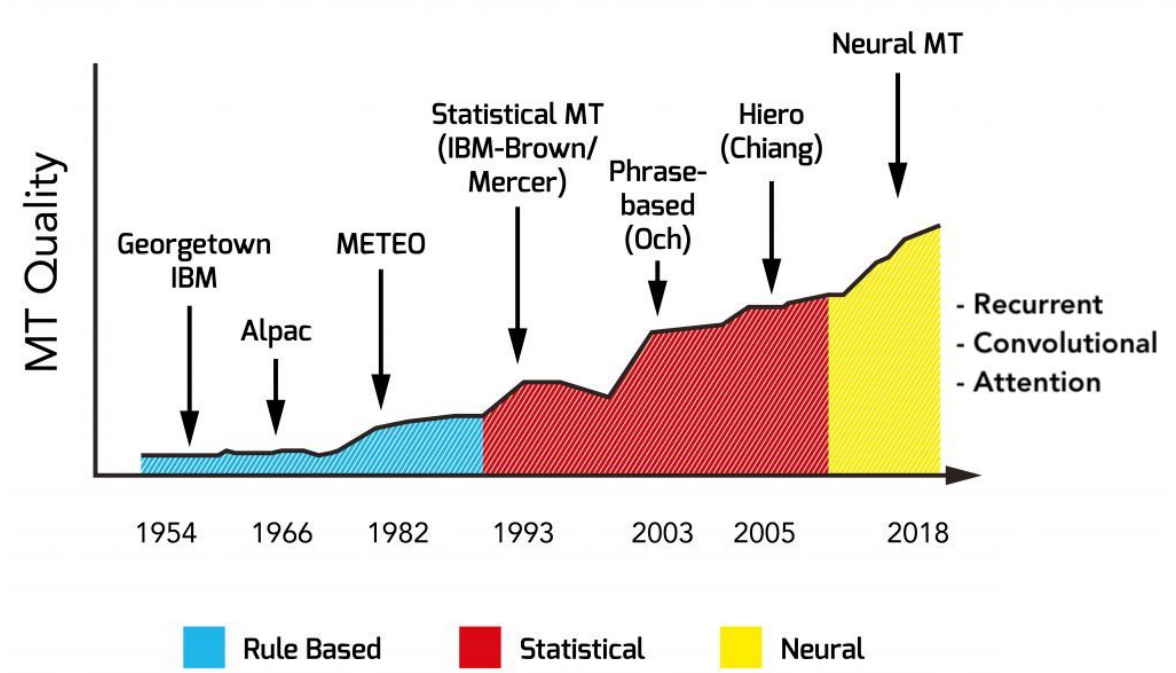
Εισαγωγή

Η *μηχανική μετάφραση* (machine tanslation - MT) είναι το υπο-πεδίο της υπολογιστικής γλωσσολογίας που μελετά την χρήση λογισμικών για την αυτόματη μετάφραση κειμένων από μία φυσική γλώσσα σε μία άλλη, χωρίς την ανάμειξη του ανθρώπου [Yang20b]. Σε ένα πρόβλημα μηχανικής μετάφρασης, η είσοδος αποτελείται από μία ακολουθία συμβόλων μίας γλώσσας και ο υπολογιστής καλείται να μετατρέψει αυτή την ακολουθία στην αντίστοιχή της σε μία άλλη γλώσσα [Good16]. Εκτός από το ερευνητικό ενδιαφέρον για την πλήρη κατανόηση των σημασιολογικών φαινομένων των γλωσσών στον τομέα της *επεξεργασίας της φυσικής γλώσσας* (natural language processing - NLP), η μηχανική μετάφραση έχει τεράστιο αντίκτυπο στη μείωση του κόστους για τη δημιουργία εφαρμογών που αποσκοπούν στην καλύτερη επικοινωνία των ανθρώπων.

Η ανάγκη επίλυσης του προβλήματος της μηχανικής μετάφρασης έχει απασχολήσει εδώ και πολλά χρόνια την ανθρωπότητα. Ανατρέχοντας στο 17^0 αιώνα, πολλοί μελετητές όπως ο Νεύτωνας και ο Ντεκάρτ είχαν εισάγει την ιδέα μιας παγκόσμια γλώσσας που θα μοιράζεται τα ίδια σύμβολα, εκφράζοντας το ίδιο νόημα σε όλες τις γλώσσες. Το 1933 ο Ρώσος Peter Troyanskii [Hutc04a] πρότεινε μια μηχανή αυτόματης μετάφρασης στην Ακαδημία Επιστημών της Σοβιετικής Ένωσης, η οποία θεωρήθηκε αρχικά ανώφελη. Οι προσπάθειες του Troyanskii αξιοποιήθηκαν αργότερα από άλλους ερευνητές και το 1954 δημιουργήθηκε για πρώτη φορά από την IBM και το Πανεπιστήμιο Georgetown μια μηχανή αυτόματης μετάφρασης που πέτυχε τη μετάφραση περισσότερων από εξήντα ρωσικών προτάσεων στα αγγλικά [Hutc04b]. Αργότερα, ο τομέας της Μηχανικής Μετάφρασης γνώρισε άνθιση σε τρεις διαφορετικές φάσεις (Σχήμα 1.1: (i) *μηχανική μετάφραση βάσει κανόνων* (rule-based machine translation -RBMT [Forc11], (ii) *στατιστική μηχανική μετάφραση* (statistical machine translation - SMT) [Koeh07, Koeh03], (iii) *μηχανική μετάφραση με χρήση νευρωνικών δικτύων* (neural machine translation - NMT) [Cho14].

Η μηχανική μετάφραση βάσει κανόνων αφορά συστήματα που βασίζονται σε γλωσσικούς κανόνες, λεξικά και γραμματικές που καλύπτουν τα κύρια σημασιολογικά, μορφολογικά και συντακτικά χαρακτηριστικά κάθε γλώσσας. Επιτρέπουν στις λέξεις να τοποθετούνται σε διαφορετικά μέρη και να έχουν διαφορετικές σημασίες ανάλογα με το περιβάλλον ώστε να χαρτογραφηθούν οι ιδιαιτερότητες της μετάφρασης μεταξύ των δύο γλωσσών. Το μειονέκτημα αυτής της μεθόδου έγκειται στο γεγονός ότι απαιτείται η δημιουργία κανόνων από γλωσσολόγους, το οποίο είναι εξαιρετικά χρονοβόρο, καθώς οι κανόνες αυτοί χαρακτηρίζουν τη συγκεκριμένη γλώσσα και μόνο. Παράλληλα, δεν υπάρχουν ισοδύναμες λέξεις σε όλες τις γλώσσες, ενώ αρκετές έχουν περισσότερες από μία σημασίες καθιστώντας τη διαδικασία αυτή μη αποδοτική.

Η στατιστική μηχανική μετάφραση είναι μια μέθοδος μηχανικής μετάφρασης όπου οι μεταφράσεις δημιουργούνται βάσει στατιστικών μοντέλων, των οποίων οι παράμετροι προέρχονται από την



Σχήμα 1.1: Η εξέλιξη της μηχανικής μετάφρασης

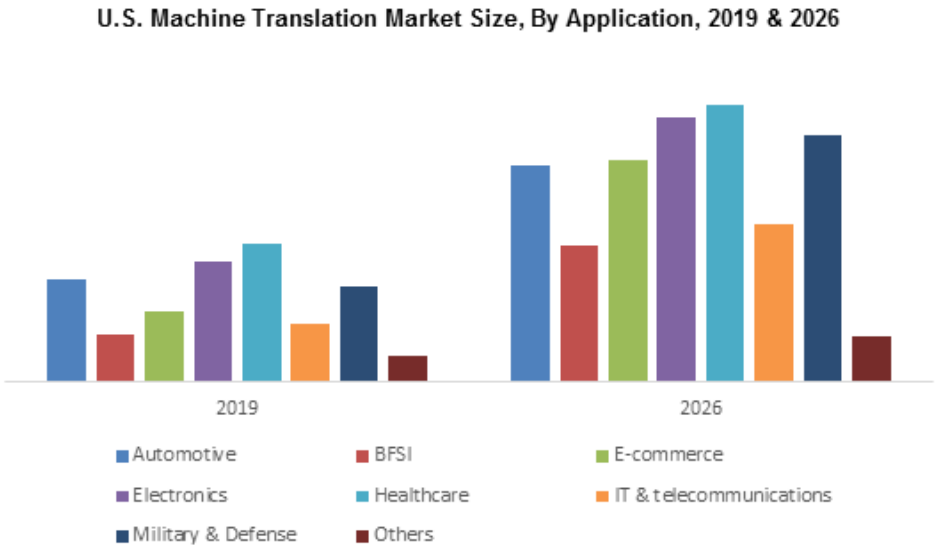
ανάλυση κειμένων του εξεταζόμενου ζεύγους γλωσσών. Η ιδέα της στατιστικής μηχανικής μετάφρασης προέρχεται από τη θεωρία της πληροφορίας, όπου μια ακολουθία λέξεων αντιστοιχίζεται σε μια μεταφρασμένη ακολουθία σύμφωνα με κατανομές πιθανοτήτων. Η μέθοδος αυτή απαιτεί την ύπαρξη μεταφρασμένων ζευγών για την εξαγωγή των στατιστικών μοντέλων, διαδικασία που είναι εξαιρετικά χρονοβόρα και ακριβή. Επίσης, τα αποτελέσματα δεν είναι ικανοποιητικά όταν πρόκειται για ζεύγη γλωσσών που δεν εμφανίζουν ομοιότητα στα γραμματικά πρότυπα που ακολουθούν (σειρά ρημάτων, επιθέτων, ουσιαστικών).

Η νευρωνική μηχανική μετάφραση είναι μία μέθοδος αυτόματης μετάφρασης που χρησιμοποιεί ένα τεχνητό νευρωνικό δίκτυο για να προβλέψει την πιθανότητα μιας ακολουθίας λέξεων, μοντελοποιώντας ολόκληρες προτάσεις σε ένα ενιαίο στατιστικό μοντέλο. Η μέθοδος αυτή έχει το πλεονέκτημα της απευθείας εκπαίδευσης στα δοσμένα κείμενα, με στόχο την πλήρη κατανόηση του περιεχομένου και της σημασίας των λέξεων σε αυτά, ώστε να αποδοθεί με τον καλύτερο δυνατό τρόπο η αντιστοίχηση των λέξεων στις δύο γλώσσες. Τα μοντέλα αυτά μπορούν επίσης να γενικεύσουν σε περισσότερες της μιας γλώσσες, δίνοντας ικανοποιητικά αποτελέσματα.

1.1 Η τρέχουσα κατάσταση στον τομέα της μηχανικής μετάφρασης

Σύμφωνα με έρευνα που διεξήχθη για τη σημαντικότητα της γλώσσας στη συμπεριφορά των καταναλωτών σε ένα σύνολο 2.400 καταναλωτών σε παγκόσμια κλίμακα [DePa06], προέκυψε ότι το 72, 1% των συμμετεχόντων σπαταλά περισσότερο χρόνο σε ιστοσελίδες όπου το περιεχόμενο εμφανίζεται στη δική του γλώσσα, ενώ το 56, 2% των ερωτηθέντων συμφωνεί ότι είναι πιο πιθανό να αγοράσει ένα προϊόν με χαρακτηριστικά στη μητρική του γλώσσα, ακόμα κι αν η τιμή του είναι πιο υψηλή. Η γλώσσα, συνεπώς, για την προτίμηση αγαθών και τη στοχευμένη διαφήμιση προϊόντων είναι καταλυτικής σημασίας, ιδίως σήμερα που η προσέλκυση του καταναλωτή βασίζεται κατεξοχήν

σε διαδικτυακές διαφημίσεις, οπότε τα συστήματα αυτόματης μετάφρασης μπορούν να επιφέρουν σημαντικό οικονομικό όφελος στις επιχειρήσεις που τα αξιοποιούν.



Σχήμα 1.2: Η αναμενόμενη διείσδυση της μηχανικής μετάφρασης ανά επιχειρηματικό κλάδο στις ΗΠΑ (Πηγή [Pree20])

Ένα άλλο παράδειγμα που καθιστά εξαιρετικά σημαντική την μηχανική μετάφραση είναι η γρήγορη επικοινωνία μεταξύ αλλογλώσσων, ιδίως σε πραγματικό χρόνο (ηλεκτρονική αλληλογραφία, συνομιλίες, διακρατικές συνεδριάσεις και άλλα). Παράδειγμα τέτοιας λειτουργίας αποτελούν οι διάφορες online μεταφραστικές υπηρεσίες, όπως το Google Translate[1], οι οποίες μάλιστα εμφανίζουν αρκετά ικανοποιητικά αποτελέσματα. Βλέπουμε, συνεπώς, πως η ύπαρξη συστημάτων αυτόματης μετάφρασης είναι καταλυτικής σημασίας για την εκπλήρωση πολλών καθημερινών μας συνηθειών σε σύντομο χρονικό διάστημα. Το πρόβλημα είναι εξίσου σημαντικό σε γλώσσες με ελάχιστα διαθέσιμα δεδομένα για εκπαίδευση. [Gu18]

Στο Σχήμα 1.2 απεικονίζεται η πρόβλεψη για την εξέλιξη της διείσδυσης της μηχανικής μετάφρασης σε διάφορους επιχειρηματικούς τομείς στις ΗΠΑ ως το 2026, από ιστορικά δεδομένα της χρονικής περιόδου 2016-2019 [Pree20]. Το 2019 η παγκόσμια βιομηχανία της Μηχανικής Μετάφρασης άγγιζε τα 550 εκατομμύρια δολάρια και υπολογίζεται έως το 2026 να φτάσει τα 1.5 δισεκατομμύρια, έχοντας σύνθετο ετήσιο ρυθμό ανάπτυξης 17%. Σύμφωνα με στατιστικά δεδομένα, αναμένεται μεγάλη διείσδυση της μηχανικής μετάφρασης στους τομείς των τηλεπικοινωνιών και των τεχνολογιών πληροφορικής, με ρυθμό αύξησης 15% ως το 2026, λόγω της ανάγκης για δημιουργία ενιαίων συστημάτων που θα διευκολύνουν τη διαγλωσσική μεταφορά των δεδομένων σε σύντομο χρόνο. Εξίσου σημαντική φαίνεται να είναι και η διείσδυση της αυτοματοποιημένης μετάφρασης στον τομέα της υγείας, που σήμερα, με την πανδημία του Covid-19, είναι καθοριστικής σημασίας η άμεση πρόσβαση σε έρευνες και αποτελέσματα για την εξαγωγή συμπερασμάτων, αλλά και την ενημέρωση όλων των ανθρώπων για φλέγοντα ζητήματα ανεξαρτήτως γλώσσας [Way20].

Το πρόβλημα αυτό, συνεπώς, αποτελεί ένα ανοιχτό ερευνητικό πεδίο στο οποίο επενδύονται πολλά χρήματα παγκοσμίως και με το οποίο ασχολούνται ολοένα και περισσότεροι ερευνητές. Στο

[1] https://translate.google.com/

πλαίσιο της παρούσας εργασίας, θα διερευνήσουμε μεθόδους που μπορούν να επιλύσουν εν μέρει το πρόβλημα και να αποκαλύψουν προκλήσεις για το μέλλον.

1.2 Η προσέγγιση της παρούσας εργασίας

Στην παρούσα διπλωματική διατυπώνουμε το πρόβλημα της αυτόματης μηχανικής μετάφρασης της αγγλικής γλώσσας στην ελληνική. Η μηχανική μετάφραση με χρήση νευρωνικών δικτύων αποσκοπεί στη δημιουργία μοντέλων μετάφρασης που κατανοούν το περιεχόμενο της πρότασης και παράγουν μεταφράσεις λαμβάνοντας υπόψη όλες τις πιθανές εξαρτήσεις στο κείμενο. Οι κύριες μεθοδολογίες για την αντιμετώπιση του συγκεκριμένου προβλήματος εστιάζουν στην επεξεργασία ακολουθιακών δεδομένων διαφορετικού μήκους εισόδου-εξόδου, γεγονός που δυσχέραινε αρχικά την επιτυχή μετάφρασή τους. Εμείς, θα διερευνήσουμε μεθόδους που αξιοποιούν ακολουθιακές αρχιτεκτονικές και δίνουν ενθαρρυντικά αποτελέσματα χρησιμοποιώντας ζεύγη προτάσεων στις δύο γλώσσες.

Για την επίλυση του προβλήματος της μηχανικής μετάφρασης αγγλικού κειμένου στην ελληνική γλώσσα, θα επιχειρήσουμε να αξιοποιήσουμε τη μηχανική μάθηση, συνδυάζοντας ευρέως χρησιμοποιούμενες τεχνικές νευρωνικών δικτύων. Έχοντας ως σύνολο δεδομένων ζεύγη προτάσεων που αντιστοιχούν στις δυο γλώσσες, θα δημιουργήσουμε χρήσιμες αναπαραστάσεις των δειγμάτων στην προσπάθειά μας να εξάγουμε όσο το δυνατό περισσότερη πληροφορία. Σημειώνεται επίσης, ότι διαθέτουμε ένα μικρό σύνολο δεδομένων και το γεγονός αυτό δυσχεραίνει τη διαδικασία της μετάφρασης, αν αναλογιστούμε ότι στην ελληνική γλώσσα υπάρχει πληθώρα γλωσσικών φαινομένων που δεν απαντώνται στην αγγλική. Για παράδειγμα, το άρθρο `the` στα αγγλικά μπορεί να αφορά οποιοδήποτε από τα τρία γένη στα ελληνικά (ο,η,το). Ένα άλλο παράδειγμα αποτελεί η κλίση των επιθέτων, δηλαδή το επίθετο `happy` δεν μπορεί να κλιθεί, σε αντίθεση με το επίθετο ευτυχισμένος-η-ο που η κατάληξή του υποδηλώνει σε κάθε περίπτωση διαφορετικό γένος. Τις διαφορές αυτές καλούμαστε να αναδείξουμε και να επιλύσουμε στην παρούσα εργασία.

Πιο συγκεκριμένα, θα χρησιμοποιήσουμε την τεχνική sequence-to-sequence, μια κλασική μέθοδο μετάφρασης, που περιλαμβάνει έναν κωδικοποιητή και έναν αποκωδικοποιητή. Στη συνέχεια θα πειραματιστούμε με την αρχιτεκτονική αυτή δοκιμάζοντας αμφίδρομα δίκτυα βραχυπρόθεσμης μνήμης ώστε να ανιχνεύσουμε εξαρτήσεις σε όλο το μήκος των ακολουθιών. Προς αυτή την κατεύθυνση της ανίχνευσής εξαρτήσεων, θα προσθέσουμε στην αρχιτεκτονική αυτή το δημοφιλή μηχανισμό προσοχής, που χρησιμοποιείται πλέον σε όλες τις σύγχρονες τεχνικές μηχανικής μάθησης. Τέλος, για να βελτιστοποιήσουμε τη διαδικασία της αποκωδικοποίησης, χρησιμοποιήσαμε τον αλγόριθμο ακτινικής αναζήτησης στο μοντέλο sequence-to-sequence, μια προσθήκη που ήταν καταλυτικής σημασίας για τη διαμόρφωση ενός αξιόλογου αποτελέσματος.

Η υπόλοιπη διατριβή διαρθρώνεται ως εξής: στο Κεφάλαιο 2, περιγράφεται το τεχνικό υπόβαθρο, με τις αντίστοιχες παραπομπές στα πρωτότυπα κείμενα, που απαιτείται για την πλήρη κατανόηση της συγκεκριμένης εργασίας. Στο κεφάλαιο αυτό, δίνεται μια εκτενέστερη αναφορά στα εργαλεία που χρησιμοποιήθηκαν στα τελικά μοντέλα και είχαν τη μεγαλύτερη συμβολή στα τελικά αποτελέσματα της πειραματικής διαδικασίας που διεξήχθη. Στο Κεφάλαιο 4, αποτυπώνεται η περιγραφή των αρχιτεκτονικών που δημιουργήθηκαν στα πλαίσια αυτής της εργασίας και τα στάδια εξέλιξης του μοντέλου, από την πρώτη υλοποίηση έως τις τελικές βελτιστοποιήσεις των αρχιτεκτονικών που επέφεραν τα τελικά αποτελέσματα. Στο Κεφάλαιο 5 δίνεται αρχικά, η περιγραφή του συνόλου των δεδομένων

που χρησιμοποιήθηκαν και η επεξεργασία τους για την εκπαίδευση και αξιολόγηση της πειραματικής διαδικασίας. Έπειτα, παρατίθενται τα αποτελέσματα και ο σχολιασμός κάθε υλοποίησης. Τέλος, στο Κεφάλαιο 6 καταγράφουμε τα βασικά συμπεράσματα της εργασίας καθώς και πιθανές μελλοντικές ερευνητικές κατευθύνσεις.

Κεφάλαιο 2

Μηχανική Μάθηση

Η Μηχανική Μάθηση αποτελεί ένα υποπεδίο της επιστήμης των υπολογιστών και συγκεκριμένα της τεχνητής νοημοσύνης που έχει ως αντικείμενο τη δημιουργία υπολογιστικών συστημάτων ικανών να μαθαίνουν βάσει προτύπων, να βελτιώνουν δηλαδή την αποτελεσματικότητά τους μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας όπως ο άνθρωπος [Mitc97]. Εμπειρία αποτελούν τα δεδομένα εισόδου, ή αλλιώς δεδομένα εκπαίδευσης, τα οποία δέχονται την απαραίτητη προ-επεξεργασία προκειμένου να έρθουν σε κατάλληλη μορφή για να μπορέσει το σύστημα να εξάγει χρήσιμη γνώση από αυτά. Οι αλγόριθμοι μηχανικής μάθησης ταξινομούνται σε τρεις κύριες κατηγορίες, όπως φαίνεται παρακάτω:

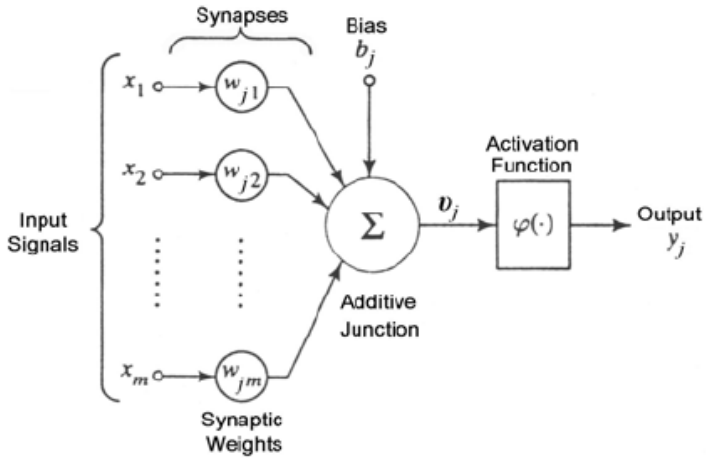
1. **Επιβλεπόμενη Μάθηση** (Supervised Learning), όπου το σύστημα τροφοδοτείται με δεδομένα εκπαίδευσης επισημειωμένα με μια *ετικέτα* (label). Η ετικέτα αυτή αποτελεί την επιθυμητή έξοδο. Ο αλγόριθμος προσπαθεί να μοντελοποιήσει τη σχέση ανάμεσα στα δεδομένα εισόδου και εξόδου , συγκρίνοντας την έξοδό του με την επιθυμητή ώστε να μαθαίνει προς τη σωστή «κατεύθυνση» και να μειώνει το λάθος ταξινόμησης κατά την διαδικασία της εκπαίδευσης.

2. **Μη-επιβλεπόμενη Μάθηση** (Unsupervised Learning), όπου, σε αντίθεση με την προηγούμενη περίπτωση, δεν υπάρχουν ετικέτες για τα δεδομένα. Το σύστημα προσπαθεί να εντοπίσει πρότυπα, να βρει κανόνες και να συσχετίσει τα δεδομένα εισόδου μεταξύ τους. Η μέθοδος αυτή εφαρμόζεται σε προβλήματα *συσταδοποίησης* (clustering), όπου ο αλγόριθμος επιδιώκει να δημιουργήσει *συστάδες* (clusters) και να τοποθετήσει τα δεδομένα που παρουσιάζουν τη μεγαλύτερη ομοιότητα στην ίδια συστάδα.

3. **Ενισχυτική Μάθηση** (Reinforcement Learning). Σε αυτή την περίπτωση το σύστημα επιλέγει επαναληπτικά μια δράση χωρίς να διαθέτει την επιθυμητή έξοδο, λαμβάνει μια ανατροφοδότηση από το περιβάλλον και μεταβαίνει σε μία κατάσταση. Από το ερέθισμα που του δίνεται κάθε φορά προσπαθεί να καταλάβει αν η επιλογή του ήταν σωστή. Η κύρια διαφορά με την επιβλεπόμενη μάθηση είναι ότι στην ενισχυτική μάθηση οι αποφάσεις λαμβάνονται ακολουθιακά. Με άλλα λόγια, η έξοδος εξαρτάται από την κατάσταση της τρέχουσας εισόδου και η επόμενη είσοδος εξαρτάται από την έξοδο της προηγούμενης εισόδου, δηλαδή οι αποφάσεις του συστήματος είναι εξαρτώμενες η μία με την άλλη. Αντίθετα, στις άλλες δύο περιπτώσεις μάθησης, οι αποφάσεις του συστήματος είναι ανεξάρτητες μεταξύ τους (για διαφορετικές εισόδους).

2.1 Τεχνητά Νευρωνικά Δίκτυα

Ένα *τεχνητό νευρωνικό δίκτυο* (Artificial Neural Network ή ANN) είναι ένα υπολογιστικό μοντέλο που προσομοιώνει τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου. Παρόλο που ο ανθρώπινος εγκέφαλος αποτελεί πεδίο ερευνών εδώ και πολλά χρόνια, το πρώτο βήμα για την μοντελοποίησή του έγινε το 1943 από τους Warren McCulloch και Walter Pitts [Palm86], όταν υλοποίησαν ένα νευρωνικό δίκτυο με ηλεκτρικά κυκλώματα. Στη συνέχεια, το 1958, ο Frank Rosenblatt εισήγαγε την έννοια του νευρώνα Perceptron, μιας δομής που αποτέλεσε βάση για τη δημιουργία των νευρωνικών δικτύων που χρησιμοποιούμε σήμερα. 'Ενας νευρώνας δέχεται ένα ερέθισμα μέσω των δενδρικών απολήξεων του και παράγει ένα σήμα εξόδου στον άξονά του. Ο άξονας του συνδέεται μέσω συνάψεων με δενδρίτες άλλων νευρώνων. Στη γενική του μορφή ένας νευρώνας Perceptron (Σχήμα 2.1) δέχεται ένα διάνυσμα $x[i]$ τιμών (σήμα εισόδου) και $w[i]$ βαρών, όπου λαμβάνοντας το σταθμισμένο μέσο τους $\sum w[i] * x[i]$ και περνώντας το αποτέλεσμα αυτό από μία *συνάρτηση ενεργοποίησης* (activation function), η έξοδος ενεργοποιείται (γίνεται 1) αν το άθροισμα αυτό είναι μεγαλύτερο από ένα κατώφλι, διαφορετικά δεν ενεργοποιείται (γίνεται μηδέν) [Frea90].



Σχήμα 2.1: Αναπαράσταση ενός τεχνητού νευρώνα

Τα δομικά στοιχεία ενός νευρώνα μπορούν να συνοψιστούν στην παρακάτω λίστα:

- **Είσοδος**: Διάνυσμα εισόδου $x[i]$ που περιέχει την πληροφορία που θα τροφοδοτηθεί στο νευρώνα.
- **Βάρη**: Τα βάρη του δικτύου $w[i]$ αναπαριστούν βαθμωτούς πολλαπλασιασμούς. Ο ρόλος τους είναι να εκτιμήσουν πόσο σημαντική είναι η κάθε είσοδος αλλά και πως η αύξηση ή η μείωση της εκάστοτε εισόδου επηρεάζει την έξοδο.
- **Αθροιστής**: Ο αθροιστής δέχεται τις εισόδους και πολλαπλασιάζοντάς τες με τα βάρη δίνει μια έξοδο.
- **Συνάρτηση ενεργοποίησης**. Καθορίζει υπό ποιες προϋποθέσεις η είσοδος του νευρώνα τον ενεργοποιεί
- **Πόλωση**. Η *πόλωση* (bias) είναι ουσιαστικά μια επιπλέον σταθερή τιμή που προστείθεται στο σταθμισμένο άθροισμα της εισόδου. Στην πράξη, η ύπαρξή της μετατοπίζει (προς τα πάνω ή προς τα κάτω) την τιμή της εισόδου και άρα επηρεάζει την ενεργοποίηση του νευρώνα.

2.1.1 Συναρτήσεις Ενεργοποίησης

Όπως αναφέρθηκε προηγουμένως, οι συναρτήσεις ενεργοποίησης καθορίζουν τις προϋποθέσεις κάτω από τις οποίες η είσοδος ενεργοποιεί τον νευρώνα. Από τις πιο ευρέως χρησιμοποιούμενες είναι η σιγμοειδής, η οποία περιορίζει την είσοδό της στο διάστημα $[0, 1]$, καθιστώντας την χρήσιμη σε περιπτώσεις που επιθυμούμε λ.χ. το δίκτυο να υπολογίζει πιθανότητες. Η σιγμοειδής συνάρτηση περιγράφεται από τον μαθηματικό τύπο της Εξίσωσης 2.1

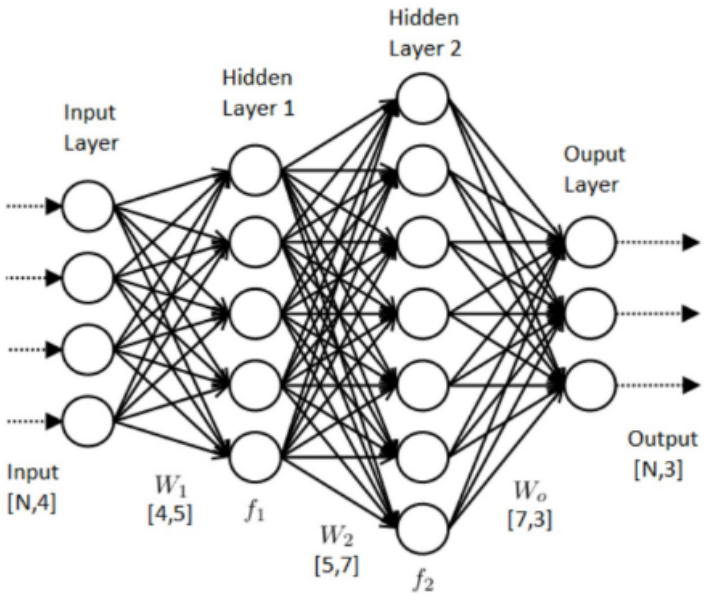
$$f(x) = \frac{1}{1 - e^{-x}} \tag{2.1}$$

Μια εναλλακτική συνάρτηση ενεργοποίησης που χρησιμοποιείται σε νευρωνικά δίκτυα, όταν οι πιθανές κλάσεις κατηγοριοποίησης των δεδομένων της εισόδου είναι παραπάνω από δύο είναι η *softmax*, που περιγράφεται στην Εξίσωση 2.2.

$$Softmax(y_i) = \frac{e^{y_i}}{\sum_{i=0}^{N} e^{y_i}} \tag{2.2}$$

2.1.2 Πολυστρωματικοί Νευρώνες (Multilayer Perceptron)

Ο απλός νευρώνας που περιγράψαμε στην Ενότητα 2.1 χρησιμοποιείται για να επιλύσει απλά προβλήματα. Η σύνδεση πολλών νευρώνων σε διαφορετικά επίπεδα δύναται, ωστόσο, να επιλύσει πιο σύνθετα προβλήματα. Έτσι κατασκευάζεται μια αρχιτεκτονική που είναι γνωστή ως *perceptron πολλαπλών επιπέδων* (multi-layer perceptron - MLP), όπου οι νευρώνες οργανώνονται σε επίπεδα, όπως φαίνεται και στο Σχήμα 2.2



Σχήμα 2.2: Πολυεπίπεδο Νευρωνικό Δίκτυο (Πηγή [Hayk09])

Τα επίπεδα ενός MLP διακρίνονται σε τρία είδη: (i) *επίπεδο εισόδου* (input layer), (ii) *κρυφά επίπεδα* (hidden layer), (iii) *επίπεδο εξόδου* (output layer). Το επίπεδο εισόδου δέχεται τα δεδομένα και τα προωθεί στα επόμενα- κρυφά επίπεδα. Κάθε νευρώνας του i επιπέδου, τροφοδοτείται από τις εξόδους του νευρώνα του $i-1$ επιπέδου. Η έξοδός του εισάγεται στους νευρώνες του επόμενου επιπέδου ή στην έξοδο του δικτύου αν ο νευρώνας ανήκει στο τελευταίο επίπεδο. Τα ενδιάμεσα επίπεδα, που

καλούνται και κρυφά, μπορούν να ποικίλλουν σε πλήθος ανάλογα με το πλήθος και των δεδομένων. Αν τα κρυφά επίπεδα είναι παραπάνω από 2, τότε μιλάμε για αρχιτεκτονικές *βαθιών νευρωνικών δικτύων* (deep neural networks) και για *βαθιά μάθηση* (deep learning). Το τελευταίο επίπεδο είναι το επίπεδο εξόδου και το πλήθος των νευρώνων που υπάρχουν σε αυτό εξαρτάται από το πρόβλημα και τον αριθμό των κλάσεων στις οποίες εντάσσονται τα δεδομένα. Για παράδειγμα, σε ένα πρόβλημα παλινδρόμησης μπορεί να υπάρχει ένα ς νευρώνας εξόδου, ενώ σε ένα πρόβλημα ταξινόμησης πολλών κλάσεων μπορούν να υπάρχουν πολλαπλοί νευρώνες εξόδου (ίσοι με το πλήθος των κλάσεων).

2.1.3 Συναρτήσεις Κόστους

Στόχος ενός νευρωνικού δικτύου είναι η εκμάθηση των παραμέτρων του (βάρη, πολώσεις) από τα δοσμένα δεδομένα εκπαίδευσης. Η διαδικασία αυτή επιτυγχάνεται ποσοτικοποιώντας το σφάλμα μεταξύ της πραγματικής τιμής της εξόδου (ετικέτας δεδομεων) και της τιμής που το δίκτυο τελικά πρόβλεψε σε κάθε βήμα. Για να μετρήσουμε την αποτελεσματικότητα του δικτύου κατά τη διαδικασία της εκπαίδευσης, χρησιμοποιούμε *συναρτήσεις κόστους* (loss functions). Μια από τις πιο συχνά χρησιμοποιούμενες συναρτήσεις κόστους είναι η *εγκάρσια εντροπία* (cross-entropy), που αθροίζει τις πιθανότητες της κάθε κλάσης που τελικά πρόβλεψε το δίκτυο ως έξοδο. Ο μαθηματικός τύπος της εγκάρσιας εντροπίας για περισσότερες από δύο κλάσεις ($M > 2$) δίνεται από την Εξίσωση 2.3

$$\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{2.3}$$

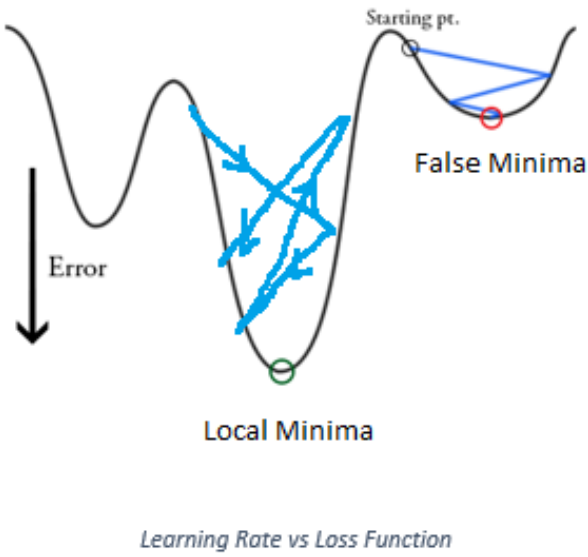
όπου, M είναι το πλήθος των κλάσεων, log ο φυσικός λογάριθμος, y δυαδικός δείκτης (0 ή 1) αν η ετικέτα της κλάσης c είναι η σωστή για την παρατήρηση o και p προβλεπόμενη πιθανότητα της παρατήρησης o για τον αν ανήκει στην κλάση c.

2.1.4 Αλγόριθμος Κατάβασης Κλίσης

Ο αλγόριθμος *κατάβασης κλίσης* (gradient descent) είναι ένας ευρέως χρησιμοποιούμενος αλγόριθμος στη μηχανική μάθηση, που έχει ως στόχο την εύρεση των βέλτιστων παραμέτρων ενός νευρωνικού δικτύου. Η βελτιστοποίηση αυτή επιτυγχάνεται μέσω της ελαχιστοποίησης της συνάρτησης κόστους $J(\theta)$ (Ενότητα 2.1.3), όπου θ είναι το διάνυσμα που αντιπροσωπεύει τις παραμέτρους του δικτύου. Για να βρεθεί το ελάχιστο της συνάρτησης κόστους πρέπει να υπολογίσουμε την κλίση της συνάρτησης ώστε να αποφανθούμε σε ποια κατεύθυνση πρέπει να κινηθούμε [Rude17]. Η κλίση υπολογίζεται παραγωγίζοντας τη συνάρτηση κόστους ως προς τις παραμέτρους του δικτύου, όπως φαίνεται στην Εξίσωση 2.4

$$\theta_j \leftarrow \theta_j - \eta \frac{\partial}{\partial \theta_j} J(\theta) \tag{2.4}$$

όπου η ο ρυθμός προσαρμογής των παραμέτρων του δικτύου σε κάθε βήμα εκπαίδευσης, γνωστός και ως *ρυθμός μάθησης* (learning rate). Η υπερπαράμετρος αυτή επιλέγεται έτσι ώστε το μοντέλο να μην μαθαίνει με εξαιρετικά αργό ρυθμό, κάνοντας χρονοβόρα τη διαδικασία της εκμάθησης, αλλά και ούτε με μεγάλο ρυθμό, καθώς υπάρχει ο κίνδυνος να «προσπεράσουμε» το ελάχιστο της συνάρτησης, εκτελώντας ταλαντώσεις γύρω από αυτό (Σχήμα 2.3). Συνήθως ο ρυθμός μάθησης είναι μεγάλος στα πρώτα βήματα και μειώνεται σταδιακά, όσο προχωράει η εκπαίδευση.



Σχήμα 2.3: Ο αλγόριθμος κατάβασης κλίσης (Πηγή [Sudh17])

2.2 Αναδρομικά Νευρωνικά Δίκτυα

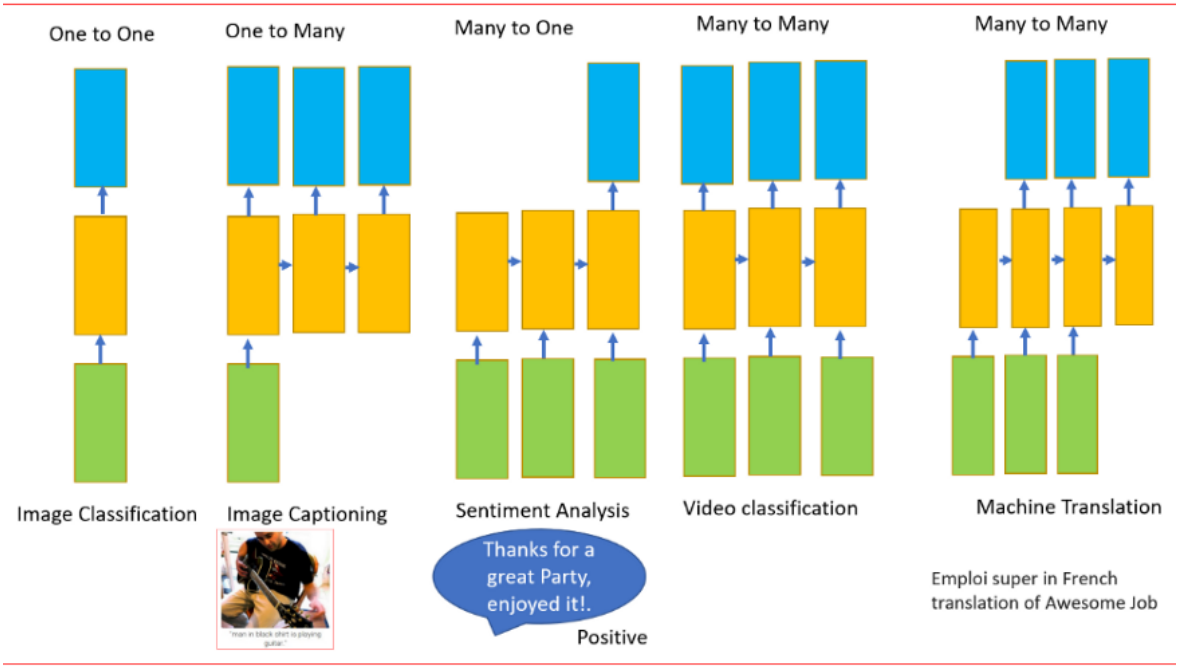
Σε μία ακολουθία, η σειρά εμφάνισης των δεδομένων αποτελεί καταλυτικό παράγοντα για την ταξινόμηση της. Η ακολουθιακή μάθηση εφαρμόζεται σε δεδομένα κειμένου, λόγου, βίντεο όπου η πληροφορία ξεδιπλώνεται στο χρόνο και η πρόβλεψη τη χρονική στιγμή t εξαρτάται από το περιεχόμενο των προηγούμενων εισόδων. Για παράδειγμα, σε μια πρόταση, η κάθε λέξη διαδέχεται μια άλλη οδηγώντας σε νοηματική εξάρτηση μεταξύ αυτών. Στην καθημερινή μας ζωή, πολλές από τις εφαρμογές που χρησιμοποιούμε όπως η μηχανές αυτόματης μετάφρασης ή διάφοροι διαλογικοί εικονικοί πράκτορες (chatbots) βασίζονται σε συστήματα ακολουθιακής μάθησης.

Τα αναδρομικά νευρωνικά δίκτυα (recurrent neural networks - RNN) είναι μία κατηγορία νευρωνικών δικτύων που έχουν ως χαρακτηριστικό να συνδέουν προηγούμενες εισόδους με την τρέχουσα είσοδο. Κάθε κόμβος του δικτύου έχει τη δυνατότητα να λαμβάνει πληροφορία τόσο από τις τιμές των κρυμμένων κόμβων όσο και από την τρέχουσα είσοδο, δίνοντας με αυτό τον τρόπο το πλεονέκτημα της μνήμης. Η ακολουθιακή πληροφορία διατηρείται στην *κρυφή κατάσταση* (hidden state). Εκεί το μοντέλο αναζητά συσχετίσεις μεταξύ των δεδομένων που έχουν λάβει χώρα σε ξεχωριστές χρονικές στιγμές, οι οποίες μπορεί να είναι αρκετά απομακρυσμένες χρονικά μεταξύ τους (long-term dependencies).

Το Σχήμα 2.4 συνοψίζει τις κατηγορίες στις οποίες εντάσσονται τα προβλήματα ακολουθιακής μάθησης Πιο συγκεκριμένα, αυτά αναλύονται σε:

- **Ένα-προς-ένα** (One-to-One). Στην κατηγορία αυτή, έχουμε μία είσοδο και μία έξοδο. Για παράδειγμα, έχοντας ως είσοδο μία εικόνα θέλουμε να την ταξινομήσουμε, δίνοντάς της μία και μόνο ετικέτα (π.χ. σκύλος ή γάτα). Η εικόνα δεν μπορεί να ανήκει σε περισσότερες από μία κλάσεις.

- **Πολλά-σε-ένα** (Many-to-One). Στην κατηγορία αυτή, έχουμε μια ακολουθία δεδομένων εισόδου και επιθυμούμε να την ταξινομήσουμε σε μία κλάση (έξοδο). Χαρακτηριστικό παράδειγμα



Σχήμα 2.4: Κατηγορίες αναδρομικών νευρωνικών δικτύων

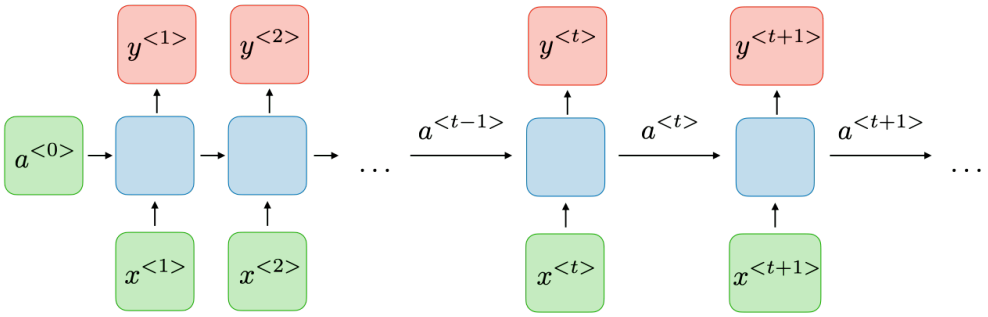
της κατηγορίας αυτής είναι μια πρόταση όπου στόχος είναι να αποφανθούμε αν το περιεχόμενο αφορά χαρά ή λύπη.

- **Ένα-σε-πολλά** (One-to-Many). Στην περίπτωση αυτή, έχουμε μία είσοδο και μια ακολουθία για έξοδο. Για παράδειγμα δίνοντας μιας εικόνα ως είσοδο, στόχος μας είναι να παράξουμε μια περιγραφή (ακολουθία-πρόταση) ως έξοδο για την εικόνα αυτή. Η έξοδος εδώ αφορά πρόταση, άρα περισσότερες της μιας λέξεις.

- **Πολλά-σε-πολλά** (Many-to-Many). Τέλος, η κατηγορία πολλά-σε-πολλά αφορά προβλήματα όπου η είσοδος και η έξοδος αποτελούνται από ακολουθίες, το μήκος των οποίων μπορεί να μεταβάλλεται. Στην κατηγορία αυτή ανήκει και το πρόβλημα της αυτόματης μηχανικής μετάφρασης με το οποίο εξετάζουμε στην παρούσα εργασία.

Στην περίπτωση που η ακολουθία εισόδου έχει το ίδιο μήκος με την ακολουθία εξόδου, η αντιστοίχηση των δύο ακολουθιών μπορεί εύκολα να επιτευχθεί με τη χρήση RNN ή *δικτύων μακράς βραχυπρόθεσμης μνήμης* (LSTMs), όπως παρουσιάστηκε στην προηγούμενη ενότητα. Οι προαναφερόμενες μέθοδοι παρουσιάζουν ικανοποιητικά αποτελέσματα για όλες τις κατηγορίες πλην της τελευταίας (πολλά-σε-πολλά). Ειδικά στην περιοχή της μηχανικής μετάφρασης που εξετάζουμε, αυτό γίνεται κατανοητό με το ακόλουθο παράδειγμα: μεταφράζοντας την πρόταση `What are you doing today?` από τα αγγλικά στα ελληνικά, η είσοδος αποτελείται από 5 λέξεις και η έξοδος από 3 (`Τι κάνεις σήμερα;`). Στο σημείο αυτό, γίνεται κατανοητό πως μία μονάδα LSTM ή GRU είναι αδύνατο να αντιστοιχίσει μία λέξη σε περισσότερες ώστε να παραχθεί η σωστή μετάφραση. Η λύση δόθηκε για πρώτη φορά το 2014 όταν η Google εισήγαγε το μοντέλο *ακολουθίας-σε-ακολουθία* (sequence-to-sequence) επιχειρώντας να αντιστοιχίσει διαφορετικού μήκους ακολουθίες [Suts14].

Το Σχήμα 2.5 συνοψίζει τα δομικά στοιχεία της αρχιτεκτονικής ενός RNN. Σε κάθε χρονική στιγμή



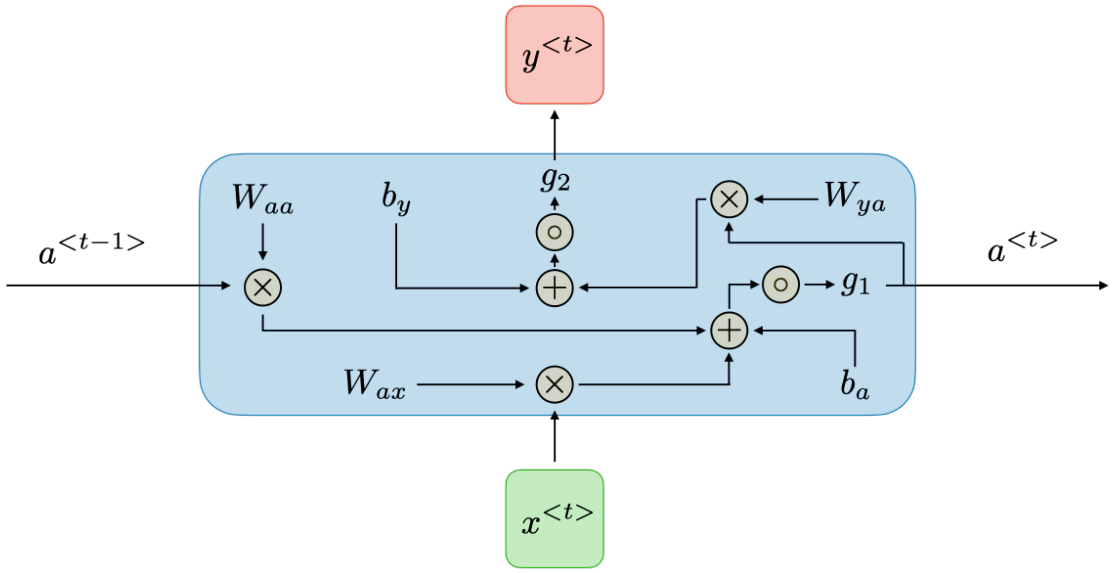
Σχήμα 2.5: Αρχιτεκτονική ενός αναδρομικού νευρωνικού δικτύου (Πηγή [Amid20])

t, το κρυφό επίπεδο $a^{<t>}$ και η έξοδος $y^{<t>}$ δίνονται από τις Εξισώσεις 2.5 και 2.6 αντίστοιχα:

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a) \tag{2.5}$$

$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y) \tag{2.6}$$

όπου W_{ax}, W_{aa}, W_{ya} είναι τα διανύσματα βαρών του δικτύου, b_a, b_y είναι οι τιμές πόλωσης για την κρυφή κατάσταση και την έξοδο αντίστοιχα, g_1,g_2 οι συναρτήσεις ενεργοποίησης και x το διάνυσμα εισόδου. Η εσωτερική δομή μιας μονάδας ενός δικτύου RNN δίνεται στο Σχήμα 2.6



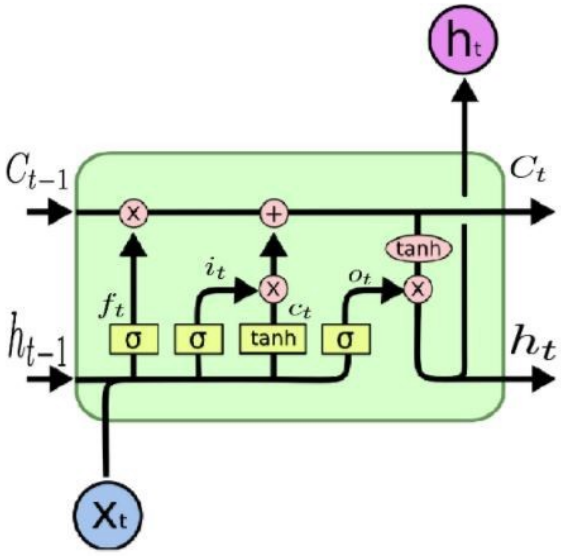
Σχήμα 2.6: Εσωτερική δομή της μονάδας ενός RNN (Πηγή [Amid20])

Στην παρούσα διπλωματική εργασία, θα ασχοληθούμε εκτενώς με τα RNN και ειδικότερα, με την αρχιτεκτονική πολλά-προς-πολλά, όπου το μήκος της ακολουθίας εισόδου διαφέρει από αυτό της εξόδου ($T_x \neq T_y$) για την επίλυση του προβλήματος της Μηχανικής Μετάφρασης.

2.2.1 Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης

Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης, στα οποία αναφερθήκαμε πρώτη φορά στην προηγούμενη υπό-ενότητα, ανήκουν και αυτά στα αναδρομικά νευρωνικά δίκτυα και δίνουν τη δυνατότητα μάθησης μεγάλου μήκους ακολουθιών. Είναι γνωστό πως τα RNN αδυνατούν να διατηρήσουν

την πληροφορία σε μεγάλου μήκους ακολουθίες, εκεί δηλαδή που υπάρχει εξάρτηση από εισόδους που εμφανίστηκαν πολλά χρονικά βήματα πριν. Το γεγονός αυτό, οδηγεί στο γνωστό πρόβλημα των *εκλειπόμενων και εκρυγνυόμενων κλίσεων* (vanishing and exploding gradients), όπου μη γνωρίζοντας τη σωστή εξάρτηση μεταξύ των δεδομένων, δεν μπορούμε να αποφανθούμε από την κλίση σε ποια κατεύθυνση θα μεταβάλλουμε τα βάρη. Συνοπτικά, οι παράγωγοι των επιπέδων των νευρωνικών δικτύων πολλαπλασιάζονται μεταξύ τους με αποτέλεσμα, αν οι ποσότητες αυτές είναι πολύ μικρές, να έχουμε εξάλειψη των παραγώγων (εκλειπόμενες κλίσεις), ενώ αν πολλαπλασιάζουμε συνεχώς με έναν αριθμό μεγαλύτερο της μονάδας, να έχουμε ως αποτέλεσμα μια μεγάλη αύξηση (εκρυγνυόμενες κλίσεις). Η λύση στο πρόβλημα αυτό ήρθε για πρώτη φορά το 1997 από τους Hochreiter & Schmidhuber με τα LSTM [Hoch97].



Σχήμα 2.7: Ένας LSTM νευρώνας (κύτταρο) (Πηγή [Olah15])

Ένας LSTM κύτταρο διαθέτει εσωτερικούς μηχανισμούς-*πύλες* (gated cells), όπως φαίνεται στο Σχήμα 2.7, που ορίζουν ποια πληροφορία θα απολεσθεί και ποια θα αποθηκευτεί ώστε να προωθηθεί στο επόμενο επίπεδο. Οι μαθηματικοί μετασχηματισμοί εντός ενός κελιού LSTM περιγράφονται στις Εξισώσεις 2.7 -2.12 περιγράφουν τους μαθηματικούς μετασχηματισμούς σε κάθε επίπεδο.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \qquad (Forget\ Gate) \tag{2.7}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \qquad (Input\ Gate) \tag{2.8}$$
$$C_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \qquad (Cell\ Input) \tag{2.9}$$
$$C_t = f_t * C_{t-1} + i_t * C_t \qquad (Cell\ State) \tag{2.10}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \qquad (Output\ Gate) \tag{2.11}$$
$$h_t = o_t * \tanh(C_t) \qquad (Hidden\ State) \tag{2.12}$$

Σε αντίθεση με τα κύτταρα μνήμης ενός ηλεκτρονικού υπολογιστή αυτές οι πύλες είναι αναλογικές και υλοποιούνται με πολλαπλασιασμούς (ανά στοιχείο) σιγμοειδών συναρτήσεων με εύρος $[0, 1]$. Το πλεονέκτημα των αναλογικών πυλών είναι η παραγωγισιμότητα τους, σε αντίθεση με τις ψηφιακές,

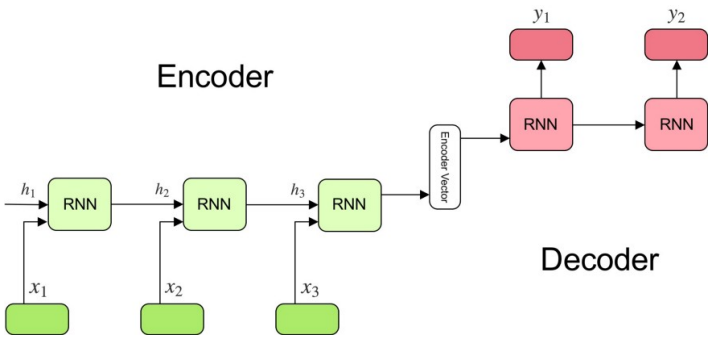
κάτι που τις καθιστά κατάλληλες για εφαρμογή του *αλγορίθμου οπίσθιας διάδοσης του σφάλματος* (back propagation) κατά τη διαδικασία εκπαίδευσης του δικτύου. Αυτές οι πύλες δρουν πάνω στα σήματα που δέχονται και παρόμοια με τους κόμβους του νευρωνικού δικτύου, είτε αποκλείουν είτε επιτρέπουν την διέλευση της πληροφορίας, αποφασίζοντας βασιζόμενες σε δικές τους συστοιχίες βαρών. Τα βάρη αυτά, όπως ακριβώς και τα άλλα βάρη του νευρωνικού δικτύου, αναπροσαρμόζονται από τη μέθοδο μάθησης. Με πιο απλά λόγια, η πύλη «μαντεύει» ποια είσοδο θα «κόψει» ή θα «αφήσει» να περάσει μέσω των βαρών και έπειτα ο αλγόριθμος μάθησης θα αναπροσαρμόσει τα βάρη, ώστε το σφάλμα να μειωθεί.

2.2.2 Αμφίδρομα LSTM

Τα αμφίδρομα LSTM (bi-directional LSTM ή biLSTM) είναι μία ειδική κατηγορία αναδρομικών νευρωνικών δικτύων που συνδυάζει δύο ανεξάρτητες μονάδες LSTMs ώστε η πληροφορία να τροφοδοτείται στο δίκτυο από την είσοδο προς την έξοδο αλλά και αντίστροφα. Η πληροφορία του κρυφού επιπέδου σε κάθε βήμα περιέχει γνώση, τόσο από τις προηγούμενες χρονικές στιγμές, όσο και από τις επόμενες, προσδίδοντας καλύτερη κατανόηση του περιεχομένου της ακολουθίας.

2.3 Το μοντέλο ακολουθίας-σε-ακολουθία

Το μοντέλο ακολουθίας-σε-ακολουθία (sequence-to-sequence ή seq2seq) αποτελείται από τρεις βασικές μονάδες: (i) τον *κωδικοποιητή* (encoder), (ii) ένα ενδιάμεσο στάδιο κωδικοποίησης (hidden vector ή thought vector) και (iii) τον *αποκωδικοποιητή* (decoder). Ο κωδικοποιητής αποτελείται από μονάδες LSTM ή GRU, οι οποίες διαβάζουν σειριακά την ακολουθία των λέξεων της εισόδου όπως φαίνεται στο Σχήμα 2.8. Κάθε μονάδα «μαθαίνει» ένα μέρος της πληροφορίας ώστε διαδίδοντας το στις επόμενες, να παραχθεί τελικά ένα διάνυσμα αναπαράστασης που θα περιέχει κωδικοποιημένο το περιεχόμενο ολόκληρης της ακολουθίας εισόδου. Σε κάθε βήμα t, το διάνυσμα του ενδιάμεσου σταδίου h ανανεώνεται σύμφωνα με την είσοδο-λέξη στο βήμα $X[i]$. Όταν όλες οι λέξεις που απαρτίζουν την ακολουθία διαβαστούν, το τελευταίο ενδιάμεσο στάδιο του κωδικοποιητή περιλαμβάνει ολόκληρο το περιεχόμενο της ακολουθίας, κωδικοποιημένο σε μια κρυφή αναπαράσταση.

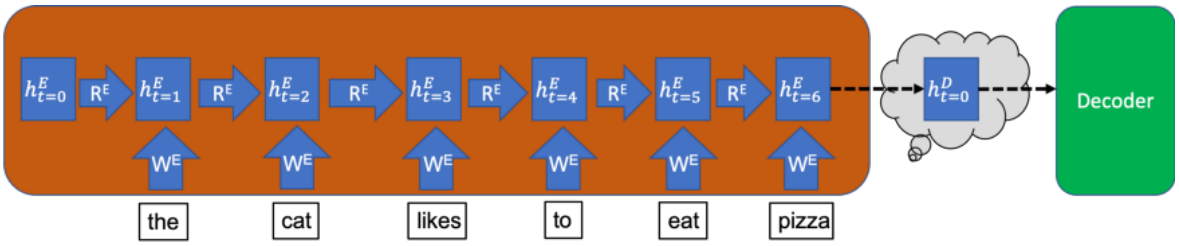


Σχήμα 2.8: Το μοντέλο ακολουθίας-σε-ακολουθία (Πηγή [Kost19])

Ας θεωρήσουμε ότι η είσοδος του κωδικοποιητή είναι η πρόταση `The cat likes to eat pizza` που απεικονίζεται στο Σχήμα 2.9. Η πρόταση αποτελείται από έξι λέξεις, άρα θα έχουμε έξι βήματα σε καθένα από τα οποία η κρυφή αναπαράσταση h θα ανανεώνεται λαμβάνοντας υπόψη την

προηγούμενη κρυφή αναπαράσταση (κωδικοποίηση του περιεχομένου όσων λέξεων έχει επεξεργαστεί μέχρι αυτή τη χρονική στιγμή) και την τρέχουσα είσοδο-λέξη. Τα μπλε βέλη αντιστοιχούν στα βάρη του δικτύου που ανανεώνονται κατά τη διαδικασία της εκπαίδευσης ώστε να επιτευχθεί μια όσο το δυνατόν πιο ακριβής μετάφραση.

Στο πρώτο βήμα t_1, το ενδιάμεσο στάδιο h_1 αρχικοποιείται σε μια τυχαία τιμή. Κάθε επίπεδο έχει δύο εξόδους: (i) το ενδιάμεσο στάδιο και (ii) την έξοδο για το επόμενο επίπεδο. Οι έξοδοι των αναδρομικών δικτύων αγνοούνται, με αποτέλεσμα ώστε σε κάθε βήμα να διαδίδεται μόνο η κρυφή αναπαράσταση, «αυτά που το δίκτυο έχει μάθει ως τώρα» δηλαδή, στο επόμενο επίπεδο.



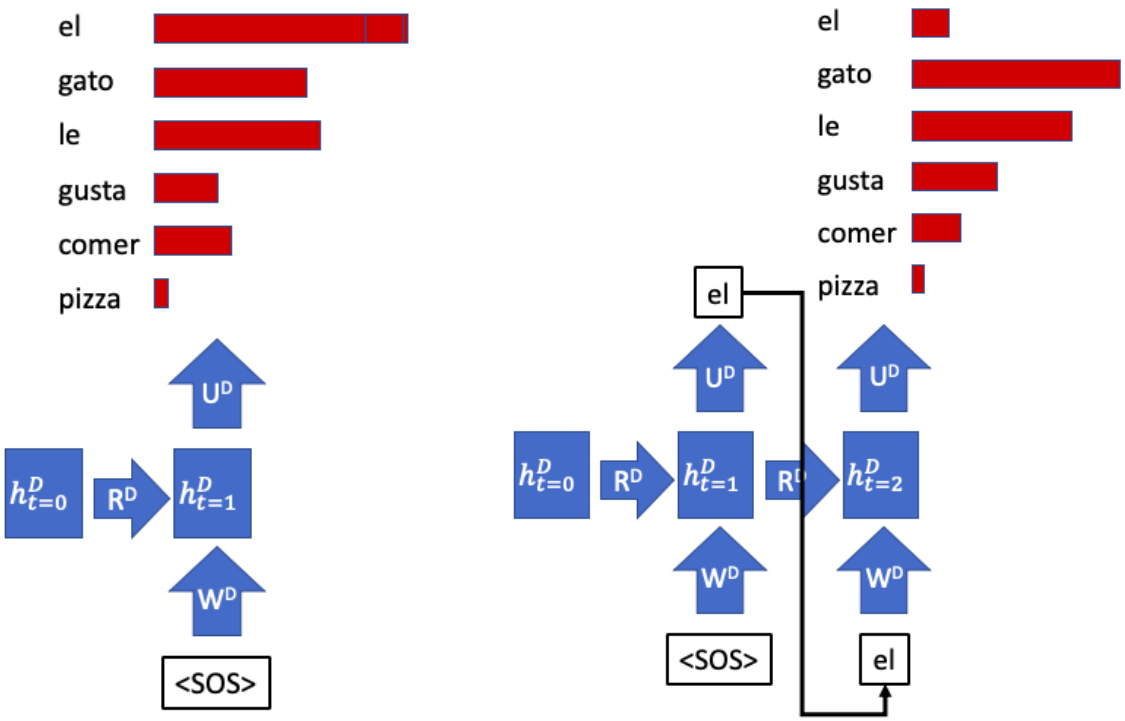
Σχήμα 2.9: Κωδικοποίηση της πρότασης «The cat likes to eat pizza» (Πηγή [Lann19])

Η έξοδος του κωδικοποιητή είναι το ενδιάμεσο στάδιο της τελευταίας μονάδας RNN και αποτελεί την τελική αναπαράσταση της κωδικοποιημένης εισόδου (που στόχος είναι να συγκεντρώσει την πληροφορία από όλη την ακολουθία). Κάθε μονάδα δέχεται ως είσοδο την κρυφή αναπαράσταση (ενδιάμεσο στάδιο) της προηγούμενης της και έχει ως έξοδο μια λέξη-πρόβλεψη και μία κρυφή αναπαράσταση. Η έξοδος υπολογίζεται λαμβάνοντας υπόψη την κρυφή αναπαράσταση στο τρέχον βήμα καθώς και τα αντίστοιχα βάρη $W(S)$. Από την κατανομή πιθανότητας των πιθανών λέξεων εξόδου, με χρήση της softmax (Ενότητα 2.1.1), επιλέγουμε τη λέξη με τη μεγαλύτερη πιθανότητα. Για να ξεκινήσει η μετάφραση, προστίθεται στην είσοδο του αποκωδικοποιητή στο πρώτο βήμα, η συμβολοσειρά `<SOS>`, ή αλλιώς start of sentence.

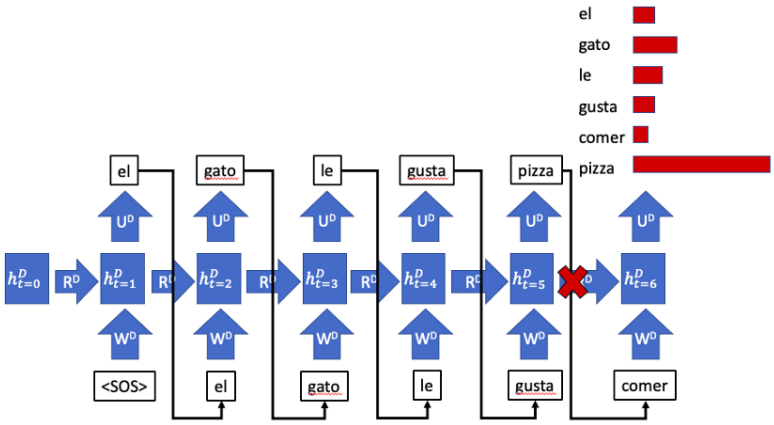
Τη χρονική στιγμή $t = 0$ ο αποκωδικοποιητής θα ανανεώσει την κρυφή του κατάσταση και με τα βάρη του δικτύου θα «επιλέξει» ως έξοδο τη λέξη με τη μεγαλύτερη πιθανότητα από το διαθέσιμο λεξιλόγιο που έχει δημιουργηθεί. Στο παράδειγμά μας, όπως δίνεται στο Σχήμα 2.10, η λέξη με τη μεγαλύτερη πιθανότητα είναι η `el` και άρα θα αποτελέσει το πρώτο στοιχείο της μεταφρασμένης πρότασης. Το δίκτυο αντιστοιχίζει βάρη σε όλες τις λέξεις του λεξιλογίου, ωστόσο εδώ απεικονίζονται οι λέξεις της συγκεκριμένης πρότασης με τις αντίστοιχες πιθανότητές τους για απλότητα. Στη συνέχεια, η έξοδος-πρόβλεψη εισάγεται στο επόμενο επίπεδο για να συνεχιστεί η μετάφραση. Όπως φαίνεται και στο Σχήμα 2.11, ο αποκωδικοποιητής προβλέπει λανθασμένα τη λέξη `pizza` αντί της λέξης `comer`. Στο σημείο αυτό η λανθασμένη πρόβλεψη του μοντέλου θα διαδοθεί στα επόμενα επίπεδα, αλλοιώνοντας το αποτέλεσμα της μετάφρασης. Η διαστρέβλωση της μετάφρασης μπορεί να αποφευχθεί κατά τη διαδικασία της εκπαίδευσης χρησιμοποιώντας την τεχνική *teacher forcing*, η οποία αναλύεται στην αμέσως επόμενη ενότητα.

2.3.1 Teacher Forcing

Η μέθοδος αυτή επινοήθηκε με σκοπό να επιδιορθώνει τα λάθη του αποκωδικοποιητή κατά τη διαδικασία της μετάφρασης. Όπως προαναφέρθηκε, όταν η έξοδος μιας αναδρομικής μονάδας είναι λανθασμένη σε κάποιο βήμα, η επόμενη μονάδα που τροφοδοτείται από αυτή θα «μάθει» από τη λαν-



Σχήμα 2.10: Σχηματική αναπαράσταση της διαδικασίας της αποκωδικοποίησης (Πηγή [Lann19])

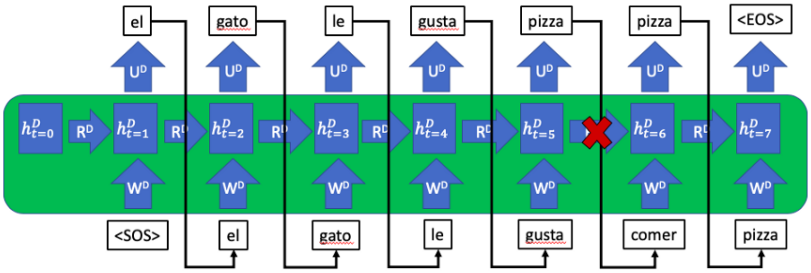


Σχήμα 2.11: Ο αποκωδικοποιητής προβλέπει λανθασμένα τη λέξη `pizza` (Πηγή [Lann19])

θασμένη λέξη, με αποτέλεσμα να αλλοιώνεται η διαδικασία της εκπαίδευσης και κατά συνέπεια, της μετάφρασης. Έτσι, προκειμένου να βελτιώσουμε τη διαδικασία της εκμάθησης του δικτύου, τροφοδοτούμε στην επόμενη μονάδα τη σωστή λέξη, διορθώνοντας την πρόβλεψη όπως φαίνεται στο Σχήμα 2.12. Η διαδικασία αυτή θυμίζει το «δάσκαλο που διορθώνει το μαθητή», όταν αυτός εκτίθεται σε ένα άγνωστο αντικείμενο που καλείται να μάθει. Η μέθοδος αυτή εφαρμόζεται μόνο κατά τη διαδικασία της εκπαίδευσης και όχι κατά τη διαδικασία του ελέγχου.

2.3.2 Μηχανισμός Προσοχής

Το βασικότερο πρόβλημα στην ακολουθιακή μάθηση με χρήση RNN, όπως είδαμε, είναι η συσχέτιση των δεδομένων μεγάλου μήκους ακολουθιών. Η χρήση των LSTMs βελτίωσε την απόδοση των απλών RNNs χωρίς όμως να δώσει πλήρη λύση στο πρόβλημα των εκλειπόμενων και εκρυγνυόμενων



Σχήμα 2.12: Η μέθοδος teacher forcing (Πηγή [Lann19])

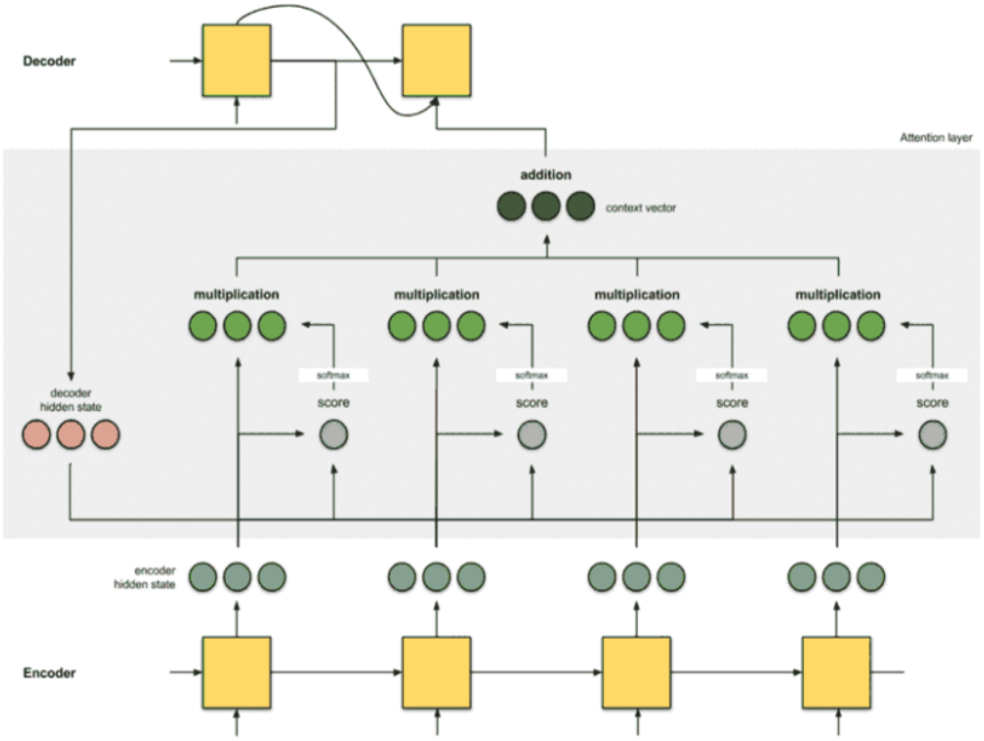
κλίσεων. Στην παρούσα εργασία θα μελετήσουμε την χρήση επιπέδων προσοχής (attention layers), στο πρόβλημα της μηχανικής μετάφρασης με χρήση seq2seq μοντέλων.

Στο μοντέλο seq2seq που περιγράφεται στην Ενότητα 2.3, ο κωδικοποιητής επεξεργάζεται την ακολουθία εισόδου και την ενσωματώνει-κωδικοποιεί σε ένα διάνυσμα (context vector). Ο αποκωδικοποιητής, με αυτό τον τρόπο, δέχεται ως είσοδο μόνο το τελευταίο κρυφό επίπεδο. Το διάνυσμα αυτό αποτελεί την αναπαράσταση ολόκληρης της ακολουθίας εισόδου. Για να κατανοήσουμε καλύτερα το πρόβλημα, ας φανταστούμε ένα κείμενο εκατό λέξεων, το οποίο καλούμαστε μονομιάς να μεταφράσουμε από τα Ελληνικά στα Αγγλικά. Η διαδικασία αυτή είναι αδύνατο να υλοποιηθεί χωρίς να «ξεχάσουμε» κάποιες λέξεις, ιδιαίτερα εκείνες που εμφανίζονται στην αρχή του κειμένου. Το ίδιο συμβαίνει και με την μετάφραση της πρότασης στον αποκωδικοποιητή. Η λύση έγκειται στο να ανακαλύπτουμε «ευθυγραμμίσεις» (alignments) μεταξύ των λέξεων του αρχικού κειμένου και του μεταφρασμένου, ώστε όλες οι λέξεις να καθορίζουν το αποτέλεσμα της μετάφρασης. Στο επίπεδο προσοχής, κάθε λέξη επισημαίνεται με ένα σκορ (γνωστό ως *alignment score function*) και όλα τα σκορ της ακολουθίας περνούν από μια softmax συνάρτηση (Ενότητα 2.1.1), ώστε να αθροίζουν στη μονάδα. Στη συνέχεια, πολλαπλασιάζοντας όλα τα κρυφά επίπεδα με την έξοδο της softmax έχουμε ένα διάνυσμα για κάθε λέξη που ενσωματώνει τόσο το περιεχόμενο όσο και τη σημαντικότητα της λέξης για κάθε βήμα της μετάφρασης (alignment vector). Τέλος, αθροίζοντας όλα τα διανύσματα αυτά, δημιουργούμε την τελική αναπαράσταση της εισόδου, που θα τροφοδοτηθεί στον αποκωδικοποιητή. Η διαδικασία αυτή απεικονίζεται στο Σχήμα 2.13

2.4 Ο αλγόριθμος ακτινικής αναζήτησης

Όπως παρουσιάστηκε στην Ενότητα 2.3, το seq2seq μοντέλο χρησιμοποιεί έναν κωδικοποιητή και έναν αποκωδικοποιητή, αποτελούμενο από μονάδες LSTM ή GRU. Στόχος είναι να προβλέψουμε την αντιστοίχηση των λέξεων από τη μια γλώσσα στην άλλη, διατηρώντας πλήρως το περιεχόμενο της αρχικής. Ένας τρόπος, όπως είδαμε, για να το πετύχουμε αυτό είναι να επιλέγουμε κάθε φορά, δεδομένης της εισόδου, τη λέξη με τη μεγαλύτερη πιθανότητα από το διαθέσιμο λεξικό της γλώσσας-στόχου, κάνοντας δηλαδή μια *άπληστη* (greedy) επιλογή. Το ερώτημα που προκύπτει στην περίπτωση αυτή είναι αν υπάρχει μία μόνο μετάφραση σε κάθε βήμα, ή αν μπορούμε διαλέγοντας διαφορετικές λέξεις και συνδυασμούς αυτών στα επόμενα βήματα, να έχουμε ένα νοηματικά και συντακτικά καλύτερο αποτέλεσμα.

Ο αλγόριθμος *ακτινικής αναζήτησης* (beam search) [Medr77] κατασκευάζει διαφορετικά μεταξύ τους μονοπάτια για μια ακολουθία εισόδου, επιλέγοντας σε κάθε βήμα t μια λέξη βάσει της δεσμευ-



Σχήμα 2.13: Μηχανισμός Προσοχής στο seq2seq μοντέλο (Πηγή [Kari19])

μένης πιθανότητας να έχουμε ως πρόβλεψη τη λέξη αυτή, δεδομένης της εισόδου και της πρόβλεψης στο βήμα $t-1$. Ο αριθμός των εναλλακτικών μονοπατιών καθορίζεται από μια υπερπαράμετρο k που ονομάζεται *πλάτος αναζήτησης* (beam width).

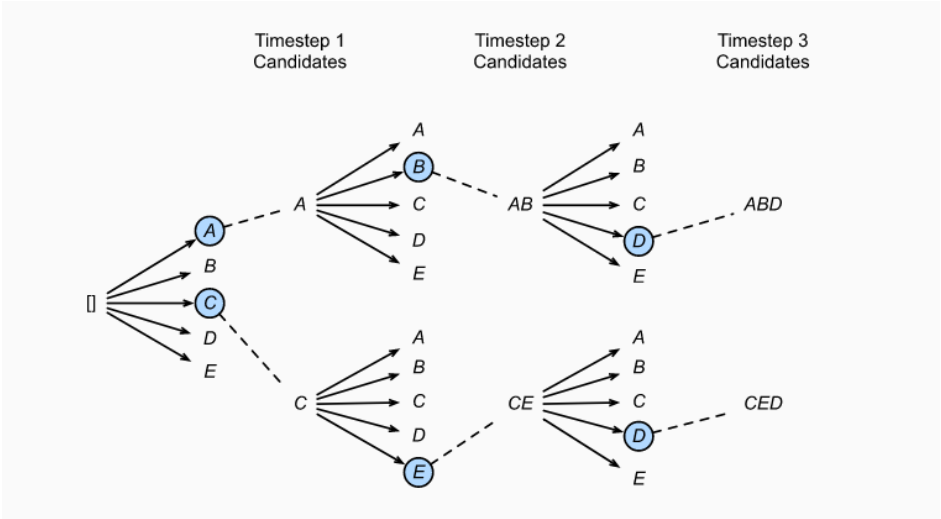
Με τη βοήθεια του Σχήματος 2.14, θα περιγραφεί η λειτουργία του αλγορίθμου ακτινικής αναζήτησης. Σε κάθε βήμα, ο αλγόριθμος επιλέγει k συνδυασμούς ($k = 2$ στο συγκεκριμένο παράδειγμα). Ας υποθέσουμε ότι έχουμε ένα λεξικό 5 λέξεων, το $Y = \{A, B, C, D, E\}$, όπου το σύμβολο $< eos >$ αντιστοιχεί στο D. Τότε, τα βήματα του αλγορίθμου έχουν ως ακολούθως:

- **Βήμα 1**: Επιλέγουμε από το λεξικό των 5 λέξεων τις $k = 2$ λέξεις με τη μεγαλύτερη πιθανότητα $\Pr(y_1 \mid c)$ Οι πιθανότητες έχουν προκύψει από τη softmax. Έστω ότι οι λέξεις αυτές είναι οι A, C.

- **Βήμα 2** : Για τις δυο επιλεγμένες λέξεις, υπολογίζουμε τις πιθανότητες $Pr(A, y_2 \mid x) = \Pr(A \mid x) \Pr(y_2 \mid A, x)$ και $\Pr(C, y_2 \mid x) = \Pr(C \mid x) \Pr(y_2 \mid C, x), \quad \forall y_2 \in Y \backslash \{A, C\}$. Από τις υπολογισμένες πιθανότητες διαλέγουμε τις $k = 2$ μεγαλύτερες, έστω $\Pr(A, B \mid x)$ και $\Pr(C, E \mid x)$

- **Βήμα 3**: Υπολογίζουμε τις πιθανότητες $\Pr(A, B, y_3 \mid x) = \Pr(A, B \mid x) \Pr(y_3 \mid A, B, x)$ και $\Pr(C, E, y_3 \mid x) = \Pr(C, E \mid x) \Pr(y_3 \mid C, E, x), \forall \forall y_3 \in Y$ και επιλέξουμε πάλι τις $k = 2$ μεγαλύτερες, έστω $\Pr(A, B, D \mid x)$ και $\Pr(C, E, D \mid x)$.

Ο αλγόριθμος επαναλαμβάνεται συνήθως μέχρι ένα προκαθορισμένο βήμα T, ή έως ότου έχουμε n ολοκληρωμένες προτάσεις. Ως ολοκληρωμένη ορίζεται μια ακολουθία όταν βρεθεί το σύμβολο $< eos >$. Οι πιθανές ακολουθίες εδώ είναι: (i) A, (ii) C, (iii) A, B, (iv) C, E, (v) A, B, D και



Σχήμα 2.14: Ο αλγόριθμος ακτινική αναζήτησης για $k = 2$ μέγεθος εξόδου 3 (Πηγή [Zhan20a])

(vi) C, E, D. Στις υποψήφιες αυτές ακολουθίες θα υπολογίσουμε την πιθανότητα ως εξής (Εξίσωση 2.13)

$$score(y_1, ...y_t) = \log \Pr(y_1, ..., y_t \mid x) = \sum_{i=1}^{t} \log \Pr(y_i \mid y_1, ..., y_{i-1}, x) \tag{2.13}$$

Η Εξίσωση 2.13 ωστόσο, μπορεί να κανονικοποιηθεί χρησιμοποιώντας το μήκος της ακολουθίας ώστε να αποφευχθεί η εμφάνιση μικρότερων scores για ακολουθίες μεγάλου μήκους (Εξίσωση 2.14)

$$score(y_1, ...y_t) = \frac{1}{t} \sum_{i=1}^{t} \log \Pr(y_i \mid y_1, ..., y_{i-1}, x) \tag{2.14}$$

Κεφάλαιο 3

Επεξεργασία Κειμένου

Στην παρόν Κεφάλαιο θα εξετάσουμε την επεξεργασία δεδομένων κειμένου γραμμένα σε φυσική γλώσσα και θα αναλύσουμε μεθόδους για την όσο το δυνατόν καλύτερη ψηφιακή αναπαράστασή τους. Τα δεδομένα κειμένου μπορούν να προέρχονται από πολλές πηγές, όπως άρθρα, ιστολόγια ακόμα και tweets. Το κοινό χαρακτηριστικό τους είναι η έλλειψη δομής (unstructured data), πράγμα το οποίο σημαίνει ότι δεν μπορούν να τροφοδοτηθούν σε αλγορίθμους μηχανικής μάθησης ως έχουν. Συνεπώς, εδώ θα εξεταστούν τεχνικές αναπαράστασης των κειμένων σε κατάλληλη (αριθμητική) μορφή, η οποία μπορεί να χρησιμοποιηθεί από τα συστήματα που θα εξετάσουμε στη συνέχεια.

3.1 Σύνολα από λέξεις

Το μοντέλο *συνόλων από λέξεις* (bag of words - BOW) [Zhan10] είναι ένας τρόπος αναπαράστασης κειμένων, ο οποίος έχει χρησιμοποιηθεί με επιτυχία σε πολλά προβλήματα επεξεργασίας φυσικής γλώσσας, όπως η κατηγοριοποίηση κειμένων. Ένα μοντέλο συνόλου από λέξεις είναι ένας τρόπος να εξάγουμε χαρακτηριστικά σε σταθερού μήκους διανύσματα από ένα κείμενο, με σκοπό τη χρήση τους σε αλγορίθμους μηχανικής μάθησης. Στη βάση της πρόκειται για μία αναπαράσταση που καταγράφει την συχνότητα εμφάνισης μιας λέξης σε ένα κείμενο. Για το λόγο αυτό απαιτείται ένα δεδομένο λεξιλόγιο και ένας τρόπος να μετριούνται τα στιγμιότυπα μιας γνωστής λέξης στο κείμενο. Η μέθοδος αυτή εστιάζει στο πλήθος των εμφανίσεων μιας λέξης σε ένα κείμενο χωρίς να λαμβάνει υπόψη τη δομή και την αλληλουχία των λέξεων σε αυτό, για παράδειγμα το σημείο στο οποίο βρίσκεται η λέξη. Η ιδέα πίσω από αυτόν τον αλγόριθμο είναι ότι παρόμοια κείμενα έχουν μεγάλο αριθμό κοινών λέξεων και ότι αυτές οι λέξεις καθορίζουν το νόημα του κειμένου.

Παρακάτω δίνεται ένα παράδειγμα με τα βήματα που ακολουθεί ο αλγόριθμος για να μετατρέψει δεδομένα κειμένου σε διανύσματα.

- **Βήμα** 1: Συλλογή Δεδομένων

  Έστω έχουμε μια συλλογή κειμένων, που αποτελείται από τις ακόλουθες 3 προτάσεις: (i) «Ήταν οι καλύτερες των εποχών», (ii) «Ήταν οι χειρότερες των εποχών», (iii) «Ήταν η στιγμή της σοφίας».

- **Βήμα** 2: Σχεδιασμός του λεξιλογίου

  Δημιουργείται μια λίστα με τις μοναδικές λέξεις που περιέχει το σύνολο των κειμένων, η οποία αποτελεί ένα λεξιλόγιο 10 λέξεων από μία συλλογή κειμένων που περιέχει 15 λέξεις.

- **Βήμα** 3: Δημιουργία Διανυσμάτων (vectors)

Ο στόχος αυτού του βήματος είναι η μετατροπή του κειμένου σε ένα διάνυσμα το οποίο θα δοθεί ως είσοδος ή έξοδος σε ένα μοντέλο μηχανικής μάθησης. Στο παράδειγμά μας, αφού το μέγεθος του λεξικού είναι 10, το διάνυσμα αναπαράστασης θα έχει κι αυτό 10 θέσεις, με κάθε μια να αντιστοιχεί σε μια δυαδική τιμή (0 αν δεν υπάρχει η λέξη στο συγκεκριμένο κείμενο και 1 αν υπάρχει). Για το πρώτο κείμενο (πρώτη γραμμή) θα είναι:

«Ήταν» = 1
«οι» = 1
«καλύτερες» = 1
«των» = 1
«εποχών» = 1
«χειρότερες» = 0
«η» = 0
«στιγμή» = 0
«της» = 0
«σοφίας» = 0
Και ως διάνυσμα πλέον εκφράζεται με τον ακόλουθο τρόπο : $[1, 1, 1, 1, 1, 0, 0, 0, 0, 0]$.

Ομοίως εκφράζονται και τα υπόλοιπα κείμενα (γραμμές). Αν μια λέξη εμφανίζεται δύο φορές στο κείμενο, στην αντίστοιχη θέση του διανύσματος θα εμφανίζεται ο αριθμός δύο.

Συνεπώς, κάθε κείμενο θα αποτελείται από ένα διάνυσμα σταθερού μεγέθους ίσου με το πλήθος των λέξεων του λεξικού (στο παράδειγμα μας διάστασης 10) και θα αποτελείται από το πλήθος εμφάνισης κάθε λέξης στην αντίστοιχη θέση του διανύσματος.

3.1.1 Σύνολα n-grams

Ένα από τα κυριότερα προβλήματα της προηγούμενης μεθόδου παρουσιάζεται όταν η συλλογή των κειμένων είναι πολύ μεγάλη, με συνέπεια το λεξιλόγιο και κατ' επέκταση η διάσταση των διανυσμάτων να είναι τεράστια. Αυτό οδηγεί σε αραιούς πίνακες, που αποτελούνται από μηδενικά και ελάχιστες μονάδες. Η διαχείριση αυτών των διανυσμάτων είναι «ακριβή» με την έννοια ότι απαιτεί πόρους όπως υπολογιστική ισχύ και μνήμη. Μια από τις προσεγγίσεις επίλυσης αυτού του προβλήματος είναι η ομαδοποίηση των λέξεων σε φράσεις μεγέθους δυο, τριών ή περισσότερων (n) λέξεων (gram).

Έστω το 2-gram «Άνοιξέ μου». Θα υπολογίσουμε την πιθανότητα μιας λέξης δεδομένης της λέξης που εμφανίστηκε πριν από αυτή, ή στο παράδειγμα μας την πιθανότητα εμφάνισης της λέξης «μου» μετά την εμφάνιση της λέξης «άνοιξε».

$$\Pr(W_n \mid {W_1}^{n-1}) \approx \Pr(W_n \mid W_{n-1}) \tag{3.1}$$

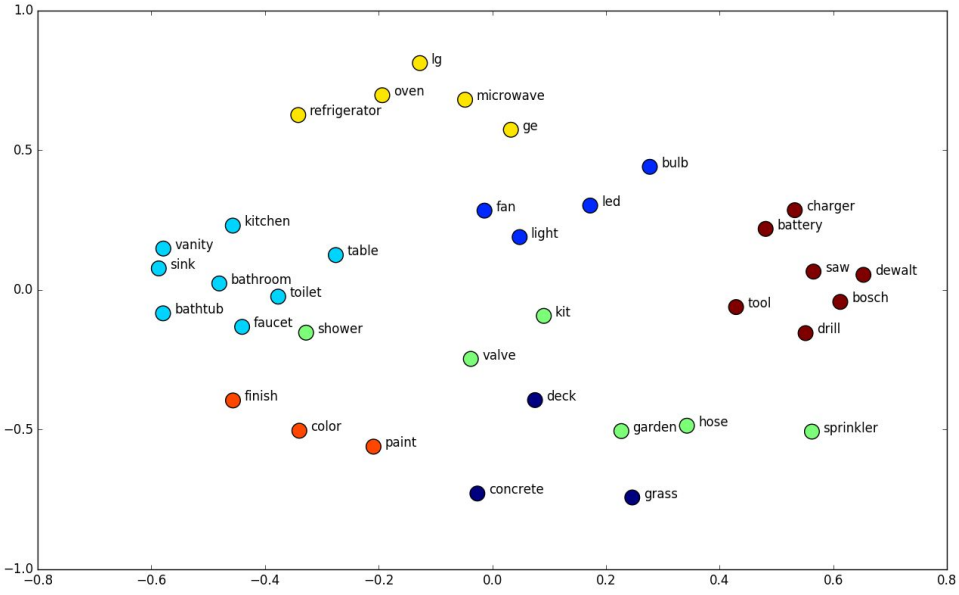
Η παραδοχή ότι μια λέξη εξαρτάται μόνο από την προηγούμενη της είναι γνωστή ως Μαρκοβιανή υπόθεση. Η Εξίσωση 3.1 γενικεύεται για 3-gram, 4-gram, κ.λ.π., λαμβάνοντας υπόψη τις n προηγούμενες λέξεις κάθε φορά. Η αδυναμία αυτού του αλγορίθμου έγκειται στο ότι εξαρτάται από την εμφάνιση των λέξεων στο συγκεκριμένο κείμενο, η οποία μπορεί να διαφοροποιείται ανάλογα το δείγμα. Επίσης, η πιθανότητα κάθε λέξης μεταβάλλεται για διαφορετικές τιμές του n.

3.1.2 Term Frequency-Inverse Term Frequency

Η τεχνική *term frequency-inverse document frequency* (TF-IDF) [Qais18] έχει ως στόχο να αντιμετωπίσει το πρόβλημα των λέξεων που εμφανίζονται σε πάρα πολλά κείμενα με μεγάλη συχνότητα, χωρίς να προσδίδει η παρουσία τους κάποια πληροφορία για το περιεχόμενο. Ο αλγόριθμος χωρίζεται σε δύο μέρη το *term frequency* (TF), όπου υπολογίζεται η συχνότητα της λέξης στο κείμενο και το *inverse document frequency* (IDF), όπου υπολογίζεται πόσο σπάνια εμφανίζεται η λέξη σε όλη τη συλλογή των κειμένων. Αυτοί οι υπολογισμοί διαμορφώνουν ένα βάρος για το κατά πόσο σημαντική είναι η κάθε λέξη.

3.1.3 Word embeddings

Η ψηφιακή αναπαράσταση των δεδομένων προϋποθέτει την μετατροπή τους σε αριθμούς ώστε να μπορούν να επεξεργαστούν από τα συστήματα μηχανικής μάθησης που περιγράφηκαν στο Κεφάλαιο 2 και γίνει εφικτή η εξαγωγή χαρακτηριστικών από αυτά. Ένα πολύ σημαντικό πρόβλημα στην αναπαράσταση κειμένων με μοντέλα n-grams, όπως είδαμε στην Ενότητα 3.1.1, είναι η αύξηση των διαστάσεων όταν πρόκειται για διακριτές μεταβλητές (λέξεις), αφού για τον υπολογισμό ενός 10-gram με λεξικό πλήθους 100.000 λέξεων έχουμε $10000^{10} - 1 = 10^{50} - 1$ ελεύθερες παραμέτρους. Η μοντελοποίηση με χρήση συνεχών τιμών μπορεί να περιορίσει το πρόβλημα αυτό και να οδηγήσει σε ευκολότερη γενίκευση [Beng03].



Σχήμα 3.1: Διανυσματική αναπαράσταση λέξεων (word embeddings). Λέξεις με παρόμοιο νοηματικά περιεχόμενο εμφανίζουν μικρότερη απόσταση μεταξύ τους (Πηγή [Ruge20]).

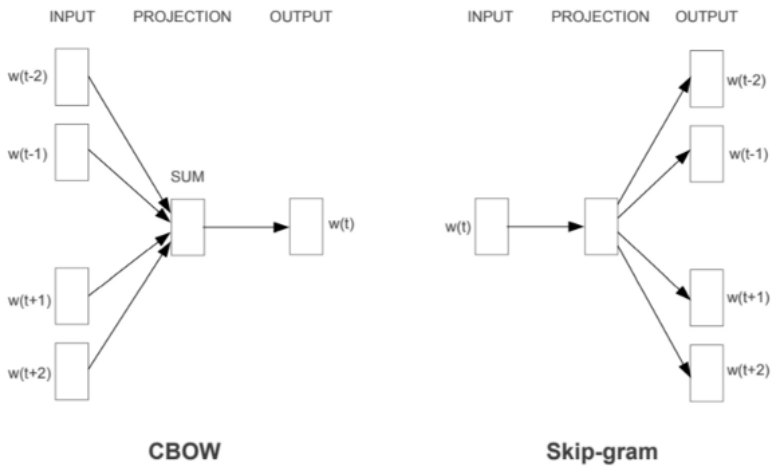
Τα *word embeddings* συνιστούν διανυσματικές αναπαραστάσεις των λέξεων και είναι ευρέως χρησιμοποιούμενα στον τομέα της επεξεργασίας φυσικής γλώσσας. Αποτελούνται από συνεχείς τιμές και το μέγεθος των διανυσμάτων αυτών ποικίλλει, ανάλογα την πολυπλοκότητα των δεδομένων $(100, 200, 300$και άλλα$)$. Συνήθως, ένα μικρότερο μέγεθος οδηγεί σε γρήγορη επεξεργασία και εκ-

παίδευση, ενώ έχει αποδειχθεί ότι μεγαλώνοντας τη διάσταση της αναπαράστασης επιτυγχάνουμε καλύτερα αποτελέσματα με μεγαλύτερο όμως κόστος.

Το σημαντικότερο χαρακτηριστικό των word embeddings είναι ότι παρόμοιες νοηματικά λέξεις εμφανίζουν μεγάλη διανυσματική ομοιότητα όπως φαίνεται στο Σχήμα 3.1 (που συνήθως μετριέται με μετρικές όπως η Ευκλείδεια απόσταση ή την ομοιότητα συνημιτόνου). Για παράδειγμα, οι λέξεις «μαμά» και «μπαμπάς» θα έχουν μικρότερη απόσταση από ότι οι λέξεις «μαμά» και «βούτυρο», διότι εμφανίζουν μεγαλύτερη νοηματική ομοιότητα. Τα διανύσματα αυτά δημιουργούνται από το διαθέσιμο προς επεξεργασία κείμενο με τις τεχνικές που πρόκειται να περιγραφούν αμέσως παρακάτω, ή συχνά λαμβάνονται προ-εκπαιδευμένα από μεγαλύτερα κείμενα για περισσότερη πληροφορία.

3.1.4 Ο αλγόριθμος Word2Vec

Ο αλγόριθμος Word2Vec αποτελεί τον πιο διαδεδομένο αλγόριθμο δημιουργίας embeddings. Όπως είδαμε προηγουμένως, η απλή αναπαράσταση των κειμένων με διανύσματα συχνότητας εμφάνισης λέξεων οδηγεί σε μεγάλη σπατάλη χώρου, καθώς για ένα λεξικό 10.000 λέξεων, η απεικόνιση κάθε κειμένου ένα διάνυσμα 10.000 στοιχείων, αποτελούμενο ως επί το πλείστον από μηδενικά. Η διαδικασία αυτή αυξάνει κατά πολύ την πολυπλοκότητα του αλγορίθμου ενώ ακόμα, η χρήση των διανυσμάτων συχνότητας δεν λαμβάνει υπόψη τη σημασιολογική ερμηνεία τον λέξεων. Το ίδιο συμβαίνει και με την αναπαράσταση με n-grams. Το πρόβλημα αυτό επιλύει η χρήση του αλγορίθμου Word2Vec.



Σχήμα 3.2: Ο αλγόριθμος Word2vec, με τις δύο διαφορετικές τεχνικές εκπαίδευσης (Πηγή [Miko13])

Το Word2Vec αποτελεί ένα νευρωνικό δίκτυο με δύο κρυφά επίπεδα το οποίο επεξεργάζεται ένα μεγάλο κείμενο για να εξάγει διανύσματα χαρακτηριστικών για τις λέξεις του κειμένου. Μια βασική διαφορά του Word2Vec με τους προγενέστερους αλγορίθμους αποτελεί η χρήση γραμμικών συναρτήσεων ενεργοποίησης στους νευρώνες του δικτύου. Λαμβάνοντας υπόψη τα *συμφραζόμενα* (context) και χωρίς ανθρώπινη παρέμβαση ο αλγόριθμος Word2Vec εκπαιδεύεται από τις γειτονικές σχέσεις λέξεων, υπολογίζοντας τη συχνότητα *συνεμφάνισης* (co-occurance) τους και δημιουργώντας αριθμητικές αναπαραστάσεις, τις οποίες προσθέτει σε ένα συμπαγές λεξιλογικό διάνυσμα. Το διάνυσμα αυτό περιέχει αριθμητικές τιμές με την μορφή πιθανοτήτων, με χρήση της συνάρτησης softmax, για την ομοιότητα και τη συσχέτιση μεταξύ των λέξεων Δύο τεχνικές χρησιμοποιούνται κυρίως για την εκπαίδευση των word2vec αναπαραστάσεων: (i) *continuous bag of words* (CBOW) και (ii) *skip-gram*

(Σχήμα 3.2). Ο αλγόριθμος CBOW λαμβάνοντας υπόψη το γειτονικό περιεχόμενο μιας λέξης, επιχειρεί να προβλέψει τη λέξη αυτή. Αντίθετα, στον αλγόριθμο skip-gram, δεδομένης μιας λέξης προσπαθούμε να προβλέψουμε την κατανομή των γειτονικών λέξεων που συνιστούν το περιεχόμενο της λέξης αυτής.

3.1.5 Μέθοδος Δημιουργίας Διανυσμάτων Λέξεων GloVe

Η μέθοδος Word2Vec είναι πολύ αποδοτική και σε ορισμένα προβλήματα έχει δείξει εντυπωσιακά αποτελέσματα. Παρόλα αυτά, ο υπολογισμός του διανύσματος λαμβάνει υπόψη συμφραζόμενα μόνο κοντά στη ζητούμενη λέξη. Η μέθοδος GloVe (global vectors) μπορεί να αντιμετωπίσει αυτό το πρόβλημα καθώς είναι σχεδιασμένη κατά τέτοιο τρόπο ώστε να μπορεί να υπολογίζει περισσότερες στατιστικές πληροφορίες για τις λέξεις, όπως την εμφάνιση λέξεων παραπάνω από μία φορές μέσα σε μια μεγάλη συλλογή κειμένων [Penn14]. Η βασική ιδέα για την λειτουργία αυτής της μεθόδου είναι η εξαγωγή σημασιολογικών σχέσεων από τον πίνακα συνεμφάνισης των λέξεων.

Πιο συγκεκριμένα, δεδομένης μιας συλλογής κειμένου η οποία διαθέτει συνολικά V λέξεις, ο *πίνακας συνεμφάνισης* (co-occurance matrix) θα είναι διαστάσεων $V * V$, όπου το στοιχείο X στην i γραμμή και στη j στήλη δηλώνει πόσες φορές η λέξη i εμφανίστηκε μαζί με τη λέξη j . Ένα παράδειγμα φαίνεται στο Σχήμα 3.3 όπου η πρόταση είναι η εξής «the cat sat on the mat».

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Σχήμα 3.3: Παράδειγμα πίνακα συνεμφάνισης (Πηγή [Penn14])

Το ζήτημα που ανακύπτει εδώ είναι πως προσεγγίζουμε τη μετρική που υπολογίζει τη σημασιολογική ομοιότητα. Αυτό γίνεται περισσότερο κατανοητό στο ακόλουθο παράδειγμα. Ας ορίσουμε τον

Probability and Ratio	k = *solid*	k = *gas*	k = *water*	k = *fashion*
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Σχήμα 3.4: Παράδειγμα που δείχνει τη μέθοδο εύρεσης της σημασιολογικής ομοιότητας (Πηγή [Penn14])

λόγο $\frac{P_{i,k}}{P_{j,k}}$ όπου $P_{i,k} = \frac{X_{i,k}}{X_i}$. Η $P_{i,k}$ δηλώνει την πιθανότητα να εμφανιστεί η λέξη i και k μαζί, και υπολογίζεται διαιρώντας τον αριθμό $X_{i,k}$ με τις συνολικές εμφανίσεις της λέξης i σε ολόκληρη τη συλλογή των κειμένων (X_i). Όπως φαίνεται από τον παραπάνω πίνακα (Σχήμα 3.4), δεδομένων δύο λέξεων π.χ ice και $steam$, αν η τρίτη λέξη k είναι: (i) αρκετά παρόμοια με τη λέξη ice αλλά άσχετη

με το $steam(k = solid)$, η αναλογία $\frac{P_{i,k}}{P_{j,k}}$ θα είναι πολύ υψηλή (> 1), (ii) αρκετά παρόμοια με το $steam$ αλλά άσχετη με το $ice(k = gas)$, η αναλογία $\frac{P_{i,k}}{P_{j,k}}$ θα είναι πολύ μικρή (<1), (iii) παρόμοια ή άσχετη και με τις δύο λέξεις, τότε $\frac{P_{i,k}}{P_{j,k}}$ θα προσεγγίζει τη μονάδα

Για να δημιουργήσουμε τα διανύσματα των λέξεων πρέπει να λύσουμε τρία προβλήματα. Πρώτον, δεν γνωρίζουμε την εξίσωση παρά μόνο μία έκφραση της μορφής $F(i, j, k) = \frac{P_{i,k}}{P_{j,k}}$. Δεύτερον, τα διανύσματα των λέξεων πρέπει να έχουν αρκετές διαστάσεις, ενώ η μετρική $\frac{P_{i,k}}{P_{j,k}}$ είναι βαθμωτό μέγεθος. Άρα, υπάρχει αναντιστοιχία διαστάσεων. Τρίτον, υπάρχουν τρεις οντότητες που εμπλέκονται (i, j, k) οι οποίες πρέπει να μειωθούν σε δύο για να μπορούν να είναι διαχειρίσιμες από τη συνάρτηση σφάλματος (loss function). Θα χρησιμοποιήσουμε την ακόλουθη σημειογραφία: (i) w, u δύο ξεχωριστά επίπεδα διανυσμάτων, (ii) $w^\top$ ανάστροφο διάνυσμα του w, (iii) X πίνακας συνεμφάνισης, (iv) b_w και b_u πόλωση του w και του u αντίστοιχα.

Για την επίλυση του προβλήματος απλά θα υποθέσουμε ότι υπάρχει μια συνάρτηση F, η οποία λαμβάνει τα διανύσματα των λέξεων i j και k. δίνοντας σαν έξοδο την αναλογία που ζητάμε $F(w_i, w_j, u_k) = \frac{P_{i,k}}{P_{j,k}}$. Γενικότερα, η δημιουργία δυο επιπέδων διανυσμάτων w και u βοηθάει το μοντέλο στο να μειώσει το πρόβλημα της *υπερεκπαίδευσης* (overfitting). Τα διανύσματα των λέξεων είναι γραμμικά συστήματα. Για παράδειγμα, μπορεί να πραγματοποιηθεί η ακόλουθη πράξη w_{king}–$w_{male} + w_{female} = w_{queen}$. Γι' αυτό, θα αλλάξουμε τη μορφή της συνάρτησης ως εξής: $F(w_i$–$w_j, u_k) = \frac{P_{i,k}}{P_{j,k}}$. Αυτό το κάνουμε επειδή στις περισσότερες περιπτώσεις τα διανύσματα λέξεων χρησιμοποιούνται για τον υπολογισμό αποστάσεων .

Για την επίλυση του δεύτερου προβλήματος (αναντιστοιχία διαστάσεων) θα εισάγουμε την μαθηματική μέθοδο της αναστροφής ενός πίνακα αλλά και το γινόμενο πινάκων. $F(w_i$–$w_j) * .u_k) = \frac{P_{i,k}}{P_{j,k}}$. Πιο αναλυτικά, αν υποθέσουμε ότι το διάνυσμα λέξης είναι διαστάσεων $D*1$ τότε ο πίνακας $(w_i$–$w_j)*$ θα είναι $1 \times D$ και το γινόμενο με το u_k θα δώσει βαθμωτό μέγεθος. Αν υποθέσουμε τώρα ότι η συνάρτηση F έχει μια συγκεκριμένη ιδιότητα, όπως ο *ομομορφισμός* (homomorphism) μεταξύ πρόσθεσης και πολλαπλασιασμού, τότε η εξίσωση γίνεται $F(w_i * u_k - w_j * u_k) = \frac{F(w_i * u_k)}{F(w_j * u_k)} = \frac{P_{i,k}}{P_{j,k}}$ και έτσι προκύπτει η σχέση $F(w_i * u_k) = P_{i,k}$. Θέτοντας την συνάρτηση F ίση με την εκθετική, έχουμε $e^{w_i * u_k} = P_i k = \frac{X_{i,k}}{X_i}$ και άρα $w_i * u_k = \log(X_{i,k}) - \log(X_i)$. Προσθέτοντας τις πολώσεις b_w, b_u από το νευρωνικό δίκτυο έχουμε την τελική μορφή της εξίσωσης. Σε ιδανική περίπτωση τέλειων διανυσμάτων λέξεων, η εξίσωση αυτή θα ήταν μηδέν. Τώρα ο στόχος είναι η ελαχιστοποίηση αυτής μέσω της συνάρτησης κόστους $J(w_i, w_j) = (w_i * u_j + bw_i + bu_j$—$log(X_i j))^2$.

Κεφάλαιο 4

Αρχιτεκτονικές Μοντέλων Μετάφρασης

Σε αυτό το Κεφάλαιο θα παρουσιαστούν με τη σειρά οι αρχιτεκτονικές των μοντέλων μηχανικής μάθησης που εξετάστηκαν στο πλαίσιο της διπλωματικής εργασίας, δηλαδή για την μηχανική μετάφραση κειμένων αγγλικής γλώσσας στα ελληνικά. Τα δομικά στοιχεία των παρουσιαζόμενων αρχιτεκτονικών εξετάστηκαν στο Κεφάλαιο 2, ενώ τα αποτελέσματα των πειραματικών δοκιμών θα αναλυθούν διεξοδικότερα στο Κεφάλαιο 5.

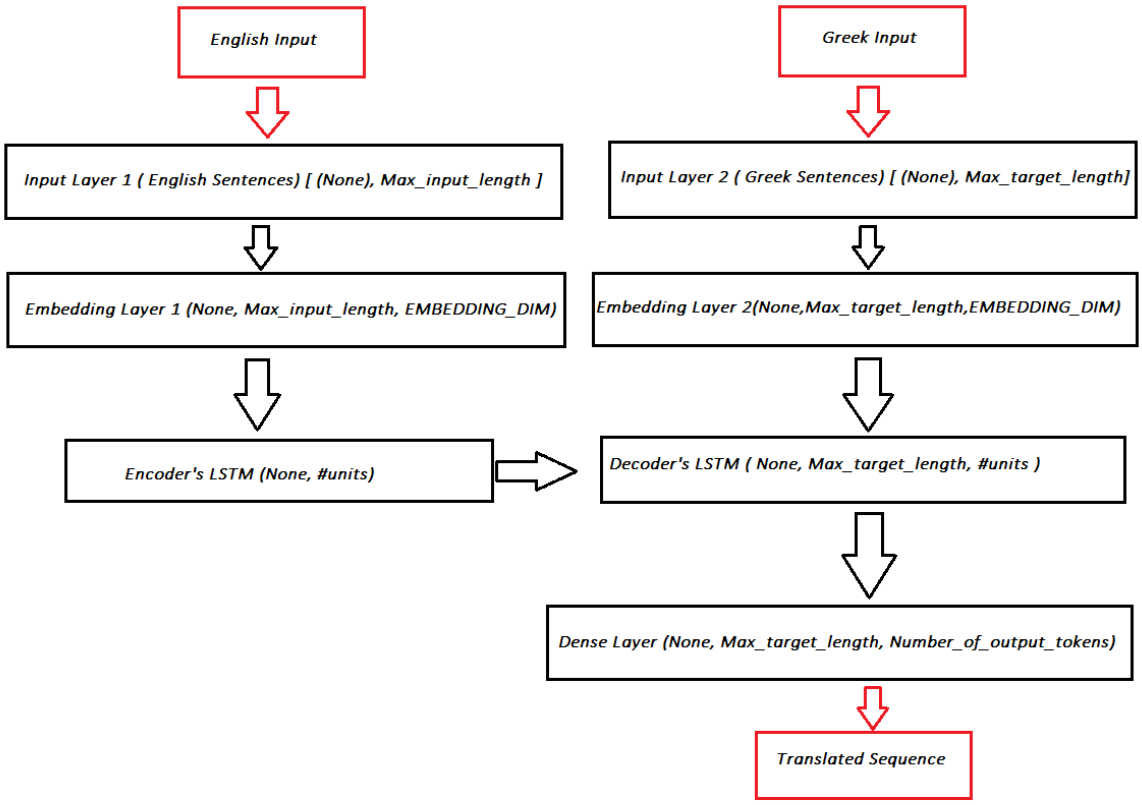
4.1 Μοντέλο seq2seq

Το πρώτο μοντέλο που εξετάστηκε αφορά την seq2seq αρχιτεκτονική (Ενότητα 2.3), η οποία είναι μία από τις πρώτες αρχιτεκτονικές που κατασκευάστηκαν για την επίλυση του προβλήματος της μηχανικής μετάφρασης. Η αρχιτεκτονική αυτή αποτελείται από δύο επίπεδα embedding, ένα για κάθε LSTM του κωδικοποιητή και του αποκωδικαοποιητή. Στη συγκεκριμένη περίπτωση χρησιμοποιήθηκαν GloVe embeddings για την απεικόνιση των λέξεων της αγγλικής γλώσας και Word2Vec embeddings για την ελληνική γλώσσα. Στην αρχιτεκτονική αυτή, επίσης, καλούμαστε να υλοποιήσουμε δύο διαφορετικούς αποκωδικοποιητές λόγω περιορισμών στην υλοποίηση, ένα για τη διαδικασία της εκπαίδευσης και ένα για τη διαδικασία του ελέγχου. Η όλη αρχιτεκτονική αναπαρίσταται στο Σχήμα 4.1.

Μετά από εκτεταμένες δοκιμές στις υπερ-παραμέτρους του μοντέλου, αυτές που επέφεραν τα καλύτερα αποτελέσματα συνοψίζονται στον Πίνακα 4.1.

Πίνακας 4.1: Υπερ-παράμετροι βέλτιστου seq2seq μοντέλου

Υπερ-παράμετρος	**Τιμή**
LSTM επίπεδα κωδικοποιητή & αποκωδικοποιητή	256
Μήκος embeddings (διανυσμάτων)	100
Βελτιστοποιητής	Adam, με ρυθμό μάθησης $0,01$
Συνάρτηση απώλειας	Κατηγορική διασταυρούμενη εντροπία
Ομαλοποίηση Βαρών	L2 με $\lambda = 0,001$
Ομαλοποίηση Πολώσεων	L2 με $\lambda = 0,0001$
Χρήστη Dropout	5% σε όλα τα κρυφά επίπεδα
Μέγεθος δέσμης	64
Εποχές εκπαίδευσης	12



Σχήμα 4.1: Σχηματική Αναπαράσταση του Sequence-to-Sequence μοντέλου

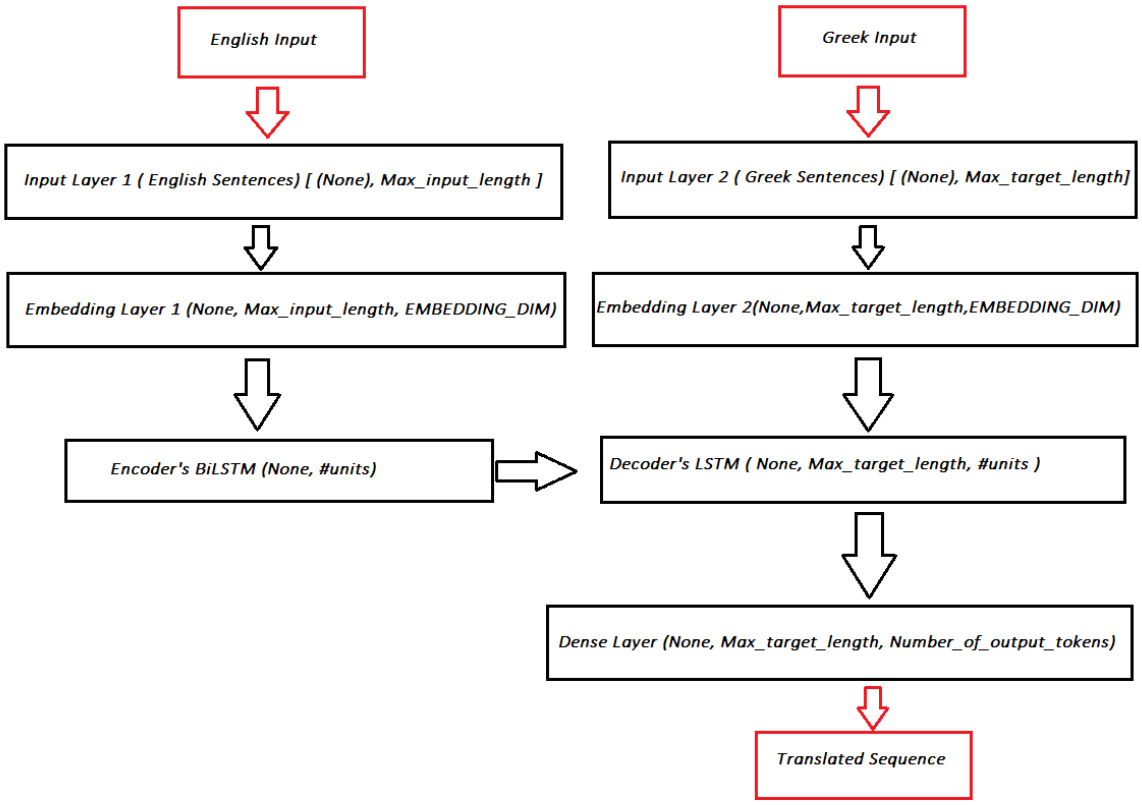
4.2 Μοντέλο seq2seq με biLSTM στον κωδικοποιητή

Για να βελτιώσουμε τα αποτελέσματα της προηγούμενης αρχιτεκτονικής, θα επιχειρήσουμε να τροποποιήσουμε την κωδικοποίηση, χρησιμοποιώντας ένα biLSTM (Ενότητα 2.2.2). Το biLSTM είναι συνδυασμός δύο LSTMs όπου το ένα ξεκινά να επεξεργάζεται προγενέστερα τμήματα της ακολουθίας και το άλλο μεταγενέστερα, διασχίζουν δηλαδή την ακολουθία ταυτόχρονα με αντίθετη φορά. Έτσι, στην προηγούμενη αρχιτεκτονική θα προσθέσουμε στον κωδικοποιητή ένα αμφίδρομο LSTM αντί του απλού. Στόχος στην αρχιτεκτονική αυτή είναι το περιεχόμενο της πρότασης να τροφοδοτείται στον κωδικοποιητή τόσο από την αρχή όσο και από το τέλος της ακολουθίας, ώστε να λαμβάνονται υπόψη εξαρτήσεις από λέξεις που εμφανίζονται σε μεταγενέστερα χρονικά βήματα. Η αναπαράσταση του μοντέλου δίνεται στο Σχήμα 4.2

Μετά από εκτεταμένες δοκιμές στις υπερ-παραμέτρους του μοντέλου, αυτές που επέφεραν τα καλύτερα αποτελέσματα συνοψίζονται στον Πίνακα 4.2.

4.3 Μοντέλο seq2seq με biLSTM και μηχανισμό προσοχής στον κωδικοποιητή

Στην αρχιτεκτονική αυτή προσθέτουμε ένα μηχανισμό προσοχής (Ενότητα 2.3.2) στο biLSTM του κωδικοποιητή, όπως φαίνεται στο Σχήμα 4.3. Σε κάθε βήμα, ο αποκωδικοποιητής δέχεται ένα



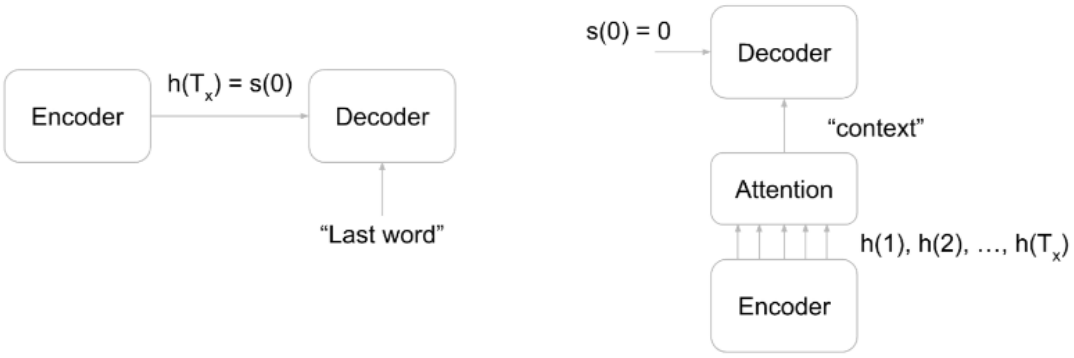
Σχήμα 4.2: Σχηματική Αναπαράσταση του seq2seq μοντέλου με biLSTM

Πίνακας 4.2: Υπερ-παράμετροι βέλτιστου seq2seq μοντέλου με χρήση biLSTM στον κωδικοποιητή

Υπερ-παράμετρος	**Τιμή**
LSTM επίπεδα κωδικοποιητή & αποκωδικοποιητή	512
Μήκος embeddings (διανυσμάτων)	100
Βελτιστοποιητής	Adam, με ρυθμό μάθησης $0,01$
Συνάρτηση απώλειας	Κατηγορική διασταυρούμενη εντροπία
Ομαλοποίηση Βαρών	L2 με $\lambda = 0,001$
Χρήστη Dropout	5% σε όλα τα κρυφά επίπεδα
Μέγεθος δέσμης	64
Εποχές εκπαίδευσης	14

διάνυσμα προς αποκωδικοποίηση (μετάφραση) που αντιστοιχεί στο σταθμισμένο μέσο των κρυφών επιπέδων των λέξεων της εισόδου με τα βάρη προσοχής. Το διάνυσμα αυτό περιέχει την πληροφορία από όλες τις κρυφές καταστάσεις και αποσκοπεί στον να «επικεντρωθεί» ο αποκωδικοποιητής στην κρυφή κατάσταση με το μεγαλύτερο βάρος, για να μεταφράσει όσο το δυνατόν καλύτερα κάθε λέξη.

Μετά από εκτεταμένες δοκιμές στις υπερ-παραμέτρους του μοντέλου, αυτές που επέφεραν τα καλύτερα αποτελέσματα συνοψίζονται στον Πίνακα 4.3.



Σχήμα 4.3: Αρχιτεκτονική seq2seq με μηχανισμό προσοχής. Αριστερά η προηγούμενη αρχιτεκτονική seq2seq και δεξιά η προσθήκη του μηχανισμού προσοχής.

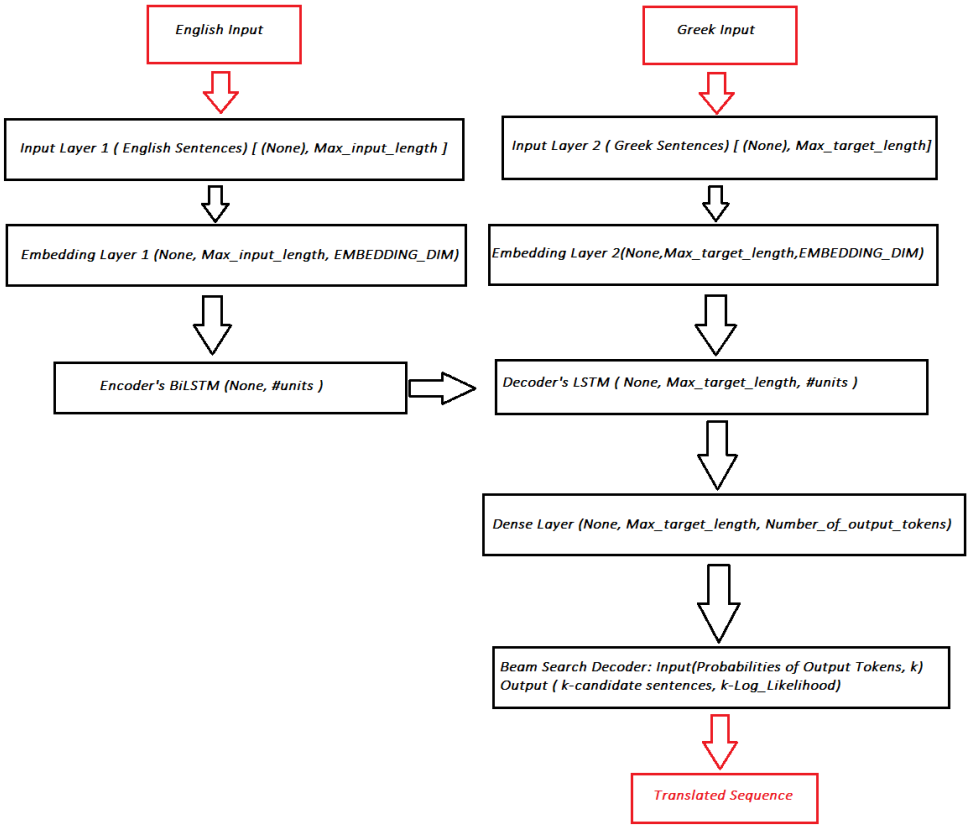
Πίνακας 4.3: Υπερ-παράμετροι βέλτιστου seq2seq μοντέλου με χρήση biLSTM και μηχανισμού προσοχής στον κωδικοποιητή

Υπερ-παράμετρος	**Τιμή**
LSTM επίπεδα κωδικοποιητή & αποκωδικοποιητή	512
Μήκος embeddings (διανυσμάτων)	100
Βελτιστοποιητής	Adam, με ρυθμό μάθησης $0,01$
Συνάρτηση απώλειας	Κατηγορική διασταυρούμενη εντροπία
Χρήστη Dropout	10% σε όλα τα κρυφά επίπεδα
Μέγεθος δέσμης	64
Εποχές εκπαίδευσης	23

4.4 Μοντέλο seq2seq με αποκωδικοποιητή ακτινικής αναζήτησης

Στην αρχιτεκτονική αυτή συνδυάζεται το μοντέλο seq2seq με έναν αποκωδικοποιητή ακτινικής αναζήτησης (Ενότητα 2.4). Η έξοδος του δικτύου τη χρονική στιγμή t_n δεν βασίζεται στην άπληστη επιλογή, δηλαδή στη λέξη με τη μεγαλύτερη πιθανότητα αλλά λαμβάνει υπόψη όλα τα πιθανά ζεύγη λέξεων των $t_0, t_1, \ldots, t_{n-1}$ χρονικών στιγμών, μεγιστοποιώντας την πιθανότητα του συνδυασμού τους. Στο συγκεκριμένο μοντέλο, για κάθε λέξη x_i που επιχειρούμε να μεταφράσουμε, λαμβάνουμε από την υλοποίηση του sequence-to-sequence μοντέλου τις πιθανότητες που το δίκτυο αντιστοιχίζει σε κάθε λέξη $x_1, x_2, \ldots, x_N$ του λεξικού και στη συνέχεια επιχειρούμε να βελτιώσουμε τη μετάφραση, με τον αλγόριθμο ακτινικής αναζήτησης. Τα αποτελέσματα αυτής της υλοποίησης ήταν πολύ ενθαρρυντικά καθώς ο αλγόριθμος επιτυγχάνει σημαντικές διορθώσεις στην έξοδο του δικτύου. Η σχηματική αναπαράσταση δίνεται στο Σχήμα 4.4

Μετά από εκτεταμένες δοκιμές στις υπερ-παραμέτρους του μοντέλου, αυτές που επέφεραν τα καλύτερα αποτελέσματα συνοψίζονται στον Πίνακα 4.4.



Σχήμα 4.4: Αρχιτεκτονική seq2seq με αποκωδικοποιητή ακτινικής αναζήτησης

Πίνακας 4.4: Υπερ-παράμετροι βέλτιστου seq2seq μοντέλου με αποκωδικοποιητή ακτινικής αναζήτησης

Υπερ-παράμετρος	**Τιμή**
LSTM επίπεδα κωδικοποιητή & αποκωδικοποιητή	256
Μήκος embeddings (διανυσμάτων)	100
Βελτιστοποιητής	Adam, με ρυθμό μάθησης $0,01$
Συνάρτηση απώλειας	Κατηγορική διασταυρούμενη εντροπία
Ομαλοποίηση Βαρών	L2 με $\lambda = 0,001$
Ομαλοποίηση Πολώσεων	L2 με $\lambda = 0,0001$
Χρήστη Dropout	5% σε όλα τα κρυφά επίπεδα
Μέγεθος δέσμης	64
Εποχές εκπαίδευσης	23

Κεφάλαιο 5

Πειραματική Διαδικασία

Σε αυτό το Κεφάλαιο περιγράφεται αναλυτικά η πειραματική διαδικασία που ακολουθήθηκε στο πλαίσιο της διπλωματικής εργασίες. Αρχικά, αναφέρονται οι συλλογές κειμένων που χρησιμοποιήθηκαν (Ενότητα 5.1), κατόπιν οι μετρικές αξιολόγησης (Ενότητα 5.2), στη συνέχεια η προ-επεξεργασία που υφίστανται τα δεδομένα πριν τροφοδοτήσουν τα συστήματα μηχανικής μάθησης καθώς και η εξαγωγή χαρακτηριστικών από αυτά (Ενότητα 5.3) και τέλος παρουσιάζονται τα αποτελέσματα της αξιολόγησης των αλγορίθμων μηχανικής μετάφρασης (Ενότητα 5.4).

5.1 Συλλογές Κειμένων

5.1.1 Tatoeba Project

Στη συγκεκριμένη διπλωματική εργασία χρησιμοποιήθηκαν δύο συλλογές κειμένων για τη διερεύνηση του προβλήματος της μηχανικής μετάφρασης της αγγλικής γλώσσας στην ελληνική: (i) ζεύγη προτάσεων από το *Tatoeba Project* [tato20] και (ii) ζεύγη προτάσεων από την Ευρωπαϊκή Ένωση [eu18]. Η πρώτη συλλογή αποτελείται από 15.966 δείγματα προτάσεων στα αγγλικά και στα ελληνικά. Τα δεδομένα είναι δημόσια διαθέσιμα και έχουν μεταφραστεί από εθελοντές Έλληνες και μη μεταφραστές, στο πλαίσιο του Tatoeba Project, μιας κοινότητας για ελεύθερη συλλογή φράσεων και μεταφράσεων. Παρόλο που ενθαρρύνεται η συμμετοχή ανθρώπων των οποίων η μητρική γλώσσα σχετίζεται με τις μεταφράσεις, για την όσο το δυνατόν καλύτερη μετάφραση και την ελαχιστοποίηση του θορύβου στα δείγματα, υπάρχουν και περιπτώσεις όπου οι μεταφράσεις δεν είναι απόλυτα ακριβείς. Τα δείγματα συλλέγονται από το *Anki* [anki20] σε μορφή αρχείου απλού κειμένου, με τις εγγραφές να είναι της μορφής `English + TAB + Greek + TAB + Attribution`. Το σύνολο δεδομένων για την ελληνική γλώσσα δεν είναι τόσο ικανοποιητικό, μιας και για άλλα ζεύγη γλωσσών υπάρχουν πολυ περισσότερα δείγματα (όπως λ.χ. για το ζεύγος αγγλικών-γερμανικών που υπάρχουν 221.533 ζεύγη). Ωστόσο, υπάρχουν και άλλα ζεύγη γλωσσών με πολύ λιγότερα δείγματα (λ.χ. αγγλικά-γεωργιανά με 141 εγγραφές).

Το μήκος των προτάσεων κυμαίνεται από 2 έως 20 λέξεις, γεγονός που όπως θα δούμε στη συνέχεια δυσχεραίνει τη διαδικασία εκμάθησης λόγω περιορισμένου αριθμού λέξεων. 'Ενα χαρακτηριστικό του συγκεκριμένου συνόλου δεδομένων είναι ότι το ίδιο δείγμα στα αγγλικά μπορεί να αντιστοιχεί σε περισσότερες από μία μεταφράσεις στα ελληνικά. Αυτό συμβαίνει διότι η αγγλική γλώσσα δεν περιέχει καταλήξεις για το γένος όπως η ελληνική, για παράδειγμα η μετάφραση της πρότασης `I'm angry` μπορεί να είναι `Είμαι θυμωμένος` αλλά και `Είμαι θυμωμένη`. Ενδεικτικά παραδείγματα δίνονται στο Σχήμα 5.1.

Επίσης, η ελληνική γλώσσα χαρακτηρίζεται από ένα εξαιρετικά πλούσιο λεξιλόγιο και το γεγονός

I'd rather be playing golf.	Προτιμώ να παίξω γκόλφ.
I'd rather be playing golf.	Προτιμώ να παίζω γκόλφ.
I'd rather be playing golf.	Θα προτιμούσα να παίξω γκόλφ.
I'd rather be playing golf.	Θα προτιμούσα να παίζω γκόλφ.
Who has it?	Ποιος το 'χει;
Who has it?	Ποιος το έχει;
Who has it?	Ποιος τον έχει;
Who has it?	Ποιος την έχει;
I want to try again.	Θέλω να ξαναδοκιμάσω.
I want to try again.	Θέλω να ξαναπροσπαθήσω.

Σχήμα 5.1: Δείγματα από το πρώτο σύνολο δεδομένων

αυτό συντελεί στην ύπαρξη πολλών συνωνύμων. Η λέξη work, για παράδειγμα, δύναται να μεταφραστεί είτε ώς εργάζομαι, είτε ως δουλεύω. Τέτοιες περιπτώσεις απαντώνται σε μεγάλο βαθμό στο σύνολο των δεδομένων μας και συντελούν σε ένα βαθμό, στην προσπάθεια κατανόησης των γλωσσικών φαινομένων που συναντώνται μεταξύ των δύο γλωσσών.

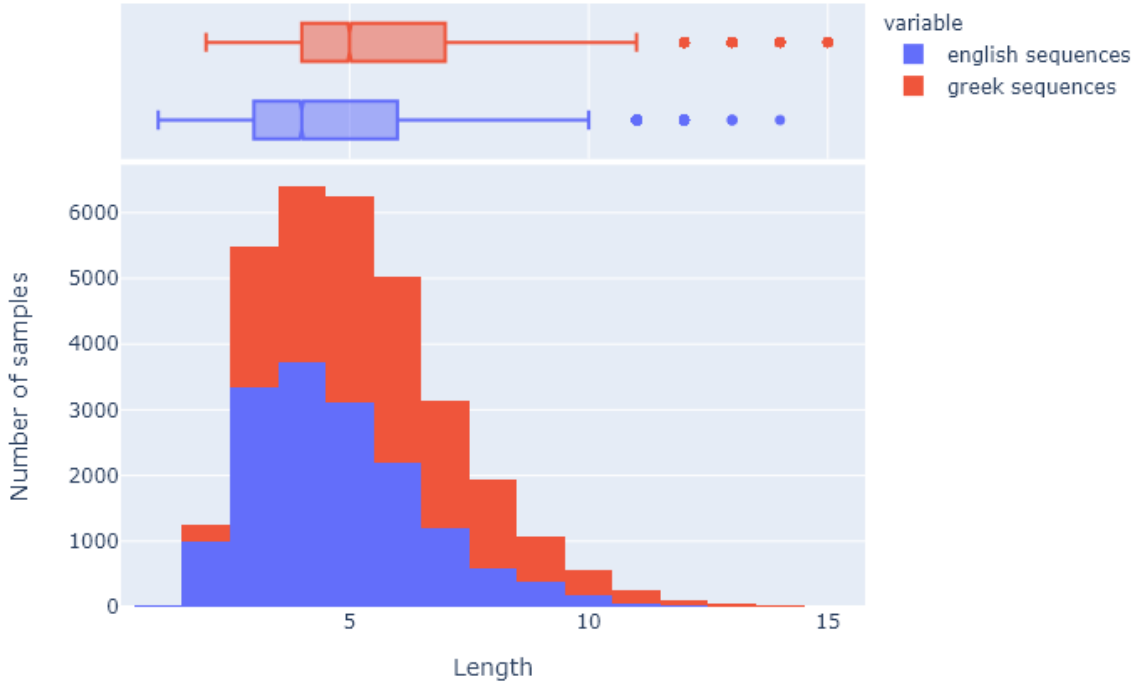
5.1.2 Μεταφράσεις κειμένων της Ευρωπαϊκής Ένωσης

Το δεύτερο σύνολο δεδομένων αποτελείται απο 12.509 δείγματα μεταφρασμένα από κείμενα της Ευρωπαϊκής Ένωσης [eu18]. Τα δεδομένα αυτά έχουν ως περιεχόμενο γεγονότα που αφορούν την Ελλάδα και την Ευρώπη στο πλαίσιο της Διακυβέρνησης και της Δημόσιας Διοίκησης, συνεπώς το λεξιλόγιο που χρησιμοποιείται είναι περιορισμένο και περιέχει ονόματα οργανισμών, πολιτικών και χωρών της ΕΕ, περιορίζοντας, όπως θα δούμε, ακόμα περισσότερο τη διαδικασία της μετάφρασης. Ενδεικτικά παραδείγματα δίνονται στο Σχήμα 5.2. Τα δεδομένα αυτά παρέχονται σε μορφή *XML* (Extensible Markup Language) και έχουν πλούσια μεταδεδομένα. Ωστόσο, στην παρούσα εργασία διατηρήθηκαν μόνο τα σχετικά με την αυτόματη μηχανική μετάφραση πεδία, δηλαδή, τα ζεύγη ελληνικών-αγγλικών ακολουθιών.

Represents the Country, in matters of international development, at the relevant international fora and organizations (European Union, OECD/DAC and other international organizations and funds of global or regional character)	*Για την εκπροσώπηση της χώρας στα θέματα αναπτυξιακής συνεργασίας εντός του πλαισίου της Ε.Ε., της Επιτροπής Αναπτυξιακής Βοήθειας (D.A.C.) του Ο.Ο.Σ.Α. και των λοιπών Διεθνών ή Περιφερειακών Οργανισμών, Διεθνών Ταμείων, Περιφερειακών Αναπτυξιακών Τραπεζών και Διεθνών Διασκέψεων που ασχολούνται με θέματα ανάπτυξης.*
The Consumer Price Index (CPI) is compiled by the National Statistical Service of Greece (NSSG) and is used for the measurement of the changes of the general level of prices of the goods and services that make up the "household basket".	*Ο Δείκτης Τιμών Καταναλωτή (ΔΤΚ) καταρτίζεται από την Ελληνική Στατιστική Αρχή (ΕΛΣΤΑΤ) και χρησιμοποιείται για τη μέτρηση των μεταβολών του γενικού επιπέδου των τιμών των αγαθών και υπηρεσιών που συνθέτουν το "καλάθι της νοικοκυράς".*

Σχήμα 5.2: Δείγματα από το δεύτερο σύνολο δεδομένων

Αξίζει να σημειωθεί το γεγονός ότι και στα δύο σύνολα δεδομένων οι ακολουθίες της ελληνικής γλώσσας αντιστοιχούν σε μεγαλύτερα μήκη σε σχέση με τις αντίστοιχες αγγλικές όπως φαίνεται και στα Σχήματα 5.3-5.4. Αυτό επιβεβαιώνει την αρχική μας παρατήρηση, ότι το μήκος της μεταφρασμένης ακολουθίας διαφέρει από το μήκος της ακολουθίας εισόδου (Σχήμα 5.5).



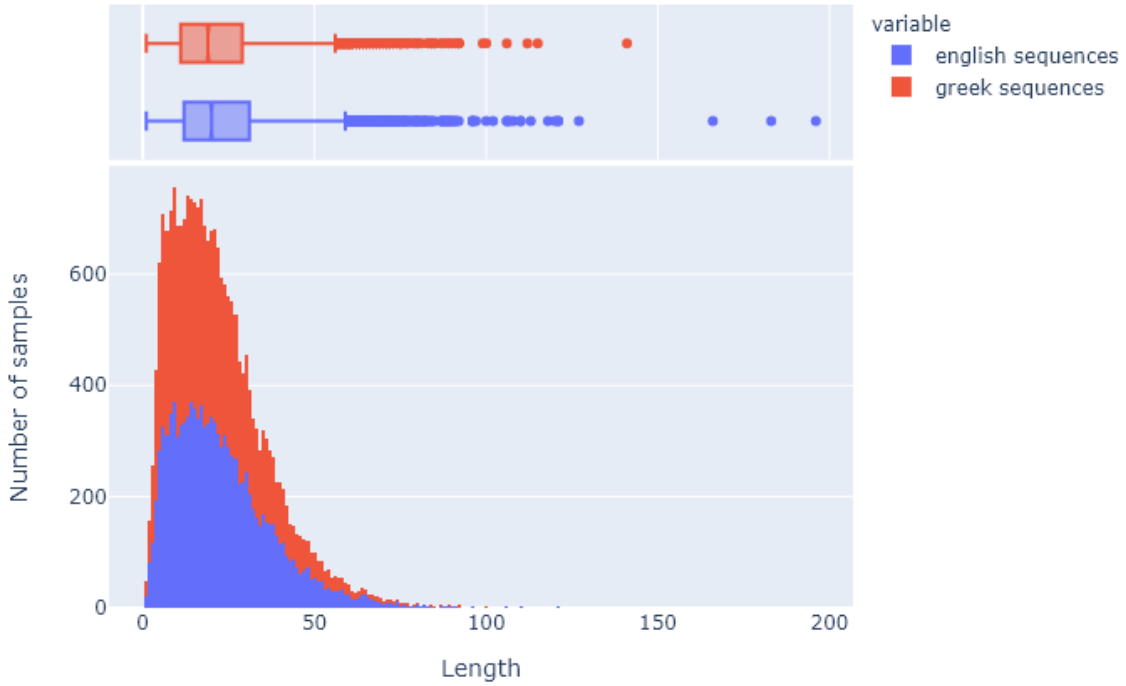
Σχήμα 5.3: Κατανομή μήκους δειγμάτων από το Tatoeba Project.

5.2 Μετρικές Αξιολόγησης

Στη διαδικασία της εκπαίδευσης, όπου διαθέτουμε την επιθυμητή έξοδο, είναι εύλογο να χρησιμοποιήσουμε ως μετρική αξιολόγησης την *ορθότητα* (accuracy), δηλαδή το ποσοστό των λέξεων που το δίκτυο πρόβλεψε σωστά συγκρίνοντας με την πραγματική μετάφραση (Εξίσωση 5.1).

$$Accuracy = \frac{Number\ of\ correct\ words\ predicted}{Total\ number\ of\ words\ predicted} \tag{5.1}$$

Στη διαδικασία του ελέγχου, ωστόσο, υπάρχει δυσκολία στο να αξιολογήσουμε αν η πρόβλεψη του δικτύου αντιστοιχεί νοηματικά στην πραγματική. Ένας τρόπος να κατανοήσουμε καλύτερα την απόδοση του αλγορίθμου μας είναι να μετρήσουμε πόση από την πληροφορία που θέλαμε να μεταφράσουμε αποδόθηκε τελικά στη γλώσσα-στόχο, κατά πόσο μεταδόθηκε σωστά η πληροφορία αυτή και τέλος, αν η τελική μετάφραση είναι γραμματικά σωστή. Ο πιο δημοφιλής τρόπος να αξιολογήσουμε το αποτέλεσμα της μετάφρασης είναι η μετρική *bilingual evaluation understudy* (BLEU) που μετρά την ομοιότητα της πρόβλεψης του δικτύου (υποψήφια-μηχανική μετάφραση) με την ανθρώπινη μετάφραση (ανθρώπινη μετάφραση-αναφορά) [Papi02].



Σχήμα 5.4: Κατανομή μήκους δειγμάτων από τις μεταφράσεις της Ευρωπαϊκή Επιτροπής.

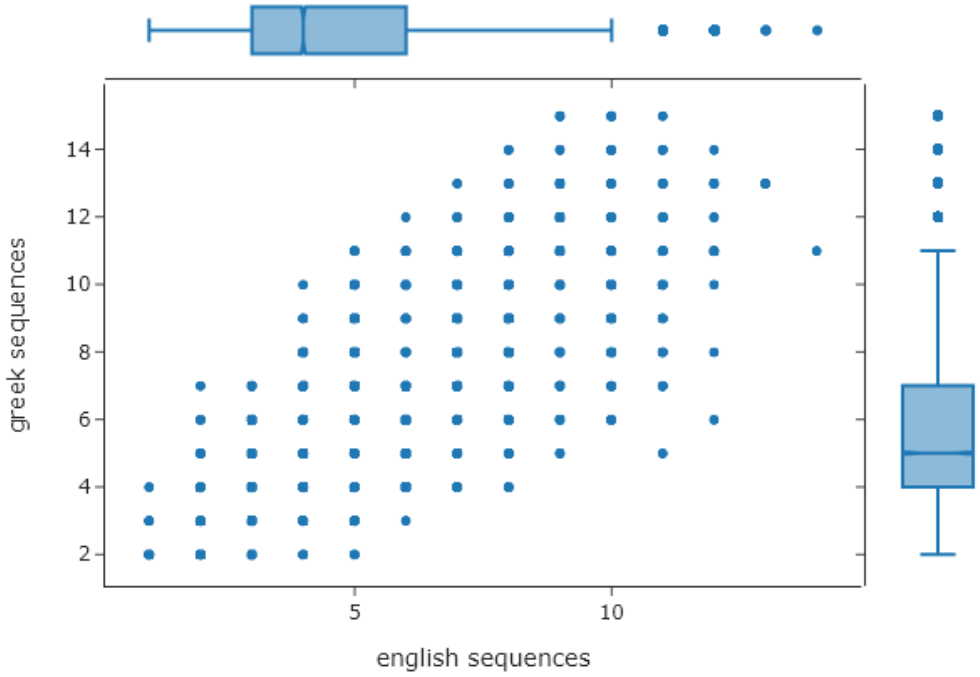
Η μετρική υπολογίζει την ομοιότητα μεταξύ των n-grams της υποψήφιας μετάφρασης και του κειμένου αναφοράς, ανεξάρτητα από τη θέση αυτών των ομοιοτήτων. Στη συνέχεια, υπολογίζει μια τροποποιημένη εκδοχή της ακρίβειας (modified n-gram precision). Ωστόσο, επειδή το μήκος της ακολουθίας της πρόβλεψης μπορεί να διαφέρει από το μήκος της αναφοράς, ο αλγόριθμος λαμβάνει υπόψη την αναντιστοιχία αυτή με το brevity penalty (BP, Εξίσωση 5.2). Το BP ισούται με 1 όταν το μήκος της υποψήφιας ακολουθίας είναι το ίδιο με αυτό της αναφοράς. Όσο πιο κοντινά είναι τα μήκη των ακολουθιών, τόσο καλύτερη είναι η μετάφραση.

$$BP = \begin{cases} 1 & \text{if } r \leq c \\ e^{1-\frac{r}{c}} & \text{if } c \leq r \end{cases} \tag{5.2}$$

όπου r το πλήθος λέξεων στην πραγματική μετάφραση-αναφορά και c το πλήθος λέξεων στην υποψήφια-μηχανική μετάφραση. Έτσι, η μετρική BLEU διαμορφώνεται όπως στην Εξίσωση 5.3 παρακάτω.

$$BLEU = BP \cdot e^{\sum_{n=1}^{N} w_n \log p_n} \tag{5.3}$$

όπου N ο αριθμός των n-grams (π.χ. 2-gram, 3-gram, 4-gram κλπ), w_n το βάρος κάθε τροποποιημένης ακρίβειας (συνήθως για $N = 4$, είναι $w_n = \frac{1}{4} = 0,25$) και p_n η τροποποιημένη ακρίβεια. Η μετρική BLEU κυμαίνεται στο $[0, 1]$, λαμβάνοντας την τιμή 1 όταν η πρόταση-έξοδος του δικτύου είναι ταυτόσημη με την πραγματική μετάφραση.



Σχήμα 5.5: Μήκος ακολουθίας εισόδου-εξόδου (αγγλικών-ελληνικών) από στο Tatoeba Project

5.3 Προ-επεξεργασία Δεδομένων και Εξαγωγή Χαρακτηριστικών

Τα ζεύγη προτάσεων διαχωρίστηκαν στις αγγλικές ακολουθίες λέξεων που θα αποτελέσουν την είσοδο του δικτύου, και στις αντίστοιχες μεταφράσεις τους στα ελληνικά. Σε κάθε ακολουθία-στόχο (πρόταση που τελικά το δίκτυο θέλουμε να προβλέψει στα ελληνικά), προστίθεται στην αρχή η συμβολοσειρά `<sos>` ή αλλιώς *start of sequence*, και στο τέλος η συμβολοσειρά `<eos>` ή *end of sequence*, ώστε όταν η πρόταση θα τροφοδοτηθεί στο δίκτυο του αποκωδικοποιητή, να υπάρχει ένδειξη για την πρώτη και την τελευταία λέξη προκειμένου να σηματοδοτηθεί η έναρξη ή η λήξη της μετάφρασης. Στη συνέχεια, διασπάμε τις προτάσεις σε *λεξικογραφικές μονάδες* (tokens), αντιστοιχίζοντας κάθε λέξη σε ένα μοναδικό αριθμό του λεξιλογίου που θα χαρακτηρίζει τη λέξη αυτή. Αξίζει να σημειωθεί το γεγονός ότι, στο συγκεκριμένο πρόβλημα λόγω της ύπαρξης δύο διαφορετικών γλωσσών, είναι απαραίτητη η ύπαρξη δύο διαφορετικών λεξικών για την αντιστοίχηση των λέξεων. Για το πρώτο σύνολο δεδομένων ενδεικτικά, βρέθηκαν 3.210 μοναδικές λέξεις για το συνολικό δείγμα στα αγγλικά και 8.743 για τα ελληνικά. Και πάλι παρατηρούμε, ότι το πλήθος των μοναδικών λέξεων στα ελληνικά είναι 2, 7 φορές μεγαλύτερο από το αντίστοιχο στα αγγλικά.

Προκειμένου να έχουμε ίδιο μήκος ακολουθιών στις εισόδους των LSTMs, συμπληρώνουμε μηδενικά σε κάθε ακολουθία (padding), σύμφωνα με το μήκος της μέγιστης ακολουθίας των δεδομένων. Το μήκος των ακολουθιών δίνεται στο Σχήμα 5.3. Για τη συλλογή κειμένων του Tatoeba Project χρησιμοποιήθηκε μήκος padding 12 για τις αγγλικές ακολουθίες που θα τροφοδοτηθούν στο δίκτυο προς μετάφραση και 15 για τις αναμενόμενες ακολουθίες εξόδου στα ελληνικά. Σε αρκετές περιπτώσεις, η ύπαρξη μικρού ποσοστού δειγμάτων μήκους πολύ μεγαλύτερου από το μέσο μήκος των ακολουθιών,

οδηγεί στην προσθήκη πολλών μηδενικών στα υπόλοιπα δείγματα, με αποτέλεσμα να αυξάνεται η υπολογιστική πολυπλοκότητα χωρίς να προσδίδεται πληροφορία. Το πρόβλημα αυτό όπως θα δούμε, απαντήθηκε κατά τη συνένωση των δύο διαφορετικών συλλογών κειμένων.

Για τη συλλογή κειμένων από μεταφράσεις της Ευρωπαϊκής Ένωσης χρησιμοποιήθηκε μήκος padding 40 για τις αγγλικές ακολουθίες που θα τροφοδοτηθούν στο δίκτυο προς μετάφραση και 60 για τις αναμενόμενες ακολουθίες εξόδου στα ελληνικά. Κάποιες έκτοπες τιμές στο σύνολο δεδομένων αφαιρέθηκαν ώστε να μην επηρεάσουν τη δημιουργία των χαρακτηριστικών.Με τον τρόπο αυτό πετύχαμε, χάνοντας λίγη πληροφορία, να κερδίσουμε σε υπολογιστική πολυπλοκότητα και χρόνο.

Τέλος, για κάθε λέξη δημιουργείται ένας πίνακας χαρακτηριστικών ο οποίος διαμορφώνεται ανάλογα με το ποια προ-εκπαιδευμένα διανύσματα λέξεων (word embeddings) χρησιμοποιούνται. Πιο αναλυτικά, για την ελληνική γλώσσα χρησιμοποιούμε Word2Vec αναπαραστάσεις 100 διαστάσεων για κάθε λέξη [Fare17]. Για την αγγλική γλώσσα χρησιμοποιούνται διανύσματα Glove. Μετά από πειραματισμό, καταλήξαμε στο συμπέρασμα ότι από τις τέσσερις διαφορετικές διαστάσεις πινάκων που είναι διαθέσιμες (50,100,200,300) , τα καλύτερα αποτελέσματα για το πρόβλημά έδωσε η αναπαράσταση με τις 100 διαστάσεις για κάθε στοιχείο (λέξη) [Penn14].

5.4 Αποτελέσματα και Αξιολόγηση Αλγορίθμων Μετάφρασης

5.4.1 Αξιολόγηση του μοντέλου seq2seq

Στη διαδικασία της εκπαίδευσης του μοντέλου που περιγράφηκε στο Κεφάλαιο 4 ενσωματώθηκε η τεχνική teacher-forcing που περιγράφηκε στο Κεφάλαιο 2. Για το σκοπό αυτό, οι προτάσεις-στόχος υφίστανται μια ολίσθηση κατά ένα βήμα στο χρόνο, ώστε η τρέχουσα πρόβλεψη μιας λέξης στο χρόνο t, να λαμβάνει υπόψη την πραγματική λέξη στο $t-1$ και όχι την πρόβλεψη του αποκωδικοποιητή. Αντίθετα, κατά τη φάση του ελέγχου, η μέθοδος αυτή είναι αδύνατο να εφαρμοστεί, καθώς το δίκτυο καλείται να δημιουργήσει τη μετάφραση λέξη προς λέξη, χωρίς παρεμβάσεις για διόρθωση.

Πίνακας 5.1: Πίνακας αποτελεσμάτων του seq2seq μοντέλου με διαφορετικές υπερπαραμέτρους για τη συλλογή κειμένων από το Tatoeba Project

Εποχές	Πλήθος νευρώνων	Ομαλοποίηση	Βελτιστοποίηση	BLEU	Ορθότητα (σύνολο ελέγχου)
12	256	L2	Adam	0.224	**0.423**
35	128	L2	Adam	0.209	0.348
11	256	L2	SGD	0.181	0.291
13	256	L1	RMSprop	0.130	0.243

Ένα σημαντικό μειονέκτημα της υλοποίησης αυτής είναι η αδυναμία πρόβλεψης νέων δεδομένων, ο οποίος οφείλεται σε φαινόμενα υπερ-προσαρμογής κατά τη διαδικασία της εκπαίδευσης και στην αδυναμία του μοντέλου να προβλέψει τις μεταφράσεις του συνόλου ελέγχου. Προκειμένου να διορθώσουμε την υπερπροσαρμογή, διακόπτουμε την εκπαίδευση όταν το σφάλμα παύει να μειώνεται για τρεις συνεχόμενες εποχές με την τεχνική του πρόωρου τερματισμού, καθώς περαιτέρω εκπαίδευση δεν πρόκειται να οδηγήσει σε βελτίωση του αποτελέσματος.

Επιπρόσθετα, εφαρμόζουμε μεθόδους ομαλοποίησης L_1 και L_2 και εισάγουμε επίπεδα dropout που σύμφωνα με τη βιβλιογραφία μειώνουν την υπερεκπαίδευση. Ένας πολύ σημαντικός παράγο-

ντας για την αδυναμία γενίκευσης του μοντέλου είναι η έλλειψη μεγάλου αριθμού δειγμάτων αφού χρησιμοποιούμε ένα σύνολο δεδομένων 15.966 δειγμάτων, εκ των οποίων το 80% χρησιμοποιείται για εκπαίδευση και το 20% για έλεγχο. Ειδικότερα στο πρόβλημά της μετάφρασης που αφορά το σύνολο των λέξεων μιας γλώσσας, είναι αδύνατο με τόσα λίγα δείγματα να επιτύχουμε ικανοποιητικό αποτέλεσμα. Ο Πίνακας 5.1 περιέχει τα αποτελέσματα που πετύχαμε για τις δύο μετρικές που εξετάζουμε (Ενότητα 5.2) και για διαφορετικές υπερ-παραμέτρους του μοντέλου (σε όλες τις περιπτώσεις το μέγεθος δέσμης ήταν 64 δείγματα).

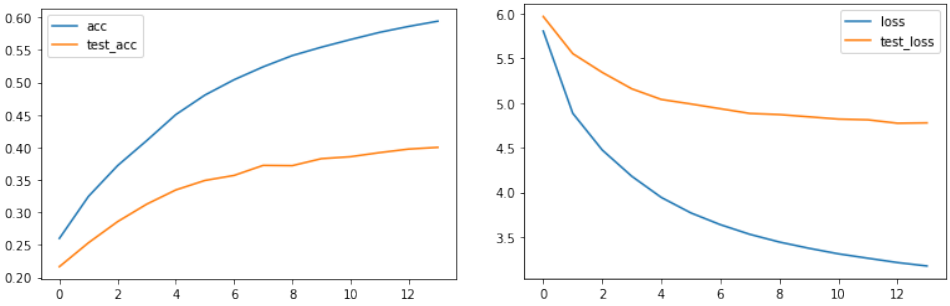
Πίνακας 5.2: Πίνακας αποτελεσμάτων του seq2seq μοντέλου με διαφορετικές υπερπαραμέτρους για τη συλλογή κειμένων από την Ευρωπαϊκή Ένωση

Εποχές	Πλήθος νευρώνων	Ομαλοποίηση	Βελτιστοποίηση	BLEU	Ορθότητα (σύνολο ελέγχου)
35	256	L2	Adam	0.13	0.192
33	256	L2	Adam	0.121	0.172
30	256	L2	SGD	0.16	0.152

Στην προσπάθεια να βελτιώσουμε την απόδοση του δικτύου μετάφρασης, επιχειρήσαμε να ενσωματώσουμε τη δεύτερη συλλογή κειμένων (από τις μεταφράσεις της Ευρωπαϊκής Ένωσης). Ωστόσο, αντιμετωπίσαμε πρόβλημα λόγω διαφορετικού μήκους των προτάσεων. Συγκεκριμένα, στην πρώτη συλλογή κειμένων το 50% των δειγμάτων εισόδου αποτελείται από 4 λέξεις και η μέγιστη ακολουθία από 15 λέξεις, όπως φαίνεται στην κατανομή του Σχήματος 5.3. Αντίθετα, στη δεύτερη συλλογή κειμένων, το 50% των δειγμάτων αποτελείται από 21 λέξεις, δηλαδή περίπου 5 φορές μεγαλύτερο μήκος (Σχήμα 5.4). Δημιουργώντας κοινού μήκους αναπαραστάσεις των 50, 60 και 70 λέξεων για να ενώσουμε τα δεδομένα των δύο συλλογών, τα αποτελέσματα δεν ήταν ικανοποιητικά, λόγω εισαγωγής μεγάλου αριθμού μηδενικών για να «στοιχηθούν» οι αναπαραστάσεις, πράγμα που έκανε τη διαδικασία εξαιρετικά χρονοβόρα και υπολογιστικά πολύπλοκη. Για το λόγο αυτό παρουσιάζονται τα αποτελέσματα για κάθε σύνολο δεδομένων χωριστά (στον Πίνακα 5.2 παρουσιάζονται τα αποτελέσματα για τη δεύτερη συλλογή κειμένων).

5.4.2 Αξιολόγηση του μοντέλου seq2seq με biLSTM στον κωδικοποιητή

Παρόλο που αναμέναμε καλύτερο αποτέλεσμα με την προσθήκη του biLSTM, τα αποτελέσματα είναι σχεδόν τα ίδια με αυτά του απλού seq2seq μοντέλου για τη συλλογή κειμένων του Tatoeba Project, αφού πετύχαμε ορθότητα 38, 75% έναντι 42, 3% και BLEU score 20, 3% και έναντι 22, 4%. Για τη συλλογή κειμένων της Ευρωπαϊκής Ένωσης, ωστόσο, όπου οι ακολουθίες είναι σημαντικά μεγαλύτερες, τα αποτελέσματα είναι λίγο καλύτερα, δίνοντας ορθότητα 26, 43%. Αυτό είναι αναμενόμενο, καθώς το biLSTM αποσκοπεί στην καλύτερη εύρεση των εξαρτήσεων μεταξύ των λέξεων σε μεγαλύτερου μήκους ακολουθίες, όπου το απλό δίκτυο LSTM είναι αδύνατο να «θυμηθεί». Η διακύμανση των μετρικών κατά τη διαδικασία της εκπαίδευσης και του ελέγχου για τα κείμενα της συλλογής του Tatoeba Project δίνεται στο Σχήμα 5.6, ενώ οι βέλτιστες τιμές των υπερ-παραμέτρων και για τις δύο συλλογές κειμένων δίνονται στον Πίνακα 5.3 (και στις δύο περιπτώσεις χρησιμοποιήθηκε ομαλοποίηση L_2 και ο βελτιστοποιητής Adam).



Σχήμα 5.6: Διακύμανση μετρικών κατά την εκπαίδευση του seq2seq μοντέλου με biLSTM στον κωδικοποιητή στα κείμενα του Tatoeba Project

Πίνακας 5.3: Πίνακας αποτελεσμάτων του seq2seq μοντέλου με biLSTM στον κωδικοποιητή για διαφορετικές υπερπαραμέτρους

Συλλογή κειμένων	Εποχές	Πλήθος νευρώνων	BLEU	Ορθότητα (σύνολο ελέγχου)
Tatoeba Project	14	256	**0.203**	**0.3875**
Ευρωπαϊκής Ένωσης	19	512	**0.145**	**0.2643**

5.4.3 Αξιολόγηση του μοντέλου seq2seq με μηχανισμό προσοχής

Ένα σημαντικό μειονέκτημα της προηγούμενης υλοποίησης είναι το γεγονός ότι ολόκληρη η ακολουθία εισόδου κωδικοποιείται στο τελευταίο κρυφό επίπεδο του κωδικοποιητή. Είδαμε, ότι ακόμα και αν τροφοδοτήσουμε την ακολουθία στον κωδικοποιητή και από τις δύο κατευθύνσεις χρησιμοποιώντας ένα δίκτυο biLSTM για να ανιχνεύσουμε εξαρτήσεις από το μέλλον, το αποτέλεσμα δεν βελτιώθηκε αρκετά. Να σημειωθεί στο σημείο αυτό, ότι το LSTM του αποκωδικοποιητή δε μπορεί να γίνει διπλής κατεύθυνσης καθώς μεταφράζει μία λέξη σε κάθε βήμα, συνεπώς, δεν υπάρχει διαθέσιμη ακολουθία για να τροφοδοτηθεί προς τα πίσω και η τροποποίηση αυτή δεν θα είχε νόημα στο πρόβλημά μας.

Στην αρχιτεκτονική που εξετάζουμε τώρα, προσθέτουμε ένα επίπεδο προσοχής στο biLSTM του κωδικοποιητή όπως παρουσιάστηκε στην Ενότητα 2.3.2. Σε κάθε βήμα, ο αποκωδικοποιητής δέχεται ένα διάνυσμα προς μετάφραση, που αντιστοιχεί στο σταθμισμένο μέσο των κρυφών επιπέδων των λέξεων της εισόδου με τα βάρη προσοχής. Το διάνυσμα αυτό περιέχει την πληροφορία από όλες τις κρυφές καταστάσεις και αποσκοπεί στο να «επικεντρωθεί» ο αποκωδικοποιητής στην κρυφή κατάσταση με το μεγαλύτερο βάρος για να μεταφράσει όσο το δυνατόν καλύτερα κάθε λέξη. Τα αποτελέσματα αυτής της αρχιτεκτονικής δίνονται στον Πίνακα 5.4

Πίνακας 5.4: Πίνακας αποτελεσμάτων του seq2seq μοντέλου με μηχανισμό προσοχής, για διαφορετικές υπερπαραμέτρους στη συλλογή κειμένων από το Tatoeba Project

Μέγεθος δέσμης	Εποχές	Πλήθος νευρώνων	Ομαλοποίηση	Βελτιστοποίηση	BLEU	Ορθότητα (σύνολο ελέγχου)
64	23	256	L2	Adam	**0.219**	**0.401**
64	22	256	L1	Adam	0.13	0.321
32	11	256	L2	RMSprop	0.145	0.280

Η αρχιτεκτονική αυτή επέφερε ελάχιστα καλύτερο αποτέλεσμα σε σχέση με το μοντέλο seq2seq

με BiLSTM δίνοντας ακρίβεια $40,1\%$ έναντι $38,75\%$ και BLUE score $21,9\%$ στη συλλογή κειμένων του Tatoeba Project. Το αποτέλεσμα, ωστόσο, δεν ξεπέρασε το αρχικό μοντέλο seq2seq.

Μια εξήγηση θα μπορούσε να είναι ότι το μοντέλο γίνεται αρκετά πολύπλοκο προσθέτοντας το μηχανισμό προσοχής, και αυξάνεται ο αριθμός των παραμέτρων εκπαίδευσης (σε πάνω από 4 εκατομμύρια) με μικρό αριθμό δειγμάτων. Επίσης, στις μεταφρασμένες ακολουθίες παρατηρείται σε κάποιες περιπτώσεις, επανάληψη ορισμένων λέξεων ακόμα και στην ίδια ακολουθία. Το φαινόμενο αυτό, όπως θα δούμε, οφείλεται στη διαδικασία της αποκωδικοποίησης και θα προσπαθήσουμε να το επιλύσουμε με την επόμενη αρχιτεκτονική.

5.4.4 Αξιολόγηση του μοντέλου seq2seq με χρήση του αλγορίθμου ακτινικής αναζήτησης στον αποκωδικοποιητή

Ένα συχνό φαινόμενο που παρατηρήθηκε στις εξόδους των προηγούμενων υλοποιήσεων είναι η ύπαρξη γραμματικών αστοχιών στις μεταφράσεις. Το πρόβλημα αυτό οφείλεται στην άπληστη διαδικασία αποκωδικοποίησης, καθώς η επιλογή της λέξης με τη μεγαλύτερη πιθανότητα σε κάθε βήμα t, δεν εξασφαλίζει ότι η συνολική έξοδος θα είναι γραμματικά ορθή και δεν δίνεται η δυνατότητα τροποποίησης της συνολικής εξόδου (μετάφρασης) σε επόμενα βήματα $t+1, t+2, ..., t+N$. Επίσης, απαντάται συχνά η πρόβλεψη ορισμένων μόνο λέξεων, καθώς κάποιες λέξεις, λόγω συχνής εμφάνισής τους στο δείγμα εκπαίδευσης, αντιστοιχούν σε μεγαλύτερες πιθανότητες, σε μεγάλο ποσοστό των προβλέψεων. Τα προβλήματα αυτά οφείλονται στη διαδικασία της αποκωδικοποίησης και θα προσπαθήσουμε να τα επιλύσουμε λαμβάνοντας υπόψη τον αλγόριθμο ακτινικής αναζήτησης. Το πλεονέκτημα αυτής της υλοποίησης είναι η ευελιξία που παρέχει, ώστε ακόμα κι αν μια λέξη που έχει ήδη προβλεφθεί σε προηγούμενο χρονικά βήμα εσφαλμένα ($t-1$) και δεν οδηγεί στη μέγιστη πιθανότητα της συνολικής πρόβλεψης στο t, να μπορεί να διορθωθεί.

Στον Πίνακα 5.5 παρατίθενται μερικά αποτελέσματα και η βελτιστοποιημένη έξοδος που επιτυγχάνεται με χρήση του αλγορίθμου ακτινικής αναζήτησης. Η είσοδος είναι η πρόταση που αντιστοιχεί στην αγγλική γλώσσα, η έξοδος είναι η μετάφραση του νευρωνικού δικτύου χωρίς τον αλγόριθμο ακτινικής αναζήτησης, ενώ δίνονται και οι $k=5$ πιο πιθανές έξοδοι με χρήση του αλγορίθμου καθώς και ο λογάριθμος της πιθανοφάνειάς τους. Έτσι, όπως φαίνεται και στον Πίνακα 5.5 στο τρίτο παράδειγμα, παρόλο που το δίκτυο έχει προβλέψει εσφαλμένα τη μετάφραση της πρότασης `You were happy` ως `Ήσουν απασχολημένος`, η πρόβλεψη διορθώνεται από τον αλγόριθμο με τη σωστή λέξη ως `Ήσαστε ευτυχισμένοι` δίνοντας τη μέγιστη πιθανοφάνεια $6,3661$. Φυσικά στο παράδειγμα αυτό, είναι εμφανής η αδυναμία κατανόησης του γένους, αφού βλέπουμε ότι οι πιθανές προτάσεις αντιστοιχούν τόσο σε ενικό όσο και σε πληθυντικό στα ελληνικά, χωρίς να γνωρίζουμε τον αριθμό που υπονοείται στα αγγλικά. Στην περίπτωση αυτή, είναι ακόμα και για τον άνθρωπο αδύνατο να καταλάβει αν πρόκειται για ενικό ή πληθυντικό αριθμό αν δεν δίνεται κάποια άλλη πληροφορία. Το σημαντικό είναι, ωστόσο, ότι η εσφαλμένη μετάφραση της λέξης happy στην λέξη απασχολημένος, επιδιορθώνεται και μάλιστα στις πέντε πιο πιθανές μεταφράσεις εμφανίζονται λέξεις που μπορούν να χαρακτηριστούν ως συνώνυμες (ευτυχισμένος και χαρούμενος).

Στο δεύτερο παράδειγμα του Πίνακα 5.5, παρατηρούμε ότι ο αλγόριθμος διορθώνει τη μετάφραση έχοντας ανιχνεύσει την εξάρτηση της λέξης `we` με την κατάληξη του ρήματος `χρειαζόμαστε` σε πληθυντικό. Οι λέξεις `λεφτά` και `χρήματα` είναι επίσης ορθές για τη μετάφραση της λέξης `money`. Η πρόταση αυτή δίνει και τη μέγιστη πιθανοφάνεια, σε αντίθεση με το παράδειγμα `Tom wants to stay`

Πίνακας 5.5: Η βελτίωση δικτύου seq2seq με χρήση αλγορίθμου ακτινικής αναζήτησης. Τα δύο πρώτα παραδείγματα αντιστοιχούν στην υλοποίηση seq2seq + ακτινική αναζήτηση, ενώ τα επόμενα δύο αντιστοιχούν στην υλοποίηση seq2seq + μηχανισμός προσοχής + ακτινική αναζήτηση

Είσοδος	’Εξοδος	Ακτινική Αναζήτηση (k=5)	Λογάριθμος Πιθανοφάνειας
Never tell lies.	Ποτέ δε θα ψέματα.	Ποτέ δε λέει ψέματα.	**9,6760**
		Κανείς δε θα ψέματα.	9,6611
		Ποτέ δε θα λέει.	9,6321
		Ποτέ δε θα λέει.	9,5714
		Κανείς ποτέ θα ψέματα.	9,5376
We need money.	Χρειάζομαι χρήματα.	Χρειαζόμαστε λεφτά.	**7,7024**
		Χρειάζομαι λεφτά.	6,6834
		Χρειάζομαι γαλλικά.	6,6634
		Έχουμε χρήματα.	6,5476
		Χρειάζομαι χρήματα.	6,5390
You were happy.	Ήσουν απασχολημένος.	Ήσαστε ευτυχισμένη.	**6,3661**
		Ήσασταν ευτυχισμένη.	6,3312
		Ήσουν χαρούμενη.	6,2911
		Ήσαστε χαρούμενοι.	6,2244
		Ήσασταν χαρούμενοι.	6,1822
Tom wants to stay in Boston.	Ο Τομ θέλει να πάω στη Βοστόνη.	Ο Τομ θέλει να παίζει στη Βοστόνη.	**7,8221**
		Ο Τομ θέλει να’ ναι στη Βοστόνη.	7,8190
		Ο Τομ θέλει να μείνει στη Βοστόνη.	7,8131
		Ο Τομ θέλει να σας στη Βοστόνη.	7,7732
		Ο Τομ θέλει να έρθει στη Βοστόνη.	7,7321

`in Boston`, όπου η σωστή μετάφραση δεν αντιστοιχεί στην ακολουθία με τη μεγαλύτερη πιθανοφάνεια, όπως προκύπτει από τον αλγόριθμο, αλλά από την τρίτη μεγαλύτερη πιθανοφάνεια. Βέβαια, σε αρκετές περιπτώσεις παρατηρούμε ότι η διαφορά μεταξύ των $k = 5$ πιθανών προτάσεων είναι πάρα πολύ μικρή και αυτό καθιστά τη διαδικασία της επιλογής δύσκολη, καθώς θα έπρεπε να χρησιμοποιήσουμε άλλες τεχνικές για να επιλέξουμε τη βέλτιστη από τις προτεινόμενες μετάφραση. Σε ορισμένες πιθανές ακολουθίες ο αλγόριθμος δεν δίνει ικανοποιητικά αποτελέσματα, συνήθως λόγω της αντικατάστασης μια λέξης με κάποια συντακτικά μη ορθή. Αξίζει να σημειωθεί ότι μεγαλύτερες τιμές της παραμέτρου k επιφέρουν καλύτερα αποτελέσματα, αυξάνοντας όμως την υπολογιστική πολυπλοκότητα που απαιτείται για την εκτέλεση του αλγορίθμου.

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές κατευθύνσεις

Στην παρούσα διπλωματική εργασία εξετάστηκε το πρόβλημα της αυτοματοποιημένης μετάφρασης αγγλικών κειμένων στην ελληνική γλώσσα, χρησιμοποιώντας μεθόδους μηχανικής μάθησης. Εκπαιδεύοντας μοντέλα seq2seq με δίκτυα LSTM, μηχανισμούς προσοχής και χρησιμοποιώντας τον αλγόριθμο ακτινικής αναζήτησης, καταλήξαμε στα αποτελέσματα που αναφέρθηκαν στο Κεφάλαιο 5. Τα αποτελέσματα είναι ενθαρρυντικά για την επίλυση του προβλήματος σε μικρού μήκους προτάσεις, με το μοντέλο seq2seq με τη χρήση του αλγορίθμου ακτινικής αναζήτησης να εμφανίζεται ως το πιο αποδοτικό. Θα παραθέσουμε στη συνέχεια μελλοντικές προεκτάσεις, ως λύσεις στα προβλήματα και τις προκλήσεις που αντιμετωπίσαμε στο πλαίσιο αυτής της εργασίας.

Ένας σημαντικός παράγοντας για την διερεύνηση της μετάφρασης συγκεκριμένα στην ελληνική γλώσσα είναι η ύπαρξη σημαντικού όγκου δεδομένων για την εκπαίδευση βαθιών νευρωνικών δικτύων. Η διαδικασία αυτή είναι αρκετά χρονοβόρα και επίπονη στην περίπτωση που επιθυμούμε να μεταφράσουμε κείμενα με ανθρώπινη συνεισφορά. Μια άλλη ιδέα ενίσχυσης του συνόλου δεδομένων είναι η αντικατάσταση ορισμένων λέξεων με νοηματικά συναφείς τους, ώστε να έχουμε όσο το δυνατόν περισσότερες αντιστοιχίσεις μεταξύ των δύο γλωσσών. Έχει ήδη αναφερθεί ότι η ελληνική γλώσσα εμφανίζει ένα εξαιρετικά πλούσιο λεξιλόγιο, με αποτέλεσμα πολλές λέξεις να εκλείπουν από την αγγλική, ή να εμφανίζονται ως συνώνυμα.

Μία ακόμα πιθανή προέκταση αποτελεί η τεχνική της *μεταφοράς μάθησης* (transfer learning), όπου το μοντέλο που δημιουργήσαμε θα μπορούσε να εκπαιδευτεί πρώτα στο ίδιο πρόβλημα, αλλά για μία γλώσσα με περισσότερα διαθέσιμα δεδομένα, για παράδειγμα τη γερμανική. Στη συνέχεια, θα μπορούσαμε να χρησιμοποιήσουμε τα προεκπαιδευμένα βάρη του δικτύου για να τα προσαρμόσουμε στην ελληνική γλώσσα. Η μέθοδος αυτή μπορεί να αποδειχθεί πολύ αποδοτική σε παρόμοια προβλήματα σπάνιων γλωσσών, με ελάχιστα διαθέσιμα δεδομένα για εκπαίδευση [Zoph16].

Στη σύγχρονη βιβλιογραφία, ακόμα, εμφανίζονται ολοένα και περισσότερες αρχιτεκτονικές που περιλαμβάνουν τους σύγχρονους *μετασχηματιστές* (transformers) για την επίλυση του προβλήματος. Προ εκπαιδευμένα γλωσσικά μοντέλα όπως ο *bidirectional encoder representations transformer* (BERT) [Devl18] και οι αναπαραστάσεις *embeddings from language models* (ELMo) [Pete18] αποτελούν μελλοντικές προσθήκες στο πρόβλημα της μηχανικής μετάφρασης. Ήδη η χρησιμοποίηση αναπαραστάσεων που προκύπτουν από γλωσσικά μοντέλα στην κωδικοποίηση και αποκωδικοποίηση της ακολουθίας βασιζόμενη σε σύνθετους μηχανισμούς προσοχής, δίνει σύμφωνα με τη βιβλιογραφία πολύ ενθαρρυντικά αποτελέσματα [Zhu20, Zhan20b, Yang20a]. Πρόσφατα η Facebook, δημοσίευσε το πρώτο μοντέλο μετάφρασης 100 γλωσσών που δεν χρησιμοποιεί ως ενδιάμεση γλώσσα μετάφρασης την αγγλική, όπως συνήθιζε να γίνεται ως τώρα για τα περισσότερα ζεύγη γλωσσών [Tang20]. Επίσης, το μοντέλο T5 της Google [Raff19], τα Bart [Lewi19] και MarianMT [Junc18] αποτελούν

μοντέλα κωδικοποίησης-αποκωδικοποίησης που επιτυγχάνουν μεταφράσεις με χρήση μετασχηματιστών, δίνοντας άμεσα και ικανοποιητικά αποτελέσματα. Το μειονέκτημα εκπαίδευσης τέτοιων μοντέλων είναι ότι απαιτούν τεράστιο όγκο δεδομένων και, άρα, πόρων που είναι δύσκολα διαθέσιμοι στο πλαίσιο εκπόνησης μιας διπλωματικής εργασίας.

Βιβλιογραφία

[Amid20] Afshine Amidi and Shervine Amidi, "Recurrent Neural Networks cheatsheet", online, November 2020.

[anki20] "Tab-delimited Bilingual Sentence Pairs from the Tatoeba Project (Good for Anki and Similar Flashcard Applications)", online, November 2020.

[Beng03] Yoshua Bengio, Réjean Ducharme, Pascal Vincent and Christian Janvin, "A Neural Probabilistic Language Model", *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1137–1155, March 2003.

[Cho14] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau and Yoshua Bengio, "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches", in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar, October 2014, Association for Computational Linguistics.

[DePa06] Donald A DePalma, Benjamin B Sargent and Renato S Beninatto, "Can't read, won't buy: Why language matters on global websites", Technical report, Common Sense Advisory Inc, 2006.

[Devl18] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *CoRR*, vol. abs/1810.04805, 2018.

[eu18] "Parallel corpus (Greek - English) in the public administration domain", online, February 2018.

[Fare17] Murhaf Fares, Andrey Kutuzov, Stephan Oepen and Erik Velldal, "Word vectors, reuse, and replicability: Towards a community repository of large-text resources", in *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pp. 271–276, Gothenburg, Sweden, May 2017, Association for Computational Linguistics.

[Forc11] Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez and Francis M. Tyers, "Apertium: a free/open-source platform for rule-based machine translation", *Machine Translation*, vol. 25, no. 2, pp. 127–144, 2011.

[Frea90] Marcus Frean, "The Upstart Algorithm: A Method for Constructing and Training Feedforward Neural Networks", *Neural Computation*, vol. 2, no. 2, pp. 198–209, 1990.

[Good16] Ian Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*, The MIT Press, 2016.

[Gu18] Jiatao Gu, Hany Hassan, Jacob Devlin and Victor O. K. Li, "Universal Neural Machine Translation for Extremely Low Resource Languages", 2018.

[Hayk09] Simon S Haykin, *Neural networks and learning machines*, Prentice Hall, 2009.

[Hoch97] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory", *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, November 1997.

[Hutc04a] J. Hutchins and Evgenii Lovtskii, "Petr Petrovich Troyanskii (1894–1950): A Forgotten Pioneer of Mechanical Translation", *Machine Translation*, vol. 15, pp. 187–221, 2004.

[Hutc04b] W. John Hutchins, "The Georgetown-IBM Experiment Demonstrated in January 1954", in Robert E. Frederking and Kathryn B. Taylor, editors, *Machine Translation: From Real Users to Research*, pp. 102–114, Berlin, Heidelberg, 2004, Springer Berlin Heidelberg.

[Junc18] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins and Alexandra Birch, "Marian: Fast Neural Machine Translation in C++", in *Proceedings of ACL 2018, System Demonstrations*, pp. 116–121, Melbourne, Australia, July 2018, Association for Computational Linguistics.

[Kari19] Raimi Karim, "Attn: Illustrated Attention", online, January 2019.

[Koeh03] Philipp Koehn, Franz J. Och and Daniel Marcu, "Statistical Phrase-Based Translation", in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003.

[Koeh07] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation", in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June 2007, Association for Computational Linguistics.

[Kost19] Simeon Kostadinov, "Understanding Encoder-Decoder Sequence to Sequence Model", online, February 2019.

[Lann19] Quinn Lanners, "Neural Machine Translation", online, July 2019.

[Lewi19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov and Luke Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension", *CoRR*, vol. abs/1910.13461, 2019.

[Medr77] Mark F. Medress, Franklin S Cooper, Jim W. Forgie, CC Green, Dennis H. Klatt, Michael H. O'Malley, Edward P Neuburg, Allen Newell, DR Reddy, B Ritea et al., "Speech understanding systems: Report of a steering committee", *Artificial Intelligence*, vol. 9, no. 3, pp. 307–316, 1977.

[Miko13] Tomas Mikolov, Quoc V. Le and Ilya Sutskever, "Exploiting Similarities among Languages for Machine Translation", *CoRR*, vol. abs/1309.4168, 2013.

[Mitc97] Tom M. Mitchell, "Machine Learnig", *Machine Learning*, 1997.

[Olah15] Christopher Olah, "Understanding LSTM Networks", online, August 2015.

[Palm86] G. Palm, "Warren McCulloch and Walter Pitts: A Logical Calculus of the Ideas Immanent in Nervous Activity", in Günther Palm and Ad Aertsen, editors, *Brain Theory*, pp. 229–230, Berlin, Heidelberg, 1986, Springer Berlin Heidelberg.

[Papi02] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002, Association for Computational Linguistics.

[Penn14] Jeffrey Pennington, Richard Socher and Christopher D Manning, "Glove: Global Vectors for Word Representation.", in *EMNLP*, vol. 14, pp. 1532–1543, 2014.

[Pete18] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer, "Deep contextualized word representations", *CoRR*, vol. abs/1802.05365, 2018.

[Pree20] Wadhwani Preeti and Gankar Saloni, "Machine Translation Market Size By Technology (Statistical Machine Translation (SMT), Rule-based Machine Translation (RBMT), Neural Machine Translation, Hybrid Machine Translation, Example-based Machine Translation), By Deployment Model (On-Premise, Cloud), By Application (Automotive, BFSI, E-Commerce, Electronics, Healthcare, IT & Telecommunications, Military & Defense), Industry Analysis Report, Regional Outlook, Growth Potential, Competitive Market Share & Forecast, 2020 – 2026", Technical report, Global Market Insights, 2020.

[Qais18] Shahzad Qaiser and Ramsha Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", *International Journal of Computer Applications*, vol. 181, 07 2018.

[Raff19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *CoRR*, vol. abs/1910.10683, 2019.

[Rude17] Sebastian Ruder, "An overview of gradient descent optimization algorithms", 2017.

[Ruge20] Pierrick Rugery, "Attention is all you need", online, July 2020.

[Sudh17] Shreenidhi Sudhakar, "Learning Rate Scheduler", online, August 2017.

[Suts14] Ilya Sutskever, Oriol Vinyals and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks", 2014.

[Tang20] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu and Angela Fan, "Multilingual Translation with Extensible Multilingual Pretraining and Finetuning", 2020.

[tato20] "Tatoeba: Collection of sentences and translations", online, November 2020.

[Way20] Andy Way, Rejwanul Haque, Guodong Xie, Federico Gaspari, Maja Popovic and Alberto Poncelas, "Facilitating Access to Multilingual COVID-19 Information via Neural Machine Translation", 2020.

[Yang20a] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang and Lei Li, "Towards Making the Most of BERT in Neural Machine Translation", 2020.

[Yang20b] Shuoheng Yang, Yuxin Wang and Xiaowen Chu, "A Survey of Deep Learning Techniques for Neural Machine Translation", 2020.

[Zhan10] Yin Zhang, Rong Jin and Zhi-Hua Zhou, "Understanding bag-of-words model: A statistical framework", *International Journal of Machine Learning and Cybernetics*, vol. 1, pp. 43–52, 12 2010.

[Zhan20a] Aston Zhang, Zachary C. Lipton, Mu Li and Alexander J. Smola, *Dive into Deep Learning*, 2020.

[Zhan20b] Zhebin Zhang, Sai Wu, Dawei Jiang and Gang Chen, "BERT-JAM: Boosting BERT-Enhanced Neural Machine Translation with Joint Attention", 2020.

[Zhu20] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li and Tie-Yan Liu, "Incorporating BERT into Neural Machine Translation", 2020.

[Zoph16] Barret Zoph, Deniz Yuret, Jonathan May and Kevin Knight, "Transfer Learning for Low-Resource Neural Machine Translation", 2016.

Παράρτημα Α

Ευρετήριο Ακρωνυμίων και Συντμήσεων

ANN: Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

BERT: Αναπαράσταση με αμφίδρομη κωδικοποίηση transformer (Bidirectional Encoder Representations Transformer)

BiLSTM: Αμφίδρομο Δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Bidirectional Long Short Term Memory)

ELMo: Αναπαραστάσεις από Γλωσσικά Μοντέλα (Embedding from Language Models)

LSTM: Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short Term Memory)

MLP: Πολυεπίπεδο Δίκτυο Perceptron (Multi Layer Perceptron)

MT: Μηχανική Μετάφραση (Machine Translation)

NLP: Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)

NMT: Νευρωνική Μηχανική Μετάφραση (Neural Machine Translation)

RNN: Αναδρομικό Νευρωνικό Δίκτυο (Recurrent Neural Network)

seq2seq: Μοντέλο Ακολουθίας σε Ακολουθία (Sequence-to-Sequence)

TF-IDF: (Term Frequency-Invers Term Frequency)