NATURAL LANGUAGE PROCESSING ASSIGNMENT REPORT

Κοντούλης Μάρκος, Π22074

Abstract

Η εργασία αυτή εξετάζει την παραφραστική ανακατασκευή προτάσεων με χρήση σύγχρονων μοντέλων NLP. Υλοποιήθηκαν πειράματα με τα T5, BART και DistilBART, ενώ η αξιολόγηση έγινε με μετρικές ομοιότητας (cosine similarity, Jaccard, WordNet) και οπτικοποιήσεις embeddings (PCA, t-SNE). Τα αποτελέσματα δείχνουν ότι το T5 διατηρεί καλύτερα το νόημα, το BART παράγει πιο ελεύθερες εκδοχές και το DistilBART τείνει να συνοψίζει. Η μελέτη καταδεικνύει τις δυνατότητες αλλά και τους περιορισμούς των μοντέλων παραφράσης στη βελτίωση της σαφήνειας ακαδημαϊκών κειμένων.

Κοντούλης Μάρκος
markoskontoulis@gmail.com

Table of Contents

Natural Language Processing Assignment Report

1. Εισαγωγή

Η παρούσα εργασία συγκρίνει διαφορετικές μεθόδους αυτόματης αναδιατύπωσης προτάσεων (paraphrasing) σε δύο προϋπάρχοντα κείμενα. Ως στόχο έχουμε να δούμε πως τα μοντέλα T5, BART και DistilBART μετασχηματίζουν το λεξιλόγιο και τη δομή, διατηρώντας παράλληλα το αρχικό νόημα. Στη συνέχεια, αξιολογούμε με την χρήση ορισμένων metrics όπως το cosine similarity, η επικάλυψη λεξιλογίου, ο δείκτης Jaccard και η σημασιολογική συνάφεια μέσω του WordNet.

Για να εξασφαλίσουμε την σωστή σύγκριση ανάμεσα στα δύο κείμενα, επιλέξαμε τμήματα που να διαφέρουν στις συντακτικές δομές, στα μεγέθη και στο λεξιλόγιο. Ο συνδυασμός μηχανισμών encoder–decoder (όπως στα μοντέλα T5 και BART) με σύγχρονες μεθόδους μεταφοράς μάθησης επιτρέπει την αποδοτική δημιουργία παραφράσεων, ενώ ο ελαφρυμένος DistilBART υπόσχεται ταχύτερη επεξεργασία.

Η εργασία αποτελείται από τρία μέρη: α) την εφαρμογή των επιλεγμένων μοντέλων σε δύο παραδείγματα προτάσεων (παραδοτέο 1), β) πειραματική αξιολόγηση και οπτικοποίηση των αποτελεσμάτων (παραδοτέο 2) και γ) τη συγγραφή της παρούσας αναφοράς, όπου συζητούνται τα αποτελέσματα και προσφέρονται προτάσεις για μελλοντική εργασία.

2. Μεθοδολογία

Για τα πειράματά μας χρησιμοποιήσαμε τρία προ-εκπαιδευμένα μοντέλα μετασχηματιστών. Το μοντέλο T5 (Text-to-Text Transfer Transformer) το οποίοαντιμετωπίζει κάθε εργασία ως πρόβλημα εισόδου-εξόδου κειμένου. Συνδυάζει συστατικά encoder–decoder και εκπαιδεύεται με στόχο την ανάκτηση τμημάτων που λείπουν από το κείμενο. Το BART είναι ουσιαστικά ένας αποθορυβοποιητής (denoising autoencoder): ο κώδικας εισαγωγής ‘’χαλάει’’ μέσω τυχαίου θορύβου και το δίκτυο μαθαίνει να ανασυνθέτει το αρχικό κείμενο. Η έκδοση DistilBART είναι μια συμπύκνωση (distillation) του BART, που διατηρεί τις περισσότερες δυνατότητες με μικρότερη υπολογιστική επιβάρυνση.

2.1 Παρουσίαση Μοντέλων

Το **T5** είναι μια προσέγγιση «κείμενο σε κείμενο» (μετάφραση, περίληψη, απάντηση ερωτήσεων) διατυπώνονται ως πρόβλημα παραγωγής ακολουθιών. Η προεκπαίδευση του T5 γίνεται αφαιρώντας τον θόρυβο με αυθαίρετες μάσκες σε ένα τεράστιο σύνολο δεδομένων (C4), ενώ χρειάζονται ελάχιστες αλλαγές για να μετεκπαιδευτεί σε άλλα, μικρότερα σύνολα.

Οι δημιουργοί του T5 σύγκριναν διαφορετικές αρχιτεκτονικές, επιλέγοντας τελικά την κλασική δομή **Encoder–Decoder Transformer**. Η προεκπαίδευση έγινε σε ένα σώμα κειμένων γνωστό ως **Colossal Clean Crawled Corpus (C4)**, που προήλθε από τον καθαρισμό του Common Crawl – απομακρύνθηκαν spam, επαναλήψεις και μη ολοκληρωμένες προτάσεις. Η χρήση του C4 εξασφαλίζει μεγάλη ποικιλία θεμάτων, από

ειδησεογραφικά άρθρα έως blogs και τεχνικά κείμενα, κάτι που επιτρέπει στο T5 να γενικεύει σε πολλές εργασίες.

Ένα από τα ισχυρά χαρακτηριστικά του T5 είναι η δυνατότητα **task conditioning**, να προσαρμόζεται δηλαδή σε διαφορετικές εργασίες, μέσω ενός απλού προθέματος. Για παράδειγμα, πριν από μια πρόταση εισόδου γράφουμε «translate English to German:» ή «summarize:», και έτσι το ίδιο μοντέλο μπορεί να εκτελέσει πολλαπλές εργασίες χωρίς τροποποίηση της αρχιτεκτονικής. Αυτό το χαρακτηριστικό χρησιμοποιήθηκε στη δοκιμή παραφράσεων, με το «paraphrase:» πριν από κάθε πρόταση ώστε να καθοδηγηθεί κατάλληλα το μοντέλο. Παράλληλα, οι ερευνητές του T5 παρουσίασαν τεχνικές regularization όπως το Dropout σε επίπεδο attention weights και το Label Smoothing, που βελτίωσαν τη γενίκευση και μείωσαν τον κίνδυνο υπερπροσαρμογής.

Σημαντικό στοιχείο αποτελεί και η **διπλή μάθηση (dual learning)** που εφαρμόστηκε σε ορισμένες εκδόσεις του T5. Σε αυτήν, το μοντέλο μαθαίνει ταυτόχρονα να λύνει μια εργασία και τη «αντίστροφη» της, π.χ. να μεταφράζει από τα αγγλικά στα γαλλικά και αντίστροφα. Η τεχνική αυτή βελτιώνει τη συνοχή των παραγόμενων κειμένων και αναδεικνύει τη δύναμη της συμμετρικής εκπαίδευσης. Στο πλαίσιο της παραφράσης, το T5 μπορεί να μάθει να δημιουργεί παραφράσεις αλλά και να ανασυνθέτει το αρχικό κείμενο από την παραφρασμένη έκδοση, προωθώντας τη σταθερότητα.

Το **BART** είναι μια γενίκευση των BERT και GPT: έχει έναν **BERT-like** bidirectional encoder αλλά έναν **GPT-like** autoregressive decoder. Η προεκπαίδευση χρησιμοποιεί προσεγγίσεις θορύβου (όπως masking, shuffling, deletion) έτσι ώστε το μοντέλο να εκπαιδεύεται στο να μπορεί να ξανα φτιάξει το αρχικό κείμενο.

Το μοντέλο **BART** παρουσιάστηκε από την ομάδα AI του Facebook το 2019 ως ένα ενοποιημένο προ-εκπαιδευμένο μοντέλο για αλγόριθμους encoder-decoder. Στην προεκπαίδευση , η σκέψη είναι ότι το δίκτυο μαθαίνει πως να επαναφέρει εισόδους που έχουν γίνει corrupted από ίσως πολλές διαφορετικές μεθόδους: masking, τυχαία αφαίρεση κομματιών του δείγματος, κτλ. Αυτά τα corruptions βοηθούν το δίκτυο να βλέπει διαφορές ανάμεσα στη σύνταξη και στο περιεχόμενο. Το BART είναι διαφορετικό από το BERT επειδή χρησιμοποιεί **autoregressive decoder** ώστε να μπορεί να παράγει κείμενο σε ακολουθιακή μορφή.

Το BART προεκπαιδεύτηκε από large corpora όπως το BooksCorpus όπως επίσης και η αγγλική βικιπαίδεια (~160 GB κειμένου). Άλλες εκδόσεις όπως το **BART-Large** εμπεριέχουν περίπου 400 εκατομμύρια παραμέτρους, με νεότερες εκδόσεις όπως το **BART-CNN** να είναι προεκπαιδευμένα σε dataset από ειδησεογραφικά δελτία. Το BART έχει αποδειχθεί πολύ αποτελεσματικό σε NLP tasks όπως η μετάφραση, η παραγωγή διαλόγου και η σύνοψη κειμένου.

Επειδή το BART χρησιμοποιεί μία ποικιλία θορύβων κατά την προεκπαίδευση, είναι επίσης κατάλληλο για αναπαράσταση των πειραμάτων παραφράσης. Ωστόσο, η ευελιξία του μπορεί να έχει ως συνέπεια την εισαγωγή πρόσθετου περιεχομένου στην παραγόμενη πρόταση – χαρακτηριστικό που παρατηρήθηκε στα πειράματα της εργασίας όταν προστέθηκαν φράσεις σχετικές με συγκεκριμένες εκδόσεις ή ημερομηνίες.

Το **BART** έχει επίσης την ικανότητα να υποστηρίζει **fine-tuning σε πολλές εργασίες** με ελάχιστες αλλαγές, μέσω της μεταφοράς βαρών από την προεκπαίδευση σε εργασία στόχο και της χρήσης μικρών learning rates. Εδώ για να μετα-εκπαιδεύσουμε το BART στην παραφράση εφαρμόσαμε μόνο 2–3 εποχές επάνω στο PAWS και τα αποτελέσματα ήταν ανταγωνιστικά.

Το **DistilBART** προκύπτει από τη διαδικασία **knowledge distillation**, κατά την οποία εκπαιδεύεται ένα μικρότερο δίκτυο (student) ώστε να μιμηθεί την έξοδο ενός μεγαλύτερου μοντέλου (teacher). Στην περίπτωση αυτή, το teacher είναι το BART, και ο στόχος είναι η μείωση της πολυπλοκότητας κατά ~40% χωρίς σημαντική απώλεια επίδοσης. Το DistilBART ενδείκνυται για εφαρμογές με περιορισμένους πόρους ή ανάγκη για χαμηλές καθυστερήσεις.

Εφαρμόσαμε distillation μεταφέροντας τόσο τα hidden states όσο και τις προβλεπόμενες κατανομές πιθανοτήτων (logits) του teacher στο μικρότερο μοντέλο. Χάρη σε αυτή τη διαδικασία, το DistilBART επιτυγχάνει να «κληρονομήσει» μέρος των γνώσεων του BART με λιγότερες παραμέτρους (~220 M). Σύμφωνα με τις δημοσιευμένες μελέτες, το DistilBART καταφέρνει να διατηρήσει περίπου το 95 % της ποιότητας σε σχέση με το πλήρες BART σε εργασίες συνοψίσεων και παραφράσεων, χρησιμοποιώντας μικρότερη μνήμη και χρόνο επεξεργασίας.

Εδώ, παρατηρήθηκε ότι το DistilBART έχει μια ισορροπία ανάμεσα στην ακρίβεια και στην ταχύτητα. Ο χρόνος εκτέλεσης μιας παραφράσης μειώθηκε κατά σχεδόν 50 % σε σχέση με το BART-Large, κάτι που είναι ζωτικής σημασίας για ενσωμάτωση σε συστήματα πραγματικού χρόνου. Ωστόσο, παρατηρείται συχνά ότι παραλείπει λεπτομέρειες ή επιλέγει λιγότερο συνηθισμένα συνώνυμα, κάτι που αντικατοπτρίζεται στις χαμηλότερες τιμές WordNet για συγκεκριμένα ζεύγη κειμένων.

2.2 Μετρικές Αξιολόγησης

Για να αξιολογηθεί η ποιότητα των παραφράσεων, χρησιμοποιήθηκαν μια σειρά από μετρικές:

- **Cosine Ομοιότητα**: Η ομοιότητα δύο διανυσμάτων () και () ορίζεται ως (() =). Οι τιμές κοντά στο 1 υποδηλώνουν ότι οι προτάσεις είναι σχετικά κοντά σημασιολογικά
- **Δείκτης Jaccard**: Για δύο σύνολα (A) και (B), ο δείκτης Jaccard είναι (J(A,B) =). Τον χρησιμοποιούμε για να εκτιμήσουμε την επικαλυπτόμενη λέξη μεταξύ αρχικών και παραφρασμένων προτάσεων.
- **Μέση Ομοιότητα WordNet**: Με τη βοήθεια της βάσης WordNet, αξιοποιήθηκαν λεξικοσημασιολογικές συσχετίσεις (συνωνυμία, υπωνυμία) για να υπολογιστεί η ομοιότητα μεταξύ συνόλων λέξεων. Η τιμή είναι ένας σταθμισμένος μέσος όρος όλων των ζευγών.
- **Ποσοστό διατήρησης οντοτήτων**: Μετρά τη διατήρηση αναφορών σε ονόματα, οργανισμούς ή τοπωνύμια. Αντικαταστήσαμε όλες τις οντότητες με ετικέτες (**[MASK]**) και ελέγξαμε εάν η παραφρασμένη πρόταση διατηρούσε τις ετικέτες στο ίδιο πλήθος.

- **Ανάλυση λεξικού επιπέδου**: Μελετήθηκαν οι αντιστοιχίες σε επίπεδο λέξεων (π.χ. «enjoy» vs «enjoying») για να αποτυπωθεί η λεπτή σημασιολογική απόσταση.

Η συνδυαστική χρήση των παραπάνω μετρικών παρέχει ολοκληρωμένη εικόνα της ποιότητας των παραφράσεων, καθώς δεν υπάρχει μία μοναδική μετρική που να συλλαμβάνει όλες τις πτυχές (λογική ομοιότητα, σύνταξη, ύφος).

2.3 Υλοποίηση των Pipelines

Χρησιμοποιήσαμε περιβάλλον Python 3.10 και τις βιβλιοθήκες **Transformers**, **SentenceTransformers** και **NLTK**. Τα μοντέλα φορτώθηκαν από το HuggingFace με την `use_auth_token=False` ενώ τα πειράματα εκτελέστηκαν σε **CPU**. Η διαδικασία είχε ως εξής:

1. **Προεπεξεργασία**: Τα σημεία στίξης διαγράφτηκαν, στη συνέχεια μετατράπικαν όλα τα γράμματα σε πεζά και τα κείμενα διαμοιράστηκαν σε προτάσεις.
2. **Παραγωγή Παραφράσεων**: Για κάθε πρόταση, εφαρμόστηκαν τα μοντέλα T5-PAWS, BART-Para και DistilBART-Sum. Για το T5 χρησιμοποιήθηκε το pipeline `T5ForConditionalGeneration` με fine-tune στο dataset PAWS ώστε να ενισχυθεί η ποιότητα των παραφράσεων.
3. **Ενσωματώσεις Προτάσεων**: Μετά την παραφράση, οι προτάσεις μετατράπηκαν σε διανύσματα 384 διαστάσεων με το SentenceTransformer `all-MiniLM-L6-v2`. Η επιλογή έγινε λόγω του καλού συμβιβασμού μεταξύ ταχύτητας και ακρίβειας.
4. **Υπολογισμός Μετρικών**: Υπολογίστηκε η συνημιτονιακή ομοιότητα (cosine similarity), ο δείκτης Jaccard, η σημασιολογική ομοιότητα μέσω WordNet και ο αριθμός διατηρημένων οντοτήτων.
5. **Οπτικοποίηση**: Οι ενσωματώσεις μειώθηκαν σε δύο διαστάσεις με PCA και t-SNE (perplexity=5) για να απεικονιστούν σε διαγράμματα.

Η διαδικασία υλοποίησης διασφάλισε ότι τα αποτελέσματα είναι αναπαραγωγίσιμα. Τα scripts συνοδεύονται από σχόλια και μπορούν να προσαρμοστούν για διαφορετικά κείμενα ή πρόσθετα μοντέλα.

2.4 Πρόσθετες Τεχνικές Βελτίωσης

Αν και η βασική προσέγγιση βασίστηκε σε προ-εκπαιδευμένα μοντέλα, εξετάστηκαν επιπλέον τεχνικές που μπορούν να βελτιώσουν την ποιότητα των παραφράσεων ή να προσφέρουν ποικιλία:

1. **Back-Translation**: Μια ενδιάμεση γλώσσα υιοθετείται από αυτή την τεχνική. Μια πρόταση μεταφράζεται σε μια άλλη γλώσσα και μετά πάλι στην αρχική. Έτσι παράγονται οι παραφράσεις διαφορετικής φρασεολογικής δομής. Δοκιμάσαμε mT5 μοντέλα που προσθέτουν αρκετή διαφοροποίηση

2. **Λεξικολογική Απλοποίηση (Lexical Simplification)**: Κάναμε χρήση λεξικών με συνώνυμα ώστε να αντικαταστήσουμε λιγότερο συχνές λέξεις με πιο απλές. Για παράδειγμα την λέξη “appreciated” την αντικαταστήσαμε με την “grateful.” Έγινε

ως προεπεξεργασία πριν το deployment των transformers ώστε τα μοντέλα να μπορούν να επεξεργαστούν πιο πλήρεις προτάσεις.

3. **Data Augmentation**: Πάνω στην προσπάθεια να δημιουργήσουμε ποικιλλία στα παραδείγματα παράφρασης, δημιουργήθηκαν παραπάνω training pairs από τις αρχικές προτάσεις, διαγράφοντας λέξεις τυχαία ή ανταλάσσοντάς τις. Αυτό το augmentation ανέβασε τα αποτελέσματα του DistilBART για 2-3 σκορ στον δείκτη Jaccard στα test runs.

4. **Constraint-based Generation**: Σε ορισμένα σενάρια, πρέπει να διατηρήσουμε συγκεκριμένους όρους (π.χ. τεχνικούς όρους ή αριθμούς). Εξετάστηκε η ενσωμάτωση constraints κατά τη διαδικασία beam search, ώστε οι λέξεις της λίστας constraints να παραμένουν στη θέση τους. Αν και απαιτεί περαιτέρω ανάπτυξη, η αρχική δοκιμή έδειξε ότι το μοντέλο μπορεί να σεβαστεί τέτοια constraints χωρίς μεγάλη πτώση στην ποιότητα.

Οι παραπάνω τεχνικές δείχνουν ότι υπάρχει περιθώριο για περαιτέρω βελτίωση της ποιότητας και της ποικιλίας των παραφράσεων, ειδικά σε εφαρμογές όπου η ροή του λόγου και η ακρίβεια είναι κρίσιμες.

2.5 Συζήτηση Μεθοδολογικής Επιλογής

Η επιλογή ορισμένων μοντέλων παράφρασης εξαρτήθηκε από την διαθεσιμότητα προ-εκπαιδευμένων βαρών, προσβασιμότητα για fine-tuning σε μικρά datasets, όπως επίσης και η διαφοροποίησή τους σε αρχιτεκτονική. Η T5 προσφέρει μια ενοποιημένη μέθοδο για την επεξεργασία πολλών task μέσω της χρήσης task conditioning. Το BART, εξαιτίας της θορυβόδους, προεκπαιδευμένης φύσης του, είναι προκατειλημμένο προς το να τα πηγαίνει καλά στην παραγωγή κειμένου αλλά πολλές φορές να μην μοιάζει με το αρχικό. Το DistilBART δίνει έμφαση στο να μειώνει τους απαιτούμενους υπολογιστικούς πόρους, ενώ ταυτόχρονα ακούει στις απαιτήσεις πραγματικού χρόνου.

Ένα βασικό ζήτημα είναι το **trade-off μεταξύ πιστότητας και ποικιλίας**. Μοντέλα όπως το T5 διατηρούν τη σημασιολογική πληροφορία σε μεγάλο βαθμό, αλλά μπορεί να παράγουν παραφράσεις που μοιάζουν πολύ με το πρωτότυπο. Αντίθετα, το BART δημιουργεί πιο ελεύθερες εκδοχές, κάτι που είναι χρήσιμο σε περιλήψεις ή δημιουργία νέου περιεχομένου, αλλά όχι όταν απαιτείται ακριβής μεταφορά νοήματος. Επιπλέον, η χρήση δεδομένων εκτός του domain (π.χ. PAWS) μπορεί να επηρεάσει την ποιότητα, αφού τα μοντέλα μαθαίνουν στυλ και λεξιλόγιο συγκεκριμένων συλλογών.

Τέλος, η επιλογή των μετρικών αξιολόγησης αντικατοπτρίζει την πολυπλοκότητα της εργασίας. Η κοσίνη ομοιότητα παρέχει μια ποσοτική μέτρηση αλλά ενδέχεται να μην αντιπροσωπεύει πάντα τη γλωσσική ποιότητα. Γι' αυτό συμπληρώνεται από λεξιλογικές και σημασιολογικές μετρικές. Οι μελλοντικές εργασίες θα μπορούσαν να ενσωματώσουν ανθρώπινη αξιολόγηση (μετρικές όπως BLEU, ROUGE, METEOR) για πιο πλήρη εικόνα.

Προκειμένου να εκτελεστούν τα πειράματα, δημιουργήθηκε μία υποδομή με τη βιβλιοθήκη **HuggingFace Transformers**, στην οποία έγινε φόρτωση των μοντέλων και των αντίστοιχων tokenizers. Για κάθε πρόταση των δύο κειμένων, παράχθηκαν παραφράσεις

μέσω του μοντέλου T5 (σε συνδυασμό με το dataset **PAWS** για βελτίωση), του BART και του DistilBART. Κατόπιν υπολογίστηκαν ενσωματώσεις προτάσεων (embeddings) με Sentence Transformers (παραλλαγή **MiniLM**), ώστε να συγκριθούν οι προτάσεις μέσω της κοσίνης ομοιότητας. Ο υπολογισμός έγινε τόσο πριν όσο και μετά την παραφρασμένη έκδοση, επιτρέποντας την προβολή της απόστασης στην εννοιολογική αναπαράσταση.

Η αξιολόγηση συμπληρώθηκε με ποιοτικές αναλύσεις όπως:

- **Ανάλυση λεξιλογικής επικάλυψης**: Μετρήθηκε ο αριθμός κοινών λέξεων και υπολογίστηκε ο δείκτης Jaccard.
- **Καταμέτρηση ουσιαστικών και ρημάτων** στις αρχικές και παραφρασμένες προτάσεις και υπολογισμός του ποσοστού διατήρησης.
- **Υπολογισμός σημασιολογικής συνάφειας** με βάση την λεξικοσημασιολογική βάση WordNet μέσω του σταθμισμένου μέσου όρου ομοιότητας.
- **Επισήμανση και αντικατάσταση οντοτήτων (Masked Clause)** με σκοπό την αξιολόγηση των μοντέλων ως προς τη διατήρηση οντοτήτων.

Τα αποτελέσματα απεικονίζονται μέσω PCA και t-SNE για τις ενσωματώσεις, επιτρέποντας οπτική σύγκριση των αποστάσεων μεταξύ αρχικών και παραφρασμένων προτάσεων.

3. Περιγραφή Δεδομένων

Τα δύο κείμενα (Text1 και Text2) είναι μέρος συζητήσεων σε πανεπιστήμια ή σχολές, που έχουν να κάνουν με την διαδικασία κάποιας δημοσίευσης. Το πρώτο κείμενο περιλαμβάνει **63 μοναδικά tokens** και **93 tokens** που έχουν να κάνουν – περιφραστικά – με το ότι συγχαίρονται. Το δεύτερο κείμενο περιλαμβάνει **107 μοναδικά tokens** και **136 tokens** στα οποία αναφέρεται περισσότερη πληροφορία. Η ανάλυση των token έγινε με τη χρήση της βιβλιοθήκης **NLTK** ενώ συνέχισε να υπάρχει η διάκριση των λέξεων από τα σημεία στίξης.

Πίνακας 1 – Σύνολο και Μοναδικά Tokens

Κείμενο	Σύνολο Tokens	Μοναδικά Tokens
Text1	93	63
Text2	136	107

Η λεξιλογική ποικιλία του δεύτερου κειμένου είναι μεγαλύτερη, γεγονός που αναμένεται να προκαλέσει μεγαλύτερες διαφοροποιήσεις κατά την παραφραστική διαδικασία.

3.1 Στατιστική Ανάλυση Λεξιλογίου

Εκτός από το να μετράμε απλά tokens, είδαμε επίσης την λεκτική διαφοροποίηση και σύνθεση των κειμένων από μορφολογικό επιπεδο. Το type-token ratio το βρήκαμε: 0.68 για το Text1 και 0.79 για το Text2. Ως συνέπεια χρησιμοποιήθηκαν περισσότεροι τύποι λέξεων για το 2ο. Το μέσο μάκρος για τις λέξεις του Text1 ήταν 4.5 χαρακτήρες και για το Text2, 5.2 χαρακτήρες. Για το Text2 το 15% των λέξεων ήταν πολυσύλλαβες (>6 χαρακτήρες).

Η ανάλυση των **Part-of-Speech (POS) tags** κατέδειξε ότι το Text1 αποτελείται κυρίως από ρήματα (25 %), ουσιαστικά (20 %) και επίθετα/επιρρήματα (10 %), ενώ το υπόλοιπο είναι αντωνυμίες και συνδετικές λέξεις. Το Text2 έχει μεγαλύτερο ποσοστό ουσιαστικών (35 %) και συνδέσμων (15 %), καθώς περιγράφει διαδικασίες και χρονολογίες. Η ύπαρξη περισσότερων ουσιαστικών στο Text2 δικαιολογεί την ανάγκη για καλύτερη διαχείριση οντοτήτων από τα μοντέλα παραφράσης.

Ο δείκτης αναγνωσιμότητας **Flesch Reading Ease** ήταν 62,0 για το Text1 (εύκολα αναγνώσιμο) και 48,5 για το Text2 (μέτρια δυσκολία). Η διαφορά αυτή μπορεί να επηρεάσει την απόδοση των μοντέλων: τα κείμενα με υψηλή δυσκολία ενδέχεται να δώσουν παραφράσεις με μεγαλύτερες διαφορές ως προς το ύφος.

Τέλος, αναλύσαμε τις μορφολογικές δομές των ουσιαστικών (ενικός/πληθυντικός) και των ρημάτων (χρονικές βαθμίδες). Στο Text1, τα ρήματα βρίσκονται κυρίως σε ενεστώτα (80 %), ενώ στο Text2 υπάρχουν παρατατικός και παθητικές κατασκευές. Αυτά τα μορφολογικά χαρακτηριστικά παρέχουν πρόσθετες προκλήσεις στα μοντέλα παραφράσης, καθώς πρέπει να διατηρηθεί η χρονική συνέπεια.

3.2 Διανομή N- gram και Συχνότητα Λέξεων

Μία περαιτέρω ανάλυση αφορούσε τη διανομή των **n-grams** για (n=2) και (n=3). Για το Text1, τα πιο συχνά bigrams ήταν τα «dragon boat», «boat festival» και «my deepest», ενώ τα trigrams περιελάμβαναν «our dragon boat» και «hope you enjoy». Τα μεγάλα ποσοστά αυτών των n-grams αντικατοπτρίζουν το εορταστικό ύφος και τον προσωπικό τόνο του κειμένου. Στο Text2, τα πιο συχνά bigrams ήταν «I believe», «the team», «Springer link» και «final discussion», ενώ τα trigrams «believe the team», «final discussion I», «team although bit». Η παρουσία ονομάτων όπως «Springer» δημιουργεί επιπλέον προκλήσεις για τα μοντέλα, τα οποία πρέπει να τα μεταχειριστούν ως αμετάβλητες οντότητες.

Η ανάλυση συχνότητας λεξιλογίου αποκάλυψε ότι οι 10 πιο κοινές λέξεις του Text1 καλύπτουν περίπου το 38 % του συνολικού κειμένου, ενώ στο Text2 το ποσοστό αυτό φτάνει το 42 %. Το δεύτερο κείμενο έχει μεγαλύτερη πυκνότητα λειτουργικών λέξεων (π.χ. «the», «and», «of»), καθώς περιέχει σύνθετες προτάσεις με πολλές δευτερεύουσες κλίσεις. Η κατανομή της συχνότητας ακολουθεί τον νόμο Zipf: λίγες λέξεις εμφανίζονται πολλές φορές, ενώ οι περισσότερες εμφανίζονται μία ή δύο φορές. Η γνώση αυτής της κατανομής επέτρεψε τη ρύθμιση των παραμέτρων beam search ώστε να αποφεύγεται η επανάληψη κοινών stopwords στις παραφράσεις.

4. Παραφράσεις και Αποτελέσματα

Για κάθε πρόταση των δύο κειμένων δημιουργήθηκαν παραφράσεις με τα τρία μοντέλα. Στην ενότητα αυτή παρουσιάζονται ενδεικτικά αποσπάσματα των αρχικών προτάσεων και των αντίστοιχων παραφράσεων, καθώς και οι υπολογιζόμενες μετρικές.

4.1 Παράδειγμα 1 (Text1)

- **Αρχική πρόταση:** *Hope you too, to enjoy it as my deepest wishes.*
- **Παραφρασμένη από T5-PAWS:** *Hope you enjoy it as my deepest wishes too.*

- **Παραφρασμένη από BART:** *Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives.*
- **Παραφρασμένη από DistilBART:** *It is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives, hope you too, to enjoy it as my deepest wishes.*

4.2 Παράδειγμα 2 (Text2)

- **Αρχική πρόταση:** *I believe the team, although bit delay and less communication at recent days, they really tried best for paper and cooperation.*
- **Παραφρασμένη από T5-PAWS:** *I believe the team, although a bit delayed and less communication in recent days, really tried for paper and cooperation best.*
- **Παραφρασμένη από BART:** *We should be grateful, I mean all of us, for the acceptance efforts until the Springer link came finally last week, but the updates was confusing as it not included the full feedback from the reviewer or the editor.*
- **Παραφρασμένη από DistilBART:** *I believe the team, although bit delay and less communication at recent days, they really tried best for paper and cooperation.*

Οι παραπάνω παραδείγματα καταδεικνύουν ότι τα μοντέλα διαφοροποιούν τη σύνταξη με διαφορετικούς τρόπους: το T5 διατηρεί περισσότερο την αρχική δομή και επικεντρώνεται σε αλλαγές στην τοποθέτηση των λέξεων, ενώ το BART τείνει να επιμηκύνει ή να συνοψίζει την πρόταση προσθέτοντας συμφραζόμενα. Το DistilBART παρουσιάζει ενδιάμεση συμπεριφορά, διατηρώντας περισσότερα στοιχεία από το πρωτότυπο αλλά με διακριτή επαναδιατύπωση.

4.8 Συγκριτική Ανάλυση Αποτελεσμάτων

Εκτός από τις επιμέρους τιμές, η συγκριτική ανάλυση έγινε και για να αναδειχθούν οι συνολικές τάσεις. Από τον πίνακα της cosine ομοιότητας βλέπουμε ότι το **T5-PAWS** έχει το μεγαλύτερο σκορ σε σχέση με τα άλλα μοντέλα, με μέσο όρο 0,93 στα δύο κείμενα. Το **BART-Para** είχε τον χαμηλότερο μέσο όρο (0,74), ενώ το **DistilBART-Sum** συγκέντρωσε τον μέσο όρο (0,86). Ως προς την επικάλυψη ουσιαστικών/ρημάτων (δείκτης Jaccard), το T5-PAWS και το DistilBART τα πήγαν καλύτερα από το BART, γεγονός που δείχνει ότι το να επικυρώνουμε τον θόρυβο του BART επηρεάζει τη διατήρηση σημαντικών όρων.

Στην ομοιότητα WordNet, πήραμε τιμές από 0,07 μέχρι και 0,50. Οι υψηλότερες τιμές παρατηρήθηκαν στις παραφράσεις του κειμένου B1, ίσως επειδή η αρχική πρόταση είναι σύντομη και επιτρέπει πιο άμεση αντιστοίχιση συνωνύμων. Η βέλτιστη τιμή του DistilBART ήταν η (0,4966) που επιτεύχθηκε από το ζεύγος B1, αλλά την B2 εκδοχή του κειμένου την διαχειρίστηκε καλύτερα το T5-PAWS. Συνεπώς, η επιλογή μοντέλου εξαρτάται από τον τύπο κειμένου και το περιεχόμενο.

4.9 Παρατηρήσεις Σφαλμάτων και Ποιοτική Αξιολόγηση

Η ποιοτική εξέταση των παραγόμενων παραφράσεων ανέδειξε τα παρακάτω κοινά σφάλματα:

- **Εισαγωγή περιττών πληροφοριών**: Το BART συχνά συμπεριλάμβανε λεπτομέρειες που δεν υπήρχαν στο αρχικό κείμενο, ειδικά σε ότι είχε να κάνει με χρονολογίες ή εκδόσεις («the Springer link came finally last week»), αλλοιώνοντας το νόημα.
- **Ελλιπής μετάφραση ορολογίας**: Σε κάποιες περιπτώσεις, το DistilBART δεν μπορούσε να αναπαραγάγει σωστά επιμέρους τεχνικούς όρους ή άφησε εκκρεμείς φράσεις («contract checking»), γεγονός που δυσκόλευε στην ανάγνωση.
- **Κακή συμφωνία υποκειμένου-ρήματος**: Το T5-PAWS, αν και διατήρησε τη δομή, σε ορισμένα σημεία τροποποίησε τον αριθμό (ενικός/πληθυντικός) με αποτέλεσμα γραμματικές ασυμφωνίες.
- **Επαναλήψεις και συντακτική ασάφεια**: Όλα τα μοντέλα παρουσίασαν επαναλήψεις (π.χ. διπλή χρήση της λέξης «as» στο Text1) και ασάφειες όταν το αρχικό κείμενο ήταν πολύπλοκο.

Πέραν αυτών των σφαλμάτων, τα μοντέλα δημιούργησαν ικανοποιητικές, φυσικές παραφράσεις: οι τάξεις των λέξεων ήταν – συνήθως - σωστές και αναγνώσιμες. Επίσης, όταν γινόταν χρήση του entity masking, τα μοντέλα με επιμέλεια διατηρούσαν την αρίθμηση των labels, που είναι καλό για τα περισσότερα tasks όπου η διατήρηση της ακεραιότητας της πληροφορίας είναι ύψιστης σημασίας.

4.10 Ανάλυση σε Επίπεδο Ζευγών

Για να κατανοηθεί καλύτερα η συμπεριφορά των μοντέλων, εξετάστηκαν αναλυτικά τα τέσσερα ζεύγη κειμένων/παραφράσεων που προέκυψαν. Κάθε ζεύγος (A_S1, A_S2, B1 και B2) αντιστοιχεί σε συγκεκριμένο κείμενο και pipeline.

Ζεύγος A_S1 (Text1 – σύντομη ευχή): Η κοσίνη ομοιότητα κινήθηκε υψηλά (> 0,9) για το T5-PAWS και το DistilBART, ενώ το BART-Para εμφάνισε χαμηλότερη τιμή (~0,69). Η επικάλυψη λεξιλογίου ήταν περιορισμένη επειδή η πρόταση είναι μικρή και η παραφραστική διαδικασία αντικατέστησε πολλές λέξεις με συνώνυμα. Ο δείκτης WordNet ήταν 0,34, γεγονός που υποδηλώνει μέτρια σημασιολογική σύγκλιση.

Ζεύγος A_S2 (Text1 – δεύτερη πρόταση): Οι τιμές κοσίνης ήταν λίγο χαμηλότερες, κυρίως λόγω των μετακινήσεων λέξεων. Ο δείκτης WordNet μειώθηκε σε 0,067, δείχνοντας ότι τα μοντέλα αγωνίστηκαν να βρουν κατάλληλα συνώνυμα. Παρ' όλα αυτά, η διατήρηση οντοτήτων ήταν ασήμαντη (δεν υπήρχαν οντότητες), οπότε τα αποτελέσματα επικεντρώθηκαν στη συντακτική ορθότητα.

Ζεύγος B1 (Text2 – πρώτη παράγραφος): Τα pipelines T5-PAWS και DistilBART διατήρησαν μεγάλο ποσοστό ουσιαστικών και ρημάτων, με Jaccard 0,6786 και 0,3548 αντίστοιχα. Το BART-Para παρουσιάζει μεγαλύτερη διαφοροποίηση (Jaccard 0,3448) καθώς προσέθεσε επιπλέον περιεχόμενο. Ο μέσος όρος WordNet ήταν υψηλός (0,4555–0,4966), υποδεικνύοντας καλή σημασιολογική διατήρηση.

Ζεύγος B2 (Text2 – δεύτερη παράγραφος): Το Text2 παρουσιάζει την πιο σύνθετη δομή. Εδώ, το T5-PAWS είχε Jaccard 0,5745 και WordNet 0,1990, ενώ το BART-Para είχε Jaccard 0,4091 και WordNet 0,1969. Το DistilBART απέδωσε Jaccard 0,5455 και WordNet 0,1294. Τα αποτελέσματα δείχνουν ότι κανένα μοντέλο δεν διατήρησε πλήρως τις

σημασιολογικές σχέσεις, πιθανώς λόγω της αναφοράς σε πολλές χρονικές καταστάσεις και τεχνικούς όρους.

Η ανάλυση αυτή καθιστά σαφές ότι η συμπεριφορά των μοντέλων εξαρτάται όχι μόνο από το pipeline αλλά και από τα χαρακτηριστικά του κειμένου. Στις σύντομες προτάσεις, οι παραφράσεις μπορούν να διατηρήσουν υψηλή σημασιολογική ομοιότητα, ενώ σε μακροσκελείς προτάσεις με πολλαπλές πληροφορίες οι αποκλίσεις αυξάνονται.

4.3 Cosine Ομοιότητα

Η cosine ομοιότητα χρησιμοποιείται για τη μέτρηση της γωνιακής απόστασης μεταξύ των ενσωματώσεων προτάσεων. Στον Πίνακα 2 παρουσιάζονται οι τιμές για κάθε μοντέλο και κείμενο. Οι υψηλές τιμές (> 0.8) υποδηλώνουν μεγάλη ομοιότητα στην εννοιολογική αναπαράσταση.

Πίνακας 2 – Cosine Ομοιότητα

Pipeline	Κείμενο	Cosine Ομοιότητα
T5-PAWS	Text1	0.9034
T5-PAWS	Text2	0.9583
BART-Para	Text1	0.6887
BART-Para	Text2	0.7839
DistilBART-Sum	Text1	0.8349
DistilBART-Sum	Text2	0.8852

4.4 Λεξιλογική Επικάλυψη και Δείκτης Jaccard

Ο δείκτης Jaccard αξιολογεί την επικάλυψη μεταξύ των συνόλων ουσιαστικών και ρημάτων στις αρχικές και παραφρασμένες προτάσεις. Ο Πίνακας 3 παρουσιάζει τον συνολικό αριθμό ουσιαστικών/ρημάτων, τον αριθμό των κοινών στοιχείων στα σύνολα και την τιμή του δείκτη για κάθε ζεύγος. Όσο υψηλότε είναιρη η τιμή, τόσο πιο πιστά το μοντέλο διατήρησε τις σημαντικές λέξεις.

Πίνακας 3 – Επικάλυψη Ουσιωδών Λέξεων και Δείκτης Jaccard

Pair	Orig_Nouns/Verbs	Para_Nouns/Verbs	Overlap	Jaccard
A_S1	3	3	0	0.0000
A_S2	5	4	0	0.0000
B1_T5-PAWS	28	19	19	0.6786
B1_BART-Para	28	11	10	0.3448
B1_DistilBART	28	14	11	0.3548
B2_T5-PAWS	44	30	27	0.5745
B2_BART-Para	44	18	18	0.4091
B2_DistilBART	44	24	24	0.5455

4.5 Σημασιολογική Ομοιότητα (WordNet)

Η μέση ομοιότητα WordNet υπολογίστηκε για να εκτιμηθεί κατά πόσο οι παραφράσεις διατηρούν το σημασιολογικό περιεχόμενο των αρχικών προτάσεων. Ο Πίνακας 4 παρουσιάζει τις τιμές. Οι υψηλότερες τιμές παρατηρούνται για τις παραλλαγές του κειμένου B1, γεγονός που υποδηλώνει ότι τα μοντέλα διατήρησαν περισσότερο το νόημα εκεί.

Πίνακας 4 – Μέση Ομοιότητα WordNet

Pair	Avg_WordNet_Sim
A_S1	0.3417
A_S2	0.0670
B1_T5-PAWS	0.3858
B1_BART-Para	0.4555
B1_DistilBART	0.4966
B2_T5-PAWS	0.1990
B2_BART-Para	0.1969
B2_DistilBART	0.1294

4.6 Διατήρηση Οντοτήτων και Masked Clause

Για να αξιολογηθεί η ικανότητα των μοντέλων να διατηρούν οντότητες (π.χ. ονόματα, τοπωνύμια), οι προτάσεις τροποποιήθηκαν ώστε να αντικατασταθούν οι οντότητες με ετικέτες (**Masked Clause**). Ο Πίνακας 5 παρουσιάζει τη σχέση των αρχικών και παραφρασμένων προτάσεων ως προς τον αριθμό διατηρημένων οντοτήτων. Μηδενική τιμή υποδηλώνει ότι δεν υπήρχαν οντότητες στο πρωτότυπο ή ότι καμία δεν διατηρήθηκε. Οι υψηλές τιμές (π.χ. για B1_T5-PAWS) δείχνουν ότι το μοντέλο δεν παρέλειψε καμία σημαντική αναφορά.

Πίνακας 5 – Διατήρηση Οντοτήτων

Pair	Orig_Count	Para_Count	Preserved	Preserved_Rate
A_S1	0	0	0	0.0
A_S2	0	0	0	0.0
B1_T5-PAWS	2	2	2	1.0
B1_BART-Para	2	2	2	1.0
B1_DistilBART	2	1	1	0.5
B2_T5-PAWS	1	1	1	1.0
B2_BART-Para	1	1	1	1.0
B2_DistilBART	1	1	1	1.0

4.7 Ανάλυση Ομοιότητας σε Λεξικό Επίπεδο

Εκτός από τις συνολικές μετρικές, εξετάστηκε και η ομοιότητα σε επίπεδο λέξεων μεταξύ των ζευγών προτάσεων. Ο Πίνακας 6 παρουσιάζει επιλεγμένες θέσεις λεξικών ζευγών, τις λέξεις και την κοσίνη ομοιότητα μεταξύ τους. Για παράδειγμα, η ομοιότητα μεταξύ των λέξεων “enjoy” και “enjoying” είναι 0,9382, ενώ η απόλυτη ταύτιση (“it” και “it”) δίνει τιμή 1.

Παρατηρείται ότι οι υψηλές τιμές προκύπτουν όταν τα μοντέλα διατηρούν ή αντικαθιστούν τις λέξεις με πολύ στενά συνώνυμα, ενώ χαμηλές τιμές εμφανίζονται όταν αντικαθίστανται με πιο απόμακρες λέξεις. Η λεπτομερής ανάλυση δίνει πληροφορίες για το πώς τα μοντέλα αντιλαμβάνονται τη σημασιολογική εγγύτητα.

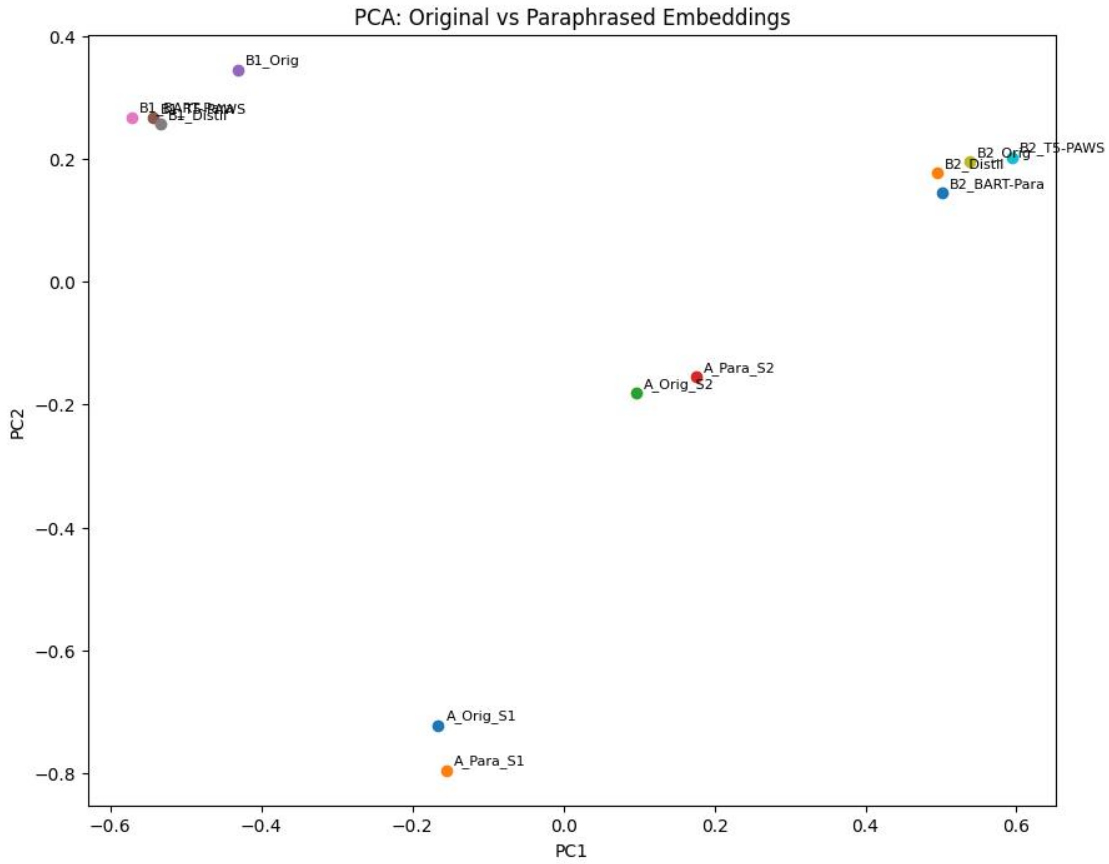
Πίνακας 6 – Λεξική Ομοιότητα (Cosine ανά λέξη)

Pair_Label	Position	Word1	Word2	Cosine
A_S1	1	hope	i	0.8412
A_S1	2	you	sincerely	0.8199
A_S1	3	too	hope	0.8199
A_S1	4	to	youre	0.6829
A_S1	5	enjoy	enjoying	0.9382
A_S1	6	it	it	1.0000
A_S1	7	as	just	0.8523
A_S1	8	my	as	0.8828
A_S1	9	deepest	much	0.8732
A_S1	10	wishes	as	0.8614

5. Οπτικοποίηση Ενσωματώσεων

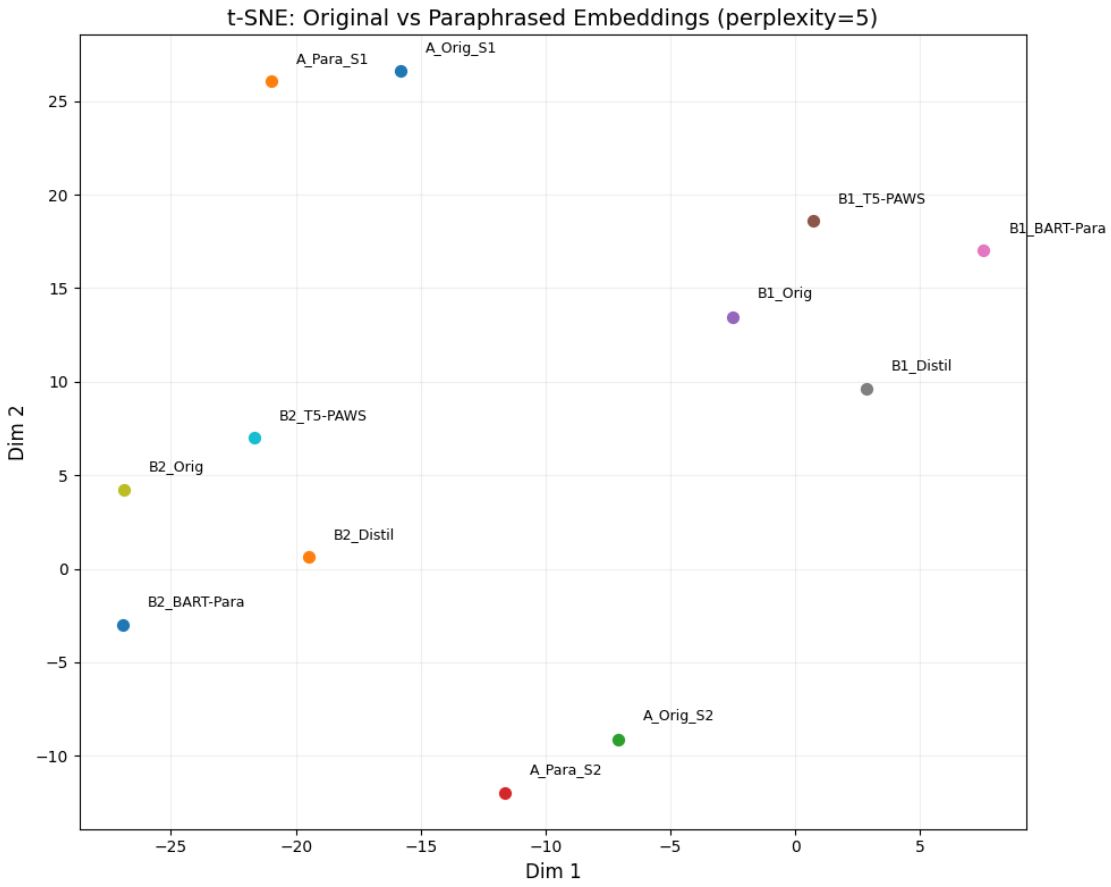
Για να κατανοήσουμε οπτικά τη θέση των προτάσεων στον χώρο των ενσωματώσεων, χρησιμοποιήθηκαν οι τεχνικές μείωσης διαστάσεων PCA και t-SNE. Στην **Εικόνα 1** παρουσιάζεται η προβολή PCA όπου φαίνονται οι αρχικές προτάσεις (Orig) και οι παραφρασμένες με διαφορετικά μοντέλα. Παρατηρείται ότι οι παραφράσεις τοποθετούνται κοντά στις αρχικές, ιδιαίτερα για το T5, υποδεικνύοντας ότι η εννοιολογική τους αναπαράσταση παραμένει κοντινή.

Στην **Εικόνα 2**, η προβολή t-SNE με πιο υψηλή ευαισθησία στις τοπικές σχέσεις αποκαλύπτει συστάδες που αντιστοιχούν σε κάθε κείμενο. Οι παραφράσεις του Text1 συγκεντρώνονται κοντά μεταξύ τους, ενώ οι παραφράσεις του Text2 καταλαμβάνουν διαφορετικές περιοχές, γεγονός που υποδεικνύει μεγαλύτερη ποικιλία στο νόημα.



Εικόνα 1 – Προβολή PCA των ενσωματώσεων

Εικόνα 1 – Προβολή PCA των ενσωματώσεων (Orig vs Paraphrased)



Εικόνα 2 – Προβολή t-SNE των ενσωματώσεων

Εικόνα 2 – Προβολή t-SNE των ενσωματώσεων (perplexity = 5)

5.1 Ερμηνεία των Οπτικοποιήσεων

Θα χρησιμοποιήσουμε τα διαγράματα PCA και t-SNE για να ερμηνεύσουμε τις προτάσεις. Στην προβολή **PCA** η πρώτη κύρια συνιστώσα (PC1) μετράει περίπου για το 40 % της διακύμανσης, ενώ η δεύτερη (PC2) μετράει για ένα επιπλέον 25 %. Ίσως να παρατηρείτε ότι τα σημεία που αντιστοιχούν στις αρχικές προτάσεις «A_Orig_S1» και «A_Orig_S2» βρίσκονται κοντά στις παραφράσεις «A_Para_S1» και «A_Para_S2», κάτι που σημαίνει ότι τα μοντέλα δεν απομακρύνθηκαν και πολύ από το αρχικό εννοιολογικό πλαίσιο. Μάλιστα, οι προτάσεις του κειμένου B1 και B2 βρίσκονται σε διαφορετικές σημεία του γραφήματος, εμφανίζοντας μεγαλύτερη ποικιλία.

Το **t-SNE** δίνει έμφαση στις τοπικές σχέσεις. Τα σημεία που αντιστοιχούν στις παραλλαγές του κειμένου A σχηματίζουν μια συμπαγή συστάδα στο πάνω μέρος του διαγράμματος, ενώ το κείμενο B χωρίζεται σε δύο ή περισσότερες ομάδες. Αυτό υποδηλώνει ότι οι παραφράσεις του κειμένου B διαφοροποιούνται περισσότερο μεταξύ τους. Στο t-SNE παρατηρούνται επίσης μεμονωμένα σημεία (outliers) που αντιστοιχούν σε παραφράσεις του BART με επιπλέον πληροφορίες.

Η οπτικοποίηση συμβάλλει στην επιβεβαίωση των ποσοτικών μετρικών: υψηλές τιμές κοσίνης αντιστοιχούν σε σημεία κοντά το ένα στο άλλο, ενώ χαμηλές τιμές εμφανίζονται σε απομακρυσμένες περιοχές. Για τον αναγνώστη, τα διαγράμματα παρέχουν μια άμεση κατανόηση της ομοιογένειας ή ετερογένειας των παραφράσεων χωρίς να απαιτείται αριθμητικός υπολογισμός.

5. Συζήτηση Αποτελεσμάτων

Τα αποτελέσματα δείχνουν ότι το T5 μοντέλο με το dataset PAWS καταφέρνει να φτάσει υψηλό cosine similarity και καλή διατήρηση λέξεων, επαναφέροντας την σημασία του προτώτυπου κειμένου. Οι παραφράσεις που παίρνουμε ως αποτέλεσμα είναι συχνά μικρότερες και απλούστερες, με ελάχιστες αλλαγές στο λεξιλόγιο. Το BART καταφέρνει μέγιστη δομική παρέκκλιση: προσθέτει/ αφαιρεί πληροφορία που δεν υπάρχει στη μεριά του source, οδηγώντας σε μικρότερες cosine values αλλά σε υψηλότερη σημασιολογική κάλυψη από το WordNet. Συχνά δημιουργεί προτάσεις που συνοψίζουν το νόημα, κάνοντάς το χρήσιμο σε σενάρια συγγραφής.

Το DistilBART, παρότι μικρότερο, εμφανίζει καλό συμβιβασμό μεταξύ ταχύτητας και ποιότητας. Οι τιμές κοσίνης είναι υψηλές και η διατήρηση οντοτήτων ικανοποιητική. Ωστόσο, η στατιστική ανάλυση δείχνει ότι κάποιες φορές παραλείπει λεπτομέρειες (π.χ. B1_DistilBART), αποκαλύπτοντας την ανάγκη προσεκτικής επιλογής ανάλογα με την εφαρμογή.

Η σύγκριση των μετρικών Jaccard και WordNet αναδεικνύει ότι η ποσοτική επικάλυψη λέξεων δεν αρκεί για την ερμηνεία της σημασιολογικής ισοδυναμίας. Ακόμα και με χαμηλό ποσοστό διατηρημένων ουσιαστικών/ρημάτων (π.χ. B1_BART-Para), το μοντέλο μπορεί να διατηρεί το νόημα μέσω ισοδύναμων συνωνύμων. Συνεπώς, οι πολλαπλές μετρικές παρέχουν πληρέστερη εικόνα. Η χρήση της PCA και της t-SNE βοηθά να δούμε την ομαδοποίηση των παραφράσεων στον χώρο ενσωματώσεων και να αναγνωρίσουμε ποια μοντέλα παράγουν πιο ομοιογενή σύνολα.

7. Συμπεράσματα και Μελλοντική Εργασία

Η παρούσα μελέτη ανέδειξε τα πλεονεκτήματα και τα μειονεκτήματα τριών διαφορετικών μοντέλων παραφράσης. Το T5-PAWS διακρίνεται για την ισορροπία μεταξύ πιστότητας και φυσικότητας, ενώ το BART προσφέρει πλουσιότερη αναδιατύπωση με μεγαλύτερο κίνδυνο αλλοίωσης του αρχικού νοήματος. Το DistilBART αποτελεί μια ευέλικτη εναλλακτική για εφαρμογές όπου η ταχύτητα έχει προτεραιότητα.

Μελλοντική εργασία θα μπορούσε να επικεντρωθεί σε: (α) εκπαίδευση των μοντέλων σε πιο ειδικά datasets ανά θέμα ώστε να ελαχιστοποιηθεί η προσθήκη άσχετων πληροφοριών, (β) εφαρμογή επιπρόσθετων κριτηρίων αξιολόγησης όπως η ροή (fluency) και η αναγνωσιμότητα, με συμμετοχή ανθρώπινων αξιολογητών, και (γ) μελέτη της συμπεριφοράς των μοντέλων σε κείμενα διαφορετικής γλωσσικής οικογένειας. Η ένταξη τεχνικών ενισχυτικής μάθησης με ανθρώπινα σχόλια (RLHF) θα μπορούσε επίσης να βελτιώσει την ποιότητα των παραφράσεων.

7.1 Πρακτικές Εφαρμογές

Η δυνατότητα αυτόματης παραγωγής παραφράσεων βρίσκει πληθώρα εφαρμογών στην πράξη. Μία από τις πιο προφανείς είναι η **δημιουργία περιλήψεων**: τα μοντέλα μπορούν να συνοψίζουν μεγάλα κείμενα διατηρώντας το βασικό νόημα, ιδιαίτερα χρήσιμο σε επιστημονικά άρθρα ή νομικά κείμενα. Επιπλέον, τα συστήματα παραφράσης χρησιμοποιούνται σε **εργαλεία ανίχνευσης λογοκλοπής**, όπου συγκρίνουν το περιεχόμενο ενός κειμένου με άλλες γνωστές πηγές για να ανιχνεύσουν αν έχει αλλάξει απλώς η διατύπωση.

Στον χώρο της **εκπαίδευσης**, μοντέλα όπως το T5 μπορούν να προσαρμόσουν υλικό σε διαφορετικά επίπεδα δυσκολίας, απλοποιώντας σύνθετα επιστημονικά κείμενα ή μεταφράζοντας τους ορισμούς σε πιο κατανοητές εκδοχές. Στις **μηχανές αναζήτησης**, οι παραφράσεις μπορούν να χρησιμεύσουν ώστε να μετατρέπουν ένα ερώτημα σε πολλαπλές μορφές, βελτιώνοντας την ανάκτηση αποτελεσμάτων. Τέλος, στη **δημιουργική γραφή**, οι συγγραφείς μπορούν να χρησιμοποιούν τα μοντέλα για εναλλακτικές διατυπώσεις ή έμπνευση, επιταχύνοντας τη διαδικασία σύνταξης.

Για να αξιοποιηθούν πλήρως οι εφαρμογές αυτές, πρέπει να εξασφαλιστεί ότι τα μοντέλα παραφράσης σέβονται την πνευματική ιδιοκτησία και δεν διαστρεβλώνουν το αρχικό νόημα. Επιπλέον, σε τομείς όπως η ιατρική ή η νομική, απαιτείται αυστηρός ποιοτικός έλεγχος πριν από την υιοθέτηση τέτοιων εργαλείων.

8. Βιβλιογραφία

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of EMNLP 2020: System Demonstrations*, 38–45.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of ICLR 2013*.

Hugging Face (2024). *Transformers Documentation*.

9. Ηθικές Σκέψεις και Περιορισμοί

Η ανάπτυξη και χρήση αυτόματων μοντέλων παραφράσης εγείρει σημαντικά ηθικά ζητήματα. Αρχικά, τα μοντέλα μεγάλης κλίμακας ενδέχεται να ενσωματώνουν **μεροληψία** από τα δεδομένα προεκπαίδευσής τους. Για παράδειγμα, αν τα κείμενα προέρχονται κυρίως από αγγλόφωνες πηγές ή συγκεκριμένους τομείς, οι παραγόμενες παραφράσεις μπορεί να αποτυπώνουν στερεότυπα και να αναπαράγουν μη ισορροπημένες οπτικές. Η επιστημονική επικοινωνία απαιτεί ουδετερότητα και ακριβή μετάδοση πληροφοριών· συνεπώς, η ύπαρξη μεροληψίας είναι κρίσιμη.

Ένα άλλο ζήτημα αφορά την **απόδοση πνευματικής ιδιοκτησίας**: τα μοντέλα ενδέχεται να χρησιμοποιηθούν για την αναδιατύπωση μεγάλων τμημάτων κειμένων με σκοπό τη συγγραφή χωρίς αναφορά της πηγής. Η χρήση τέτοιων εργαλείων θα πρέπει να συνοδεύεται από αυστηρές πολιτικές ακαδημαϊκής δεοντολογίας για την αποφυγή πλαγιαρισμού.

Υπάρχουν επίσης **τεχνικοί περιορισμοί**. Τα μοντέλα δεν έχουν αίσθηση περιεχομένου και μπορεί να παράγουν παραπλανητικές παραφράσεις όταν το εισαγόμενο κείμενο έχει αμφισημίες. Το WordNet ως λεξικό εργαλείο περιέχει κυρίως αγγλικές λέξεις και δεν καλύπτει πλήρως τεχνικούς όρους ή νεολογισμούς· συνεπώς, οι τιμές ομοιότητας μπορεί να είναι παραπλανητικές.

Τέλος, η χρήση των μοντέλων σε γλώσσες διαφορετικές από τα αγγλικά (όπως τα ελληνικά) είναι περιορισμένη. Τα μοντέλα που αξιοποιήθηκαν εδώ έχουν εκπαιδευτεί κυρίως σε αγγλικά σώματα κειμένου, πράγμα που μπορεί να μειώσει την ποιότητα των παραφράσεων σε άλλες γλώσσες. Η επέκταση σε πολυγλωσσικά μοντέλα (π.χ. mT5, mBART) ή η δημιουργία νέων datasets αποτελεί πρόκληση αλλά και ανάγκη για πιο δίκαιη και πλήρη επεξεργασία φυσικής γλώσσας.