ΑΠΑΛΛΑΚΤΙΚΗ ΕΡΓΑΣΙΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

ΠΑΡΑΔΟΤΕΟ 3: ΔΟΜΗΜΕΝΗ ΑΝΑΦΟΡΑ

ΟΜΑΔΑ:

ΑΝΝΑ ΚΙΑΜΙΛΗ, Π22069

ΙΩΑΝΝΗΣ ΣΥΡΙΑΝΟΣ, Π22166

ΓΕΩΡΓΙΟΣ ΑΛΕΞΟΠΟΥΛΟΣ, Π22008

Repository στο github: https://github.com/annakiamili/NLP_assignment

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή 4
ΜΕΘΟΔΟΛΟΓΙΑ 5
ΕΡΩΤΗΜΑ Α: ΠΑΡΑΦΡΑΣΕΙΣ ΠΡΟΤΑΣΕΩΝ 5
Στρατηγικές ανακατασκευής: 5
Υπολογιστικές τεχνικές: 5
ΕΡΩΤΗΜΑ Β: ΠΑΡΑΦΡΑΣΕΙΣ ΠΛΗΡΩΝ ΚΕΙΜΕΝΩΝ 6
Στρατηγικές ανακατασκευής: 6
Υπολογιστικές τεχνικές: 6
ΕΡΩΤΗΜΑ C: ΜΕΤΡΗΣΗ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΟΜΟΙΟΤΗΤΑΣ 6
Στρατηγικές ανακατασκευής: 6
Υπολογιστικές τεχνικές: 6
ΠΕΙΡΑΜΑΤΑ & ΑΠΟΤΕΛΕΣΜΑΤΑ 7
(1)ΠΑΡΑΔΕΙΓΜΑΤΑ ΠΡΙΝ/ΜΕΤΑ ΤΗΝ ΑΝΑΚΑΤΑΣΚΕΥΗ 7
ΠΑΡΑΔΕΙΓΜΑ 1- ΕΡΩΤΗΜΑ Α 7
ΠΑΡΑΔΕΙΓΜΑ 2- ΕΡΩΤΗΜΑ Β 7
(2) ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΑΡΑΔΟΤΕΟΥ 2 8
(2.1)ΑΠΟΤΕΛΕΣΜΑΤΑ COSINE SIMILARITY 8
(2.2) ΟΠΤΙΚΟΠΟΙΗΣΗ(PCA και t-SNE) 8
ΣΥΖΗΤΗΣΗ 12
(1) Πόσο καλά αποτύπωσαν οι ενσωματώσεις λέξεων το νόημα; 12
(2) Ποιες ήταν οι μεγαλύτερες προκλήσεις στην ανακατασκευή; 12
(3) Πώς μπορεί να αυτοματοποιηθεί αυτή η διαδικασία χρησιμοποιώντας μοντέλα NLP; 12
(4) Υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων, βιβλιοθηκών; 13
(5) Συζήτηση Ευρημάτων 13
ΣΥΜΠΕΡΑΣΜΑ 13
BONUS- MASKED CLAUSE INPUT 14
ΜΕΘΟΔΟΛΟΓΙΑ 14
ΑΠΟΤΕΛΕΣΜΑΤΑ 14

ΣΥΖΗΤΗΣΗ 14
ΣΥΜΠΕΡΑΣΜΑ BONUS 14
ΒΙΒΛΙΟΓΡΑΦΙΑ 15
ΠΑΡΑΔΟΤΕΟ 1 – Paraphrasing (Pegasus, T5, BART, Similarity) 15
ΠΑΡΑΔΟΤΕΟ 2- Word Embeddings (Word2Vec, FastText, BERT, PCA/t-SNE) 15
Bonus (Greek-BERT, Masked Clause Input) 15

Εισαγωγή

Η σημασιολογική ανακατασκευή (semantic reconstruction ή paraphrasing) αποτελεί έναν από τους πιο σημαντικούς τομείς στην Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP). Ο στόχος της είναι η δημιουργία νέων εκδοχών ενός κειμένου που εκφράζουν το ίδιο νόημα με διαφορετικό λεξιλόγιο ή συντακτική δομή. Η ικανότητα αυτή είναι κρίσιμη τόσο για την επιστημονική έρευνα όσο και για πρακτικές εφαρμογές, καθώς η γλώσσα είναι από τη φύση της πολυδιάστατη και συχνά απαιτείται η κατανόηση ή η παραγωγή κειμένου πέρα από την απλή αντιστοίχιση λέξεων.

Οι εφαρμογές της σημασιολογικής ανακατασκευής είναι πολυάριθμες:

- Αυτόματη περίληψη: η συμπύκνωση μεγάλων κειμένων σε μικρότερες, ευανάγνωστες εκδοχές.
- Μετάφραση: η δυνατότητα εναλλακτικών διατυπώσεων βοηθά στην παραγωγή πιο φυσικών μεταφράσεων.
- Ανίχνευση λογοκλοπής: οι παραφράσεις χρησιμοποιούνται για την αξιολόγηση της ομοιότητας διαφορετικών εκδοχών του ίδιου περιεχομένου.
- Εκπαίδευση μοντέλων (data augmentation): η δημιουργία παραφρασμένων δεδομένων αυξάνει την ποικιλία και βελτιώνει την απόδοση μοντέλων NLP.
- Συστήματα ερωταποκρίσεων και συνομιλίας: η κατανόηση εναλλακτικών διατυπώσεων ερωτήσεων ενισχύει την ακρίβεια των απαντήσεων.

Η ραγδαία εξέλιξη των νευρωνικών μοντέλων γλώσσας έχει επιτρέψει σημαντικές προόδους στον τομέα της παραφραστικής παραγωγής. Μοντέλα όπως το PEGASUS (Zhang et al., 2020), που σχεδιάστηκε ειδικά για paraphrasing και text summarization, το T5 (Raffel et al., 2020), το οποίο γενικεύει πλήθος εργασιών NLP σε ενιαίο πλαίσιο, και το BART (Lewis et al., 2020), που συνδυάζει τις αρχές autoencoding και autoregressive γλωσσικών μοντέλων, έχουν ανεβάσει σημαντικά την ποιότητα των ανακατασκευασμένων κειμένων. Μέσω της αρχιτεκτονικής Transformers, τα μοντέλα αυτά έχουν εκπαιδευτεί σε τεράστια σώματα κειμένων και έχουν τη δυνατότητα να κατανοούν συμφραζόμενα, να αποδίδουν νοηματικές αποχρώσεις και να παράγουν φυσική γλώσσα με υψηλή συνοχή.

Η πλατφόρμα HuggingFace Transformers έχει καταστήσει την πρόσβαση σε αυτά τα μοντέλα ιδιαίτερα εύκολη, επιτρέποντας την άμεση ενσωμάτωσή τους σε πειράματα και εφαρμογές. Μέσα από απλές κλήσεις API, οι ερευνητές και οι φοιτητές μπορούν να φορτώσουν state-of-the-art μοντέλα όπως το PEGASUS, το T5 και το BART και να εκτελέσουν paraphrasing σε προτάσεις και κείμενα, επιταχύνοντας σημαντικά τη διαδικασία αξιολόγησης διαφορετικών τεχνικών.

Η αξιολόγηση της ποιότητας των ανακατασκευών δεν μπορεί να περιορίζεται σε επιφανειακές μετρικές, όπως η ακριβής αντιστοίχιση λέξεων. Η πραγματική πρόκληση βρίσκεται στη σημασιολογική ομοιότητα. Για τον σκοπό αυτό, έχουν αναπτυχθεί μέθοδοι που αξιοποιούν ενσωματώσεις προτάσεων (sentence embeddings) ώστε να αποτυπώσουν τη σημασία σε

πολυδιάστατο χώρο. Οι Sentence-Transformers (Reimers & Gurevych, 2019), βασισμένες σε προ εκπαιδευμένα μοντέλα τύπου BERT, επιτρέπουν τον υπολογισμό ομοιοτήτων με μετρικές όπως το cosine similarity, προσφέροντας μια πιο ουσιαστική και αξιόπιστη μέτρηση του κατά πόσο η παραφρασμένη εκδοχή διατηρεί το νόημα του αρχικού κειμένου.

Παράδειγμα εφαρμογής αυτής της διαδικασίας είναι η σύγκριση δύο προτάσεων:

- Original: *The doctor edited the acknowledgments section before sending the article.*
- Paraphrased: *Before submitting the article, the doctor revised the acknowledgments part.*

Παρότι η διατύπωση είναι διαφορετική, η σημασία παραμένει ίδια. Με τη χρήση embeddings, οι δύο προτάσεις χαρτογραφούνται σε κοντινά σημεία στον διανυσματικό χώρο, αποδεικνύοντας ότι η σημασιολογική εγγύτητα διατηρείται.

Συνοψίζοντας, η σημασιολογική ανακατασκευή συνδυάζει την πρόοδο της παραγωγής φυσικής γλώσσας μέσω μοντέλων όπως το PEGASUS, το T5 και το BART, με την πρόοδο της σημασιολογικής αποτίμησης μέσω embeddings και similarity μετρικών. Η έρευνα και η πρακτική στον τομέα αυτόν έχουν άμεσες επιπτώσεις τόσο στη βελτίωση των ακαδημαϊκών μελετών όσο και στην ανάπτυξη σύγχρονων εφαρμογών NLP, από συστήματα υποβοήθησης συγγραφής έως έξυπνους βοηθούς.

ΜΕΘΟΔΟΛΟΓΙΑ

ΕΡΩΤΗΜΑ Α: ΠΑΡΑΦΡΑΣΕΙΣ ΠΡΟΤΑΣΕΩΝ

Στρατηγικές ανακατασκευής:

Στο πρώτο ερώτημα δόθηκε έμφαση στη δημιουργία εναλλακτικών εκδοχών απλών προτάσεων. Οι στρατηγικές ανακατασκευής περιλάμβαναν κυρίως συντακτικούς και λεξιλογικούς μετασχηματισμούς, δηλαδή αλλαγή της σειράς λέξεων, αντικατάσταση λέξεων με συνώνυμα ή ομόρριζες μορφές, καθώς και τροποποίηση της γραμματικής δομής χωρίς απώλεια του αρχικού νοήματος. Το μοντέλο PEGASUS εφαρμόστηκε ως γλωσσικό εργαλείο που έχει εκπαιδευτεί ειδικά σε εργασίες παραφραστικής περίληψης και μπορεί να αναδιατυπώνει προτάσεις σε διαφορετικές μορφές διατηρώντας τη σημασιολογική τους υπόσταση.

Υπολογιστικές τεχνικές:

Σε αυτό το στάδιο η αξιολόγηση έγινε ποιοτικά, με έλεγχο αν οι προτάσεις που παρήχθησαν ήταν νοηματικά κοντινές με τις αρχικές. Δεν χρησιμοποιήθηκε ακόμη κάποια μετρική, αλλά στηρίχθηκε στη γλωσσική ανάλυση του μοντέλου PEGASUS και στη θεωρητική ικανότητά του να παράγει paraphrases. Τα αποτελέσματα αποτέλεσαν τη βάση για περαιτέρω ποσοτική σύγκριση στα επόμενα βήματα.

ΕΡΩΤΗΜΑ Β: ΠΑΡΑΦΡΑΣΕΙΣ ΠΛΗΡΩΝ ΚΕΙΜΕΝΩΝ

Στρατηγικές ανακατασκευής:

Το δεύτερο ερώτημα εστίασε στην ανακατασκευή μεγαλύτερων κειμένων, όπου οι στρατηγικές δεν περιορίζονται μόνο σε απλούς συντακτικούς μετασχηματισμούς, αλλά επεκτείνονται σε πιο πολύπλοκες διαδικασίες, όπως η αναδιάρθρωση παραγράφων, η συμπύκνωση ή επέκταση περιεχομένου και η επιλογή εναλλακτικών εκφράσεων. Για να αποτυπωθούν οι διαφορετικές στρατηγικές, χρησιμοποιήθηκαν τρία διαφορετικά μοντέλα paraphrasing: PEGASUS (ειδικό για paraphrasing/summarization), T5 (γενικευμένο text-to-text μοντέλο) και BART (συνδυαστικό encoder-decoder). Με αυτό τον τρόπο φάνηκε πώς διαφορετικά μοντέλα προσεγγίζουν το ίδιο κείμενο: άλλο με τάση προς περίληψη, άλλο με πιο πιστή αναδιατύπωση, άλλο με μεγαλύτερη ελευθερία σύνθεσης.

Υπολογιστικές τεχνικές:

Η εφαρμογή των μοντέλων έγινε με τη χρήση του HuggingFace Transformers pipeline, όπου τα κείμενα εισάγονταν ως είσοδοι και παραγόταν η αντίστοιχη ανακατασκευή. Σε αυτό το στάδιο δεν εφαρμόστηκαν ακόμη αριθμητικές μετρικές, αλλά έγινε παρατήρηση των διαφορών στα outputs, με βάση τη γλωσσική δομή και την πληρότητα. Σημαντικό σημείο αποτέλεσε ο περιορισμός των tokens, όπου μοντέλα όπως το T5 και το BART είχαν δυσκολίες στην επεξεργασία πολύ μεγάλων εισόδων.

ΕΡΩΤΗΜΑ C: ΜΕΤΡΗΣΗ ΣΗΜΑΣΙΟΛΟΓΙΚΗΣ ΟΜΟΙΟΤΗΤΑΣ

Στρατηγικές ανακατασκευής:

Στο τρίτο ερώτημα δεν παράχθηκαν νέες παραφράσεις, αλλά στόχος ήταν να εκτιμηθεί η ποιότητα των ήδη παραχθέντων αποτελεσμάτων από τα (A) και (B). Η στρατηγική σε αυτό το στάδιο ήταν η αξιολόγηση της “εγγύτητας” ανάμεσα σε original και paraphrased κείμενα, με έμφαση στη σημασιολογία και όχι στη μορφολογία.

Υπολογιστικές τεχνικές:

Για τη μέτρηση χρησιμοποιήθηκαν οι Sentence-Transformers (MiniLM), που μετατρέπουν προτάσεις/κείμενα σε πολυδιάστατους διανυσματικούς χώρους (embeddings). Η ομοιότητα μεταξύ original και paraphrased εκδοχών μετρήθηκε μέσω της συνάφειας συνημίτονου (cosine similarity). Όσο μεγαλύτερη ήταν η τιμή (πιο κοντά στο 1), τόσο πιο κοντά ήταν τα νοήματα. Επιπλέον, για οπτικοποίηση χρησιμοποιήθηκε η τεχνική Principal Component Analysis (PCA) ώστε να μειωθούν οι διαστάσεις και να φανεί γραφικά αν τα παραφρασμένα κείμενα τοποθετούνται κοντά στα αρχικά στον σημασιολογικό χώρο.

ΠΕΙΡΑΜΑΤΑ & ΑΠΟΤΕΛΕΣΜΑΤΑ

(1)ΠΑΡΑΔΕΙΓΜΑΤΑ ΠΡΙΝ/ΜΕΤΑ ΤΗΝ ΑΝΑΚΑΤΑΣΚΕΥΗ

Για να αποτυπωθεί η διαδικασία της σημασιολογικής ανακατασκευής, παρουσιάζονται παρακάτω ενδεικτικά παραδείγματα από τα ερωτήματα Α και Β του Παραδοτέου 1. Τα παραδείγματα δείχνουν πώς τα μοντέλα paraphrasing (PEGASUS, T5, BART) μπορούν να αποδώσουν διαφορετικές εκδοχές των ίδιων προτάσεων και κειμένων, διατηρώντας το βασικό νόημα.

ΠΑΡΑΔΕΙΓΜΑ 1- ΕΡΩΤΗΜΑ Α

- Original: Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives.
- Rephrased: Today is the dragon boat festival in our Chinese culture and it is a great day to celebrate.

Η παραπάνω ανακατασκευή δείχνει τη δυνατότητα του PEGASUS να δημιουργεί μια πιο φυσική και συνοπτική διατύπωση, χωρίς να χάνει το βασικό νόημα της αρχικής πρότασης. Ο μετασχηματισμός γίνεται κυρίως σε επίπεδο λεξιλογίου και σύνταξης.

ΠΑΡΑΔΕΙΓΜΑ 2- ΕΡΩΤΗΜΑ Β

Original: During our final discuss, I told him about the new submission — the one we were waiting since last autumn, but the updates was confusing as it not included the full feedback from reviewer or maybe editor?

Anyway, I believe the team, although bit delay and less communication at recent days, they really tried best for paper and cooperation.

We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link came finally last week, I think.

Also, kindly remind me please, if the doctor still plan for the acknowledgments section edit before he sending again.

Because I didn’t see that part final yet, or maybe I missed, I apologize if so.

Overall, let us make sure all are safe and celebrate the outcome with strong coffee and future targets.

- [T5] Rephrased: During our final discussion, I told him about the new submission — the one we were waiting for since last autumn , but the updates were confusing as it did not include the full feedback from reviewer or maybe editor? Anyway, I believe the team, although a bit delay and less communication at recent days, really tried for paper and cooperation . We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link finally came last week , I think . Also, kindly remind me if the

- [BART] Rephrased: During our final discuss, I told him about the new submission -- the one we were waiting since last autumn, but the updates was confusing as it not included the full feedback from the reviewer or maybe editor, I apologize if so.

Η σύγκριση των παραπάνω παραφράσεων δείχνει ότι το T5 τείνει να διορθώνει γλωσσικά λάθη(discuss=> discussion) και να αποδίδει ένα πιο καθαρό ύφος, ενώ το BART κρατά μεγαλύτερη πιστότητα στην αρχική μορφή, αλλά προσθέτει επεξηγηματικά στοιχεία (I apologize if so). Αυτό αναδεικνύει ότι διαφορετικά μοντέλα υλοποιούν διαφορετικές στρατηγικές ανακατασκευής.

(2) ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΑΡΑΔΟΤΕΟΥ 2

Στο δεύτερο μέρος της εργασίας, η αξιολόγηση των παραγόμενων paraphrases έγινε ποσοτικά με τη χρήση διαφορετικών τεχνικών ενσωμάτωσης προτάσεων. Εφαρμόστηκαν τα μοντέλα **Word2Vec**, **FastText** και **BERT (MiniLM)**, και υπολογίστηκαν οι τιμές **cosine similarity** μεταξύ original και paraphrased εκδοχών. Επιπλέον, εφαρμόστηκαν οι τεχνικές **Principal Component Analysis (PCA)** και **t-SNE** για οπτικοποίηση των σχέσεων στον σημασιολογικό χώρο.

(2.1)ΑΠΟΤΕΛΕΣΜΑΤΑ COSINE SIMILARITY

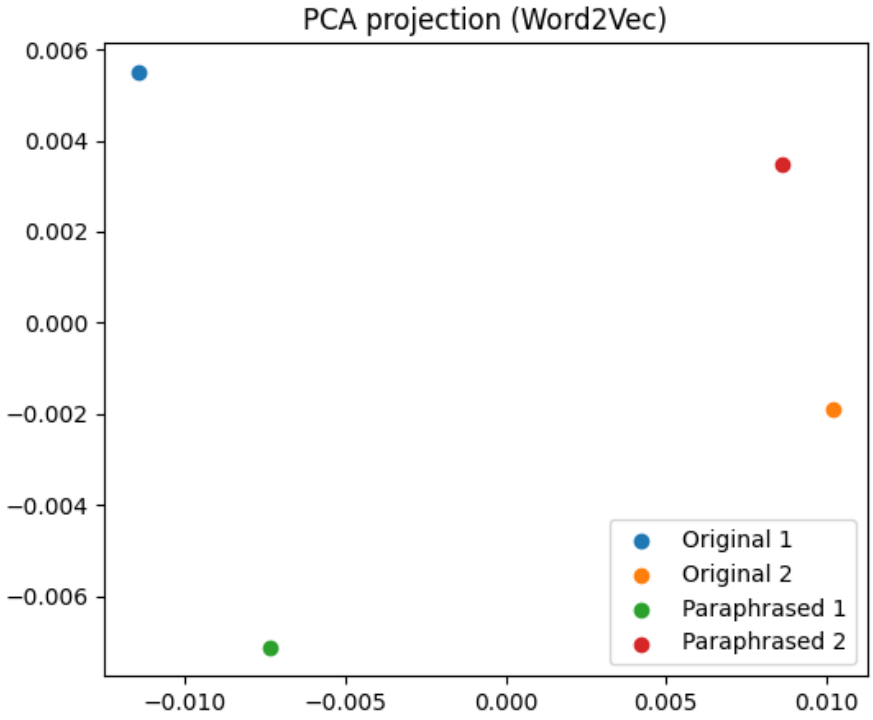
Μέθοδος	Pair 1	Pair 2
Word2Vec	0.5616	0.7756
FastText	0.7002	0.7655
BERT	0.9496	0.9474

Τα αποτελέσματα δείχνουν ότι το BERT αποδίδει τις υψηλότερες τιμές ομοιότητας, αποτυπώνοντας πιο αξιόπιστα τη σημασιολογική σχέση μεταξύ original και paraphrased εκδοχών. Αυτό ήταν αναμενόμενο, καθώς το BERT αξιοποιεί contextual embeddings, τα οποία λαμβάνουν υπόψη τα συμφραζόμενα κάθε λέξης. Αντίθετα, οι τιμές των Word2Vec και FastText είναι χαμηλότερες, γεγονός που συνδέεται με την εξάρτησή τους από το μέγεθος και την ποιότητα του corpus εκπαίδευσης. Παρόλα αυτά, το FastText υπερέχει έναντι του Word2Vec χάρη στη δυνατότητα χειρισμού μορφολογικών παραλλαγών.

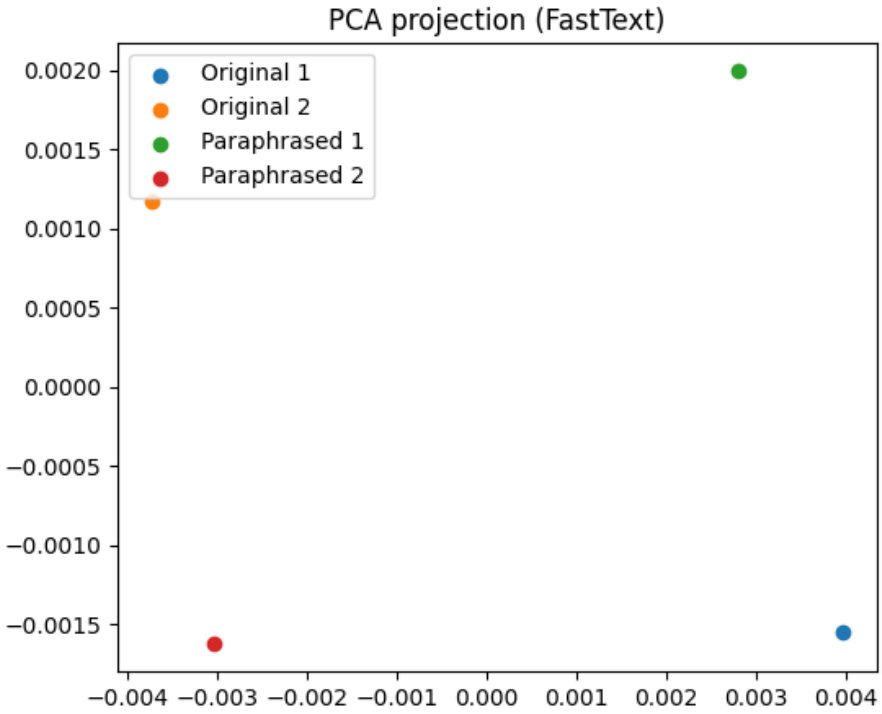
(2.2) ΟΠΤΙΚΟΠΟΙΗΣΗ(PCA και t-SNE)

Για την καλύτερη κατανόηση των αποτελεσμάτων χρησιμοποιήθηκαν γραφήματα PCA και t-SNE, τα οποία απεικονίζουν τα original και paraphrased κείμενα στον σημασιολογικό χώρο.

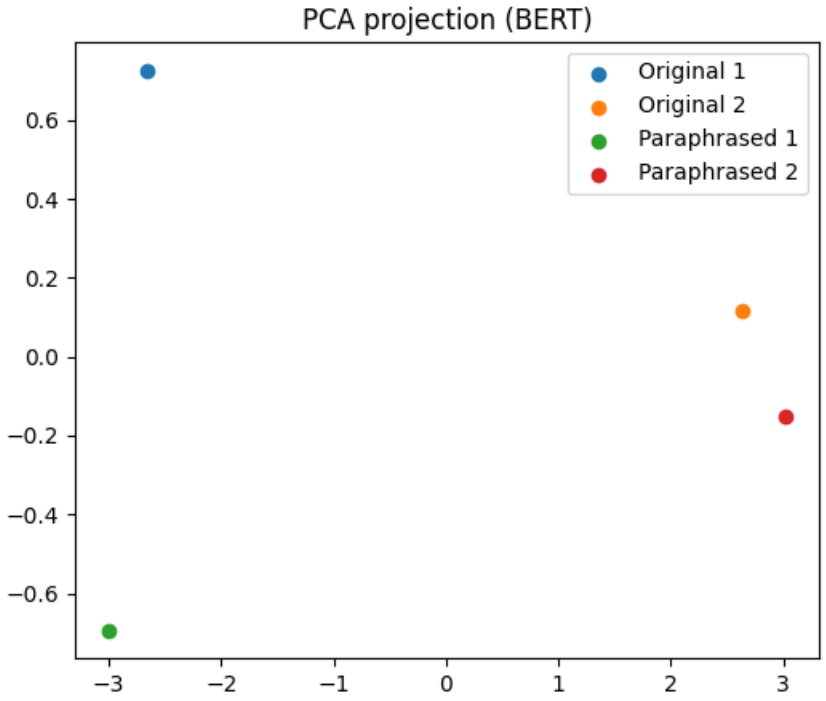
PCA(Word2Vec):



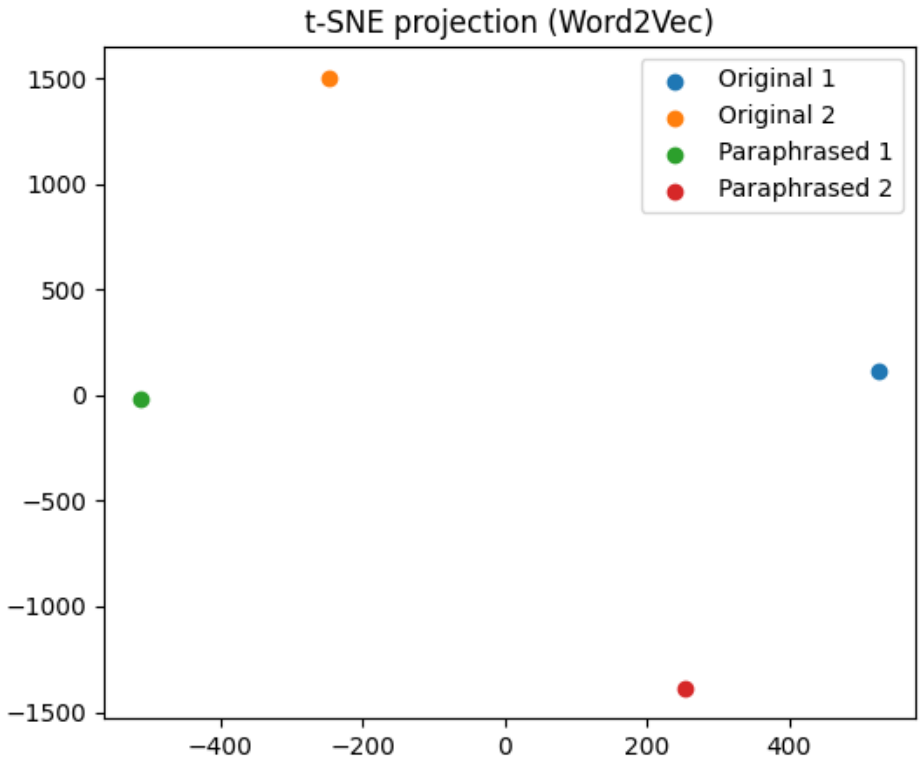
PCA(FastText):



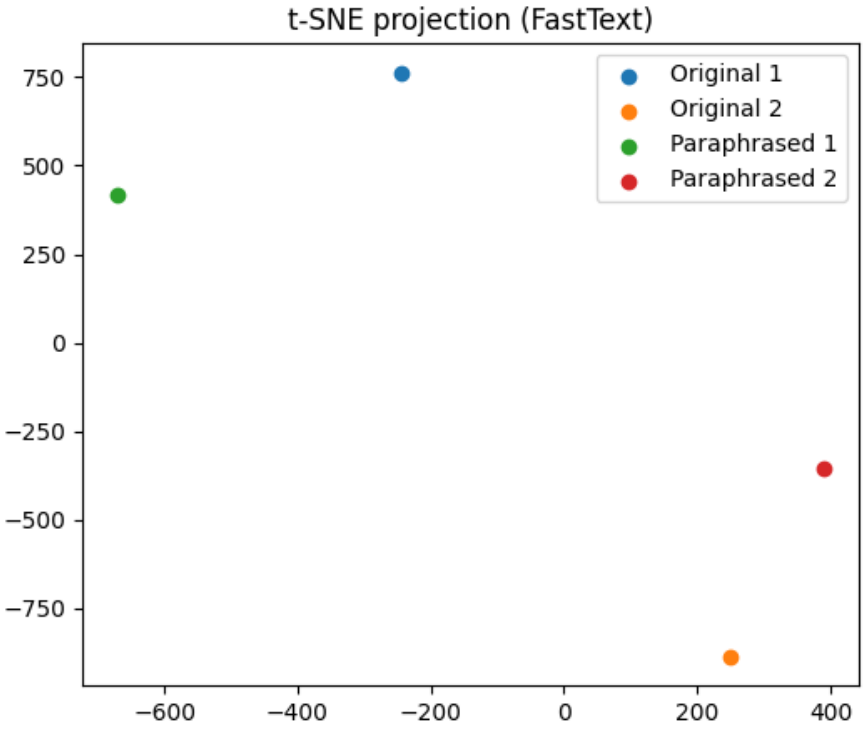
PCA(BERT):



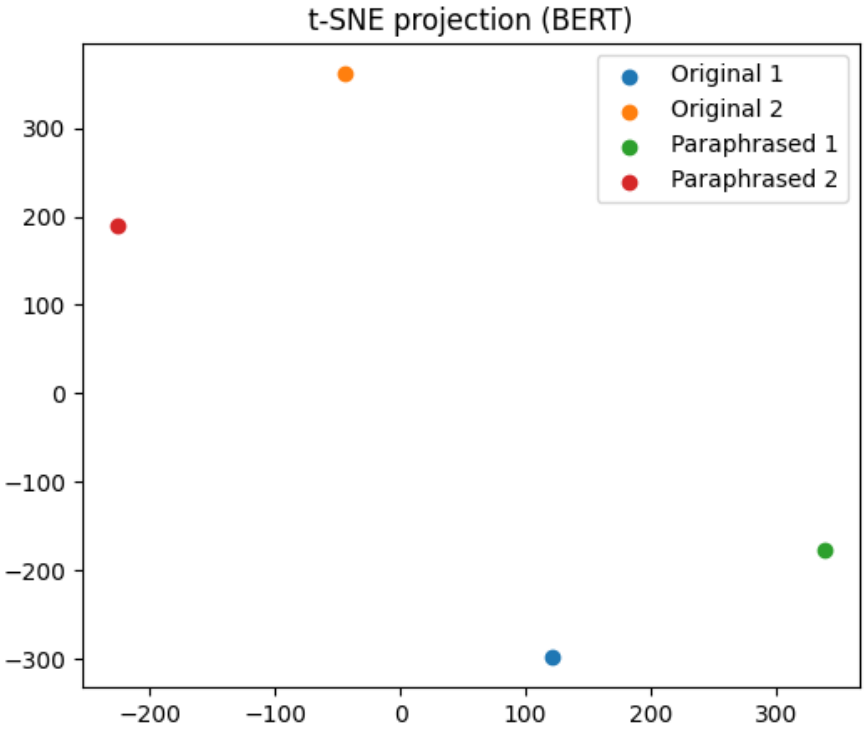
t-SNE(Word2Vec):



t-SNE(FastText):



t-SNE(BERT):



Από τα γραφήματα παρατηρείται ότι τα ζεύγη original–paraphrased βρίσκονται πολύ πιο κοντά το ένα στο άλλο όταν χρησιμοποιείται το BERT, γεγονός που επιβεβαιώνει την αριθμητική ανάλυση του πίνακα. Αντίθετα, στο Word2Vec και το FastText οι αναπαραστάσεις δείχνουν μεγαλύτερη απόσταση, στοιχείο που καταδεικνύει ότι τα παραδοσιακά embeddings αποτυπώνουν λιγότερο ακριβώς τις σημασιολογικές σχέσεις. Η χρήση PCA και t-SNE ενίσχυσε την κατανόηση των διαφορών, παρουσιάζοντας καθαρά την ανωτερότητα του BERT.

ΣΥΖΗΤΗΣΗ

(1) Πόσο καλά αποτύπωσαν οι ενσωματώσεις λέξεων το νόημα;

Οι ενσωματώσεις λέξεων λειτούργησαν με διαφορετικά επίπεδα επιτυχίας. Τα αποτελέσματα έδειξαν ότι τα παραδοσιακά μοντέλα (Word2Vec, FastText) αποτύπωσαν σε ικανοποιητικό βαθμό τη σχέση original–paraphrased, αλλά με σαφή όρια: οι τιμές cosine similarity ήταν μέτριες και δεν απεικόνιζαν πλήρως τη σημασιολογική ταύτιση. Αντίθετα, το BERT, αξιοποιώντας contextual embeddings, παρείχε πολύ υψηλότερες τιμές και τοποθέτησε τα paraphrased κείμενα σχεδόν πάνω στα original στα PCA/t-SNE γραφήματα. Αυτό δείχνει ότι τα σύγχρονα context-based μοντέλα έχουν σαφή υπεροχή στην κατανόηση του νοήματος.

(2) Ποιες ήταν οι μεγαλύτερες προκλήσεις στην ανακατασκευή;

Η βασικότερη πρόκληση αφορούσε τον έλεγχο της **πιστότητας του νοήματος** στις παραφράσεις. Κάποια μοντέλα (π.χ. BART) παρήγαγαν σωστά γλωσσικά κείμενα, αλλά μερικές φορές πρόσθεταν φράσεις που δεν υπήρχαν στο original. Επιπλέον, οι **περιορισμοί μήκους (token limits)** αποτέλεσαν πρόβλημα για μεγάλα κείμενα, όπου οι παραφράσεις κόβονταν ή γίνονταν πιο αποσπασματικές. Μια ακόμη πρόκληση ήταν η **αξιολόγηση**: η απλή σύγκριση λέξεων δεν επαρκεί, γι' αυτό χρειάστηκε να χρησιμοποιηθούν embeddings και similarity μετρικές.

(3) Πώς μπορεί να αυτοματοποιηθεί αυτή η διαδικασία χρησιμοποιώντας μοντέλα NLP;

Η διαδικασία μπορεί να αυτοματοποιηθεί πλήρως με pipelines όπως αυτά που προσφέρει η **HuggingFace Transformers**. Με απλές κλήσεις API είναι δυνατή η αυτόματη δημιουργία παραφράσεων για προτάσεις ή κείμενα, καθώς και η αξιολόγησή τους με embeddings και cosine similarity (Sentence-Transformers). Σε πιο προχωρημένες υλοποιήσεις, μπορεί να στηθεί μια αλυσίδα όπου:

- το κείμενο μπαίνει ως είσοδος,
- δημιουργούνται πολλαπλές παραφράσεις,
- υπολογίζονται αυτόματα οι βαθμοί ομοιότητας,
- και τελικά επιλέγεται η "βέλτιστη" ανακατασκευή.

(4) Υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων, βιβλιοθηκών;

Ναι, υπήρξαν σημαντικές διαφορές. Στο Παραδοτέο 1 φάνηκε ότι το PEGASUS παράγει πιο "συμπυκνωμένες" και στοχευμένες παραφράσεις, το T5 εστιάζει στη διόρθωση και καθαρότητα της γλώσσας, ενώ το BART τείνει να είναι πιο πιστό αλλά και να προσθέτει επεξηγήσεις. Στο Παραδοτέο 2, οι διαφορές στα embeddings ήταν επίσης σαφείς: το Word2Vec είναι πιο περιορισμένο, το FastText βελτιώνει τα αποτελέσματα λόγω υπολεξικής πληροφορίας, ενώ το BERT υπερέχει μακράν, αποτυπώνοντας με ακρίβεια το νόημα και την εγγύτητα των κειμένων.

(5) Συζήτηση Ευρημάτων

Τα πειράματα ανέδειξαν ότι η σημασιολογική ανακατασκευή είναι μια πολυδιάστατη διαδικασία που απαιτεί τόσο ποιοτική όσο και ποσοτική αξιολόγηση. Τα σύγχρονα γλωσσικά μοντέλα (BERT, PEGASUS, T5, BART) έχουν τη δυνατότητα να παράγουν φυσικές και νοηματικά ακριβείς παραφράσεις, όμως παραμένουν προκλήσεις, κυρίως στη διατήρηση του περιεχομένου σε μεγάλα κείμενα. Οι υπολογιστικές μετρικές (cosine similarity, embeddings, PCA/t-SNE) επιβεβαίωσαν την υπεροχή των context-based μεθόδων και ενίσχυσαν την αξιοπιστία της ανάλυσης.

ΣΥΜΠΕΡΑΣΜΑ

Η μελέτη ανέδειξε τη σημασία της σημασιολογικής ανακατασκευής ως βασικό βήμα στη σύγχρονη Επεξεργασία Φυσικής Γλώσσας. Τα πειράματα έδειξαν ότι τα state-of-the-art μοντέλα paraphrasing (PEGASUS, T5, BART) μπορούν να παράγουν ποικίλες εκδοχές κειμένων που διατηρούν το νόημα, με διαφορετικά όμως επίπεδα πιστότητας και φυσικότητας. Παράλληλα, η χρήση embeddings και μετρικών ομοιότητας κατέδειξε την υπεροχή των contextual μοντέλων, όπως το BERT, έναντι των παραδοσιακών μεθόδων Word2Vec και FastText.

Ωστόσο, η διαδικασία συνοδεύτηκε από προκλήσεις: οι περιορισμοί μήκους στις εισόδους, η τάση ορισμένων μοντέλων να προσθέτουν ή να παραλείπουν πληροφορίες, καθώς και η δυσκολία αξιολόγησης μόνο με ποιοτικά κριτήρια ανέδειξαν την ανάγκη για πολυδιάστατη προσέγγιση. Η ενσωμάτωση ποσοτικών μεθόδων (cosine similarity, PCA, t-SNE) ενίσχυσε την αξιοπιστία της μελέτης, προσφέροντας μετρήσιμα στοιχεία για την εγγύτητα των παραφράσεων στα πρωτότυπα.

Συνολικά, τα ευρήματα καταδεικνύουν ότι η αυτόματη σημασιολογική ανακατασκευή έχει μεγάλες δυνατότητες, αλλά απαιτεί συνδυασμό εργαλείων και προσεγγίσεων. Η μελλοντική εργασία μπορεί να εστιάσει σε πιο εξειδικευμένα μοντέλα για την ελληνική γλώσσα και σε αυτοματοποιημένες μεθόδους αξιολόγησης, ώστε να βελτιωθεί περαιτέρω η ακρίβεια και η αξιοπιστία των αποτελεσμάτων.

BONUS- MASKED CLAUSE INPUT

Στο πλαίσιο του Bonus μέρους της εργασίας, εξετάστηκε η δυνατότητα χρήσης μεθόδων Masked Language Modeling για την αυτόματη συμπλήρωση ελλιπών προτάσεων σε νομικά κείμενα. Το πρόβλημα που αντιμετωπίστηκε ήταν η συμπλήρωση λέξεων που λείπουν από άρθρα του Αστικού Κώδικα, με στόχο την αξιολόγηση της ικανότητας των μοντέλων να αποδώσουν ακριβές νόημα σε σχέση με το πραγματικό περιεχόμενο του κειμένου.

ΜΕΘΟΔΟΛΟΓΙΑ

Για την υλοποίηση χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο Greek-BERT, το οποίο είναι ειδικά προσαρμοσμένο στην ελληνική γλώσσα. Μέσω της βιβλιοθήκης HuggingFace Transformers αξιοποιήθηκε το pipeline *fill-mask*, στο οποίο δίνονταν ως είσοδος προτάσεις με μάσκα ([MASK]) στη θέση της λέξης που έλειπε. Το μοντέλο παρήγαγε πολλαπλές πιθανές συμπληρώσεις ταξινομημένες με βάση την πιθανότητα εμφάνισης.

ΑΠΟΤΕΛΕΣΜΑΤΑ

Τα αποτελέσματα έδειξαν ότι το Greek-BERT ήταν σε θέση να προτείνει συχνά λογικές και συντακτικά σωστές συμπληρώσεις, όπως «ακινήτου», «σκάφους», «όροι». Ωστόσο, σε αρκετές περιπτώσεις οι προτεινόμενες λέξεις δεν ταίριαζαν με τη νομική ακρίβεια που απαιτεί ένα κείμενο του Αστικού Κώδικα. Αυτό καταδεικνύει ότι, αν και το μοντέλο έχει καλή κατανόηση της γενικής ελληνικής γλώσσας, υστερεί σε πεδία που απαιτούν εξειδικευμένη γνώση, όπως η νομική ορολογία.

ΣΥΖΗΤΗΣΗ

Η μελέτη ανέδειξε τη δυναμική των μοντέλων Masked LM στη γλωσσική κατανόηση και παραγωγή. Ωστόσο, οι ελλείψεις τους γίνονται εμφανείς όταν το κείμενο απαιτεί υψηλή ακρίβεια, όπως συμβαίνει στα νομικά έγγραφα. Μια πιθανή βελτίωση θα ήταν το fine-tuning σε εξειδικευμένο corpus νομικών κειμένων, ώστε το μοντέλο να προσαρμοστεί καλύτερα στη συγκεκριμένη χρήση.

ΣΥΜΠΕΡΑΣΜΑ BONUS

Το Bonus κομμάτι ανέδειξε ότι η μέθοδος Masked Clause Input μπορεί να λειτουργήσει ως εργαλείο υποστήριξης για την επεξεργασία νομικών κειμένων, αλλά η εφαρμογή της σε πραγματικά σενάρια απαιτεί περαιτέρω εξειδίκευση. Παρά τις αδυναμίες, τα αποτελέσματα δείχνουν τη μελλοντική προοπτική τέτοιων μεθόδων σε πεδία που απαιτούν ακρίβεια και σημασιολογική συνέπεια.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΠΑΡΑΔΟΤΕΟ 1 – Paraphrasing (Pegasus, T5, BART, Similarity)

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. In Proceedings of the 37th International Conference on Machine Learning (ICML). arXiv:1912.08777. https://arxiv.org/abs/1912.08777

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Journal of Machine Learning Research, 21(140), 1–67. http://jmlr.org/papers/v21/20-074.html

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. In Proceedings of ACL 2020. arXiv:1910.13461. https://arxiv.org/abs/1910.13461

HuggingFace. (n.d.). *Transformers Documentation*.

https://huggingface.co/docs/transformers

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. In Proceedings of EMNLP-IJCNLP 2019. arXiv:1908.10084. https://arxiv.org/abs/1908.10084

ΠΑΡΑΔΟΤΕΟ 2- Word Embeddings (Word2Vec, FastText, BERT, PCA/t-SNE)

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

HuggingFace. (n.d.). *Sentence-Transformers: all-MiniLM-L6-v2*. https://www.sbert.net/

Řehůřek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora*. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (pp. 45–50). ELRA. https://radimrehurek.com/gensim/

Bonus (Greek-BERT, Masked Clause Input)

Koutsikakis, J., Papalampidi, P., Koutrika, G., & Papaspyrou, N. (2020). *Greek-BERT: The First Pre-trained Language Model for Modern Greek*. arXiv preprint arXiv:2008.12085. https://arxiv.org/abs/2008.12085

HuggingFace. (n.d.). *Fill-Mask Pipeline Documentation*. https://huggingface.co/docs/transformers/tasks/masked_language_modeling