Αναφορά Σημασιολογικής Ανακατασκευής Κειμένων

Εισαγωγή

Σκοπός της παρούσας εργασίας είναι η εφαρμογή σύγχρονων τεχνικών Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing — NLP) για την ανακατασκευή δύο κειμένων και η αξιολόγηση της αποτελεσματικότητάς τους.

Η αξιολόγηση βασίζεται σε ενσωματώσεις προτάσεων (embeddings) από το μοντέλο **Sentence-BERT (SBERT)** και στη μετρική συνημιτόνου (cosine similarity), ώστε να ελεγχθεί η διατήρηση του νοήματος μετά την ανακατασκευή.

Υλοποίηση

Α. Ανακατασκευή των δύο πρώτων προτάσεων

Για την ανακατασκευή των δύο πρώτων προτάσεων κάθε κειμένου επιλέξαμε να χρησιμοποιήσουμε το μοντέλο T5 grammar correction (vennify/t5-base-grammar-correction). Η επιλογή αυτή έγινε γιατί το T5 είναι ένα μοντέλο που έχει εκπαιδευτεί ειδικά ώστε να αναγνωρίζει και να διορθώνει γραμματικά λάθη, χωρίς να αλλάζει το νόημα του κειμένου. Στην πράξη, αυτό σημαίνει ότι οι προτάσεις παραμένουν ίδιες ως προς το περιεχόμενο και τη σημασιολογική πληροφορία που μεταφέρουν, αλλά γίνονται πιο ορθές και ευανάγνωστες.

Η διαδικασία έχει τρία στάδια. Αρχικά, χρησιμοποιείται το AutoTokenizer, το οποίο «σπάει» το κείμενο σε tokens — δηλαδή σε μικρότερα γλωσσικά κομμάτια που μπορούν να επεξεργαστούν από το μοντέλο. Στη συνέχεια, το κείμενο εισάγεται στο AutoModelForSeq2SeqLM (T5), που αναλαμβάνει να προβλέψει την πιο πιθανή γραμματικά σωστή εκδοχή κάθε πρότασης. Τέλος, με το λεγόμενο post-processing αφαιρούνται τα ειδικά τεχνικά tokens που εισάγει το μοντέλο, ώστε το αποτέλεσμα να είναι ένα φυσικό, «καθαρό» κείμενο έτοιμο προς ανάγνωση.

Με αυτόν τον τρόπο πετυχαίνουμε μια στοχευμένη διόρθωση: το περιεχόμενο δεν αλλοιώνεται, αλλά το τελικό αποτέλεσμα είναι πιο σωστό από γραμματική άποψη και πολύ πιο εύκολο στην κατανόηση από τον αναγνώστη.

Β. Ανακατασκευή ολόκληρου του κειμένου

Για την ανακατασκευή του πλήρους κειμένου αξιοποιήθηκαν τρεις διαφορετικές τεχνικές, καθεμία με διαφορετικά πλεονεκτήματα και στόχους:

- **BART (facebook/bart-large-cnn)**
  Το BART είναι ένα μοντέλο που έχει εκπαιδευτεί κυρίως για περίληψη και γενικότερη δημιουργική αναδιατύπωση κειμένων. Στην παρούσα εργασία χρησιμοποιείται ώστε να εξεταστεί πώς μεταβάλλεται το κείμενο όταν αλλάζει ουσιαστικά η μορφή και η δομή του. Το ενδιαφέρον στοιχείο είναι ότι, παρά τις αλλαγές αυτές, μπορεί να διατηρείται η βασική πληροφορία και να παραμένει υψηλή η ομοιότητα (similarity) με το αρχικό κείμενο. Η τεχνική αυτή είναι χρήσιμη όταν θέλουμε μια πιο ελεύθερη απόδοση, που όμως ενδέχεται να αποκλίνει αρκετά από το πρωτότυπο.

- **Parrot T5 (Paraphrasing)**
  Πρόκειται για fine-tuned έκδοση του T5, σχεδιασμένη αποκλειστικά για την παραγωγή παραφράσεων. Σε αντίθεση με το BART, οι παραφράσεις του Parrot T5 είναι πιο κοντά στο αρχικό νόημα και δίνουν μια φυσική ροή στο κείμενο. Είναι κατάλληλη επιλογή όταν ζητούμενο είναι η ανανέωση της διατύπωσης, χωρίς να χαθεί το περιεχόμενο ή να προκύψουν μεγάλες αποκλίσεις στη δομή. Έτσι, μπορούμε να έχουμε πιο σύγχρονη και κατανοητή μορφή του ίδιου νοήματος.

- **T5 Grammar Correction**
  Αυτή η τεχνική εστιάζει αποκλειστικά στη διόρθωση γραμματικών και συντακτικών λαθών, χωρίς να επιχειρεί παραφράσεις ή αλλαγές στη δομή. Με τον τρόπο αυτό διατηρείται σχεδόν πλήρως το νόημα και η αρχική μορφή του κειμένου, αλλά η τελική απόδοση είναι πιο «καθαρή» και επαγγελματική. Το T5 grammar correction είναι η ασφαλέστερη επιλογή όταν το ζητούμενο είναι βελτίωση της ποιότητας του κειμένου, χωρίς νοηματικές αλλοιώσεις.

ΑΝΑΛΥΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΚΩΔΙΚΑ

Η υλοποίηση πραγματοποιήθηκε σε Python με χρήση των βιβλιοθηκών Hugging Face Transformers και Sentence-Transformers. Για την εξασφάλιση αναπαραγωγιμότητας, ορίστηκε σταθερός σπόρος (SEED = 42) για τις βιβλιοθήκες `random`, `numpy` και `torch`. Το περιβάλλον ρυθμίστηκε ώστε να αποφεύγονται προειδοποιήσεις από τους tokenizers μέσω της μεταβλητής `TOKENIZERS_PARALLELISM`.

Όλα τα μοντέλα εκτελέστηκαν σε CPU, με δυνατότητα fallback σε GPU εφόσον υπάρχει διαθεσιμότητα.

Η βιβλιοθήκη `spaCy` χρησιμοποιήθηκε για τον διαχωρισμό των κειμένων σε προτάσεις.

Ο κώδικας περιλαμβάνει τις εξής βασικές συναρτήσεις:

- `grammarCorrection(text)`: Εφαρμόζει beam search στο T5 για καθαρή γραμματική διόρθωση. Το κείμενο μετατρέπεται σε tensors. Το μοντέλο επιστρέφει την διορθωμένη εκδοχή χωρίς ειδικά tokens.
- `bartFunction(text)`: Χρησιμοποιεί το pipeline του BART για παραγωγή περίληψης ή παράφραση. Ορίζονται ελάχιστο και μέγιστο μήκος εξόδου για έλεγχο της έκτασης.
- `parrot_paraphrase(text)`: Εφαρμόζει παράφραση με το μοντέλο Parrot T5. Το κείμενο προετοιμάζεται με prefix `"paraphrase:"` και μετατρέπεται σε tensors. Το μοντέλο επιστρέφει φυσική εκδοχή του κειμένου.
- `getSentenceList(text)`: Χρησιμοποιεί spaCy για διάσπαση του κειμένου σε προτάσεις, επιστρέφοντας λίστα με καθαρισμένες προτάσεις.
- `sbertSimilarityCalc(a, b)`: Υπολογίζει cosine similarity μεταξύ δύο κειμένων, μετατρέποντάς τα σε ενσωματώσεις μέσω Sentence-BERT και εφαρμόζοντας τη συνάρτηση `cos_sim`.
- `perp(n_samples)`: Υπολογίζει ασφαλή τιμή perplexity για t-SNE, ώστε να μην προκύψουν σφάλματα με μικρό αριθμό δειγμάτων.

Ανακατασκευή Κειμένων

Υλοποιήθηκαν δύο βασικές συναρτήσεις ανακατασκευής:

- `customReconstruction(text)`: Εφαρμόζει T5 Grammar Correction στις δύο πρώτες προτάσεις του κειμένου, επιτρέποντας στοχευμένη διόρθωση και αξιολόγηση σε επίπεδο πρότασης.
- `reconstructTexts(text)`: Εφαρμόζει και τις τρεις τεχνικές (BART, Parrot, T5) στο πλήρες κείμενο, επιστρέφοντας τις αντίστοιχες ανακατασκευασμένες εκδοχές.

Εκτέλεση Πειράματος

Για κάθε κείμενο στη λίστα `original_texts`, εκτελούνται τα εξής:

- Εκτύπωση του αρχικού κειμένου.
- Ανακατασκευή των πρώτων δύο προτάσεων με T5 και υπολογισμός cosine similarity.
- Ανακατασκευή ολόκληρου του κειμένου με BART, Parrot και T5.
- Υπολογισμός cosine similarity για κάθε εκδοχή.
- Αποθήκευση των embeddings για οπτικοποίηση.

Οπτικοποίηση Αποτελεσμάτων

Τα embeddings συγκεντρώνονται σε πίνακα NumPy και μειώνονται σε δύο διαστάσεις με χρήση PCA. Στη συνέχεια εφαρμόζεται t-SNE για καλύτερη απεικόνιση της σημασιολογικής απόστασης μεταξύ των εκδοχών. Τα αποτελέσματα απεικονίζονται σε δύο διαγράμματα (PCA και t-SNE), με σημάνσεις για κάθε κείμενο και κάθε μέθοδο ανακατασκευής.

Αποτελέσματα

Κείμενο 1

Η ανακατασκευή των πρώτων δύο προτάσεων με T5 Grammar Correction παρουσίασε εξαιρετικά υψηλή σημασιολογική συνέπεια, με cosine similarity 0.998. Αυτό δείχνει ότι η διόρθωση ήταν καθαρά γραμματική, χωρίς αλλοίωση του νοήματος.

Στο πλήρες κείμενο:

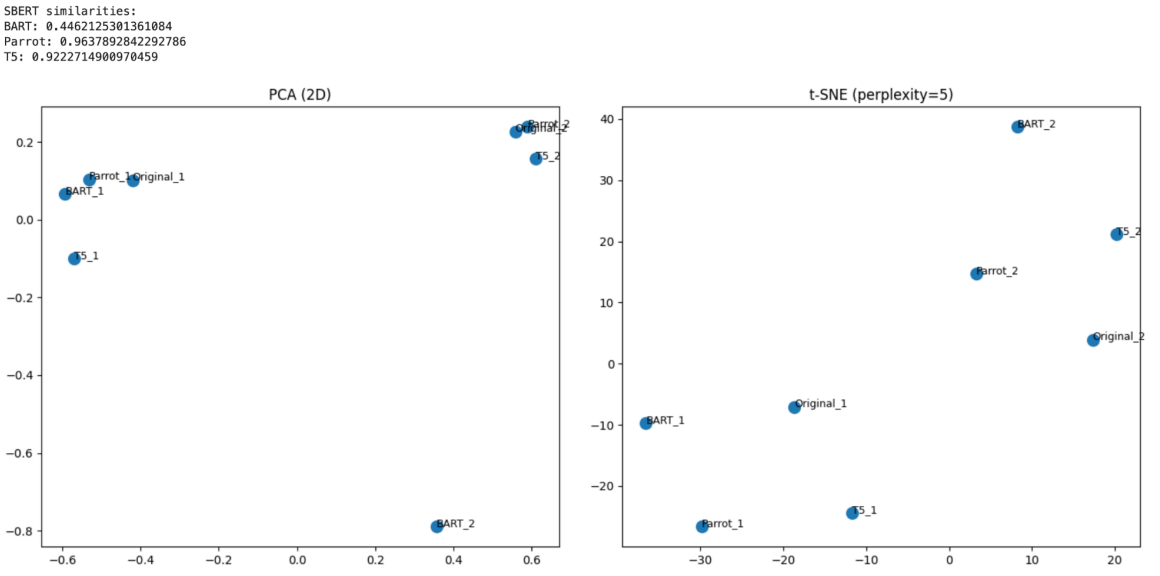
- Η ανακατασκευή με BART είχε similarity 0.894, με πιο δημιουργική διατύπωση αλλά μικρές αποκλίσεις.
- Η ε Parrot T5 παρουσίασε την υψηλότερη ομοιότητα (0.936), διατηρώντας το νόημα με πιο φυσική φρασεολογία.
- Η διόρθωση με T5 Grammar Correction παρέμεινε κοντά στο αρχικό (0.879), με εστίαση στη γλώσσα.

Κείμενο 2

Η ανακατασκευή των πρώτων δύο προτάσεων με T5 Grammar Correction είχε επίσης πολύ υψηλή ομοιότητα (0.976), επιβεβαιώνοντας την ακρίβεια της διόρθωσης.

Στο πλήρες κείμενο:

- Η ανακατασκευή με BART εμφάνισε σημαντική απόκλιση (0.446), καθώς η δημιουργική περίληψη αλλοίωσε το ύφος και τη δομή.
- Η Parrot T5 διατήρησε σχεδόν πλήρως το νόημα (0.964), με πιο φυσική διατύπωση.
- Η διόρθωση με T5 Grammar Correction παρέμεινε υψηλή (0.922), με εστίαση στη βελτίωση της γραμματικής.



Παραδείγματα Ανακατασκευής

Αρχικό (Κείμενο 1): «Thank your message to show our words to the doctor, as his next contract checking, to all of us.»

T5 Grammar Correction: «Thank you for your message to show our words to the doctor, as his next contract review for all of us.»

Parrot T5 Paraphrasing: «Thanks for your message, which conveys our words to the doctor during his next contract review with all of us.»

BART Reconstruction: «Thank your message to show our words to the doctor, as his next contract checking.»

Αρχικό (Κείμενο 2): «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?»

T5 Grammar Correction: «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 not included the full feedback from reviewer or maybe editor?»

Parrot T5 Paraphrasing: «During our final discussion I told him about the new submission — the one we were waiting since last autumn but the updates was confusing as it did not include the full feedback from reviewer or maybe editor?»

BART Reconstruction: «We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link came finally last week, I think,» he says.

Συζήτηση

Η συγκριτική ανάλυση των τριών μοντέλων ανακατασκευής ανέδειξε διαφορετικά πλεονεκτήματα και περιορισμούς ως προς τη διατήρηση του νοήματος και τη γλωσσική βελτίωση.

Το μοντέλο **T5 Grammar Correction** απέδωσε σταθερά υψηλές τιμές σημασιολογικής ομοιότητας, επιβεβαιώνοντας ότι οι διορθώσεις του είναι καθαρά γραμματικές και δεν επηρεάζουν το περιεχόμενο. Η προσέγγισή του είναι συντηρητική, εστιάζοντας αποκλειστικά στη σύνταξη και στη γραμματική, γεγονός που το καθιστά ιδανικό για εφαρμογές όπου απαιτείται ακρίβεια και τυπικότητα — όπως επιστημονικά κείμενα, νομικά έγγραφα ή επίσημη αλληλογραφία.

Το μοντέλο **Parrot T5 Paraphraser** προσέφερε πιο φυσικές και ρέουσες διατυπώσεις, διατηρώντας το νόημα με υψηλή ακρίβεια. Οι παραφράσεις του ήταν πιο ανθρώπινες, με ομαλή σύνταξη και ποικιλία στο λεξιλόγιο, γεγονός που το καθιστά ιδανικό για εφαρμογές όπως η επεξεργασία περιεχομένου, η προσαρμογή ύφους και η βελτίωση αναγνωσιμότητας. Η υψηλή τιμή cosine similarity (0.936 και 0.964 αντίστοιχα) επιβεβαιώνει ότι το μοντέλο διατηρεί το σημασιολογικό περιεχόμενο, ενώ ταυτόχρονα προσφέρει εκφραστική ευχέρεια.

Αντίθετα, το μοντέλο **BART**, αν και ισχυρό ως γεννήτρια περιλήψεων και δημιουργικών αναδιατυπώσεων, παρουσίασε σημαντική απόκλιση στο δεύτερο κείμενο (0.446), γεγονός που υποδηλώνει ότι η γενετική του φύση μπορεί να οδηγήσει σε απώλεια κρίσιμων πληροφοριών ή σε εισαγωγή ερμηνευτικών στοιχείων. Η χρήση του BART ενδείκνυται σε εργασίες όπως η παραγωγή περιεχομένου και η περίληψη κειμένου. Ωστόσο, σε περιπτώσεις όπου η ακρίβεια και η διατήρηση του νοήματος είναι απαραίτητες -όπως σε επιστημονικά ή νομικά κείμενα- η χρήση του θέλει ιδιαίτερη προσοχή.

Η σύγκριση των μοντέλων ανέδειξε τη σημασία της επιλογής εργαλείου ανάλογα με τον στόχο της ανακατασκευής. Αν η προτεραιότητα είναι το κείμενο να είναι καθαρό και με γραμματική ακρίβεια, το T5 Grammar Correction αποτελεί την πιο αξιόπιστη λύση. Αν ζητείται φυσική και εκφραστική ανακατασκευή με διατήρηση νοήματος, το Parrot T5 υπερέχει. Αντίθετα, το BART προσφέρει ευελιξία και δημιουργικότητα, αλλά με αυξημένο ρίσκο σημασιολογικής απόκλισης.

Η οπτικοποίηση των ενσωματώσεων μέσω PCA και t-SNE ενίσχυσε τη σημασιολογική ανάλυση, προσφέροντας οπτική επιβεβαίωση των αποστάσεων μεταξύ των εκδοχών. Οι ανακατασκευές με υψηλή ομοιότητα συγκεντρώθηκαν κοντά στο αρχικό κείμενο, ενώ οι πιο δημιουργικές εκδοχές (BART) εμφανίστηκαν σε μεγαλύτερη απόσταση, επιβεβαιώνοντας την ερμηνευτική τους διαφοροποίηση.

Συμπεράσματα

Η παρούσα μελέτη ανέδειξε την αποτελεσματικότητα των μετασχηματιστικών μοντέλων στην ανακατασκευή κειμένων, καθώς και τις διαφορές τους ως προς τη διατήρηση νοήματος και τη γλωσσική επεξεργασία. Η μέτρηση cosine similarity με Sentence-BERT προσέφερε αξιόπιστη και ποσοτική αξιολόγηση της σημασιολογικής συνέπειας.

Συνοψίζοντας:

- Το **T5 Grammar Correction** είναι κατάλληλο για καθαρή γραμματική διόρθωση χωρίς παραφράσεις.
- Το **Parrot T5** προσφέρει φυσικές και νοηματικά συνεπείς παραφράσεις, ιδανικές για βελτίωση ύφους και αναγνωσιμότητας.
- Το **BART** παρέχει δημιουργική αναδιατύπωση, αλλά απαιτεί προσοχή σε εφαρμογές όπου η πιστότητα είναι κρίσιμη.

Η επιλογή μοντέλου πρέπει να γίνεται με βάση τις απαιτήσεις του εκάστοτε έργου — εαν πρόκειται για τυπική διόρθωση, ανακατασκευή, ή για περίληψη. Η μελλοντική έρευνα μπορεί να εστιάσει σε συνδυαστικά μοντέλα ή σε fine-tuning για εξειδικευμένα πεδία, όπως η ιατρική νομική, κ.α.

Github Repository

Στον παρακάτω σύνδεσμο παρατίθεται το repository της εργασίας:

https://github.com/panagiotiskalogeridis/NaturalLanguageProcessing2025.git

Βιβλιογραφία

- Lewis, M. et al., BART: Denoising Sequence-to-Sequence Pre-training (2020).
- Reimers, N. & Gurevych, I., Sentence-BERT: Sentence Embeddings (2019).
- Raffel, C. et al., Exploring the limits of transfer learning with a unified text-to-text transformer (2020).
- Hugging Face Docs — https://huggingface.co/docs/transformers
- spaCy Docs — https://spacy.io/usage
- Vennify T5 model — https://huggingface.co/vennify/t5-base-grammar-correction
- Parrot T5 paraphraser — https://huggingface.co/prithivida/parrot_paraphraser