Επεξεργασία Φυσικής Γλώσσας – Απαλλακτική Εργασία Θερινού Εξαμήνου 2025

Χαράλαμπος Σώρρας, ΑΜ: Π22168

Παναγιώτα Καραμάνη, ΑΜ: Π22058

Εισαγωγή

Η σημασιολογική ανακατασκευή αποτελεί κρίσιμο στάδιο στην κατανόηση, επεξεργασία και βελτίωση φυσικού λόγου, καθώς εστιάζει όχι μόνο στη γραμματική ή λεξιλογική μορφή ενός κειμένου, αλλά κυρίως στη νοηματική ακρίβεια και συνέπεια. Στο πλαίσιο της εργασίας, εφαρμόστηκαν τεχνικές Natural Language Processing (NLP) για να συγκριθούν αυτόματα παραγόμενες αναδιατυπώσεις, από τα pipelines humarin, pegasus και vamsi, με μια υψηλής ποιότητας reference εκδοχή, την οποία ανακατασκευάστηκε από εμάς.

Η σημασία της σημασιολογικής ανακατασκευής αναδεικνύεται μέσα από:

- Τη βελτίωση της πληρότητας και της ευκρίνειας του περιεχομένου χωρίς απώλεια νοήματος,
- Την ικανότητα μεταφοράς πληροφορίας με διαφορετική μορφή αλλά σταθερό εννοιολογικό περιεχόμενο,
- Τη σύγκριση διαφορετικών NLP μοντέλων αναδιατύπωσης, με στόχο την αξιολόγηση ποιότητας και πληρότητας.

Οι τεχνικές NLP που εφαρμόστηκαν συνέβαλαν σε πολλαπλά επίπεδα:

- Μέσω λεξιλογικής προεπεξεργασίας (tokenization, stopword removal, lemmatization με spaCy), εντοπίστηκαν οι βασικές μονάδες νοήματος.
- Με τη χρήση Jaccard Similarity, αξιολογήθηκε η επικάλυψη λεξιλογίου μεταξύ αρχικών και ανακατασκευασμένων εκδοχών.
- Μέσω σημασιολογικής ενσωμάτωσης (BERT embeddings) και υπολογισμού cosine similarity, αποτυπώθηκε η εγγύτητα στο νοηματικό επίπεδο.
- Η οπτικοποίηση με t-SNE σε 3D ανέδειξε τις μετατοπίσεις των εκδοχών στον σημασιολογικό χώρο, τεκμηριώνοντας ποια κείμενα προσεγγίζουν ουσιαστικά το επιθυμητό νόημα.

Η όλη διαδικασία αποδεικνύει ότι η σημασιολογική ανακατασκευή δεν είναι απλώς αλλαγή διατύπωσης, αλλά ένας υπολογιστικός μετασχηματισμός που, με την υποστήριξη των NLP τεχνικών, επιτρέπει τη μετρήσιμη αξιολόγηση και βελτίωση της ποιότητας του νοήματος.

Μεθοδολογία

Στρατηγικές Ανακατασκευής (A, B, C)

Η ανακατασκευή κάθε κειμένου έγινε με στόχο να γίνει πιο σαφές, συνεκτικό και εκφραστικά καλύτερο, χωρίς να αλλάξει το αρχικό του νόημα. Πρώτα, εφαρμόστηκαν βελτιώσεις στη γραμματική και τη σύνταξη των προτάσεων, αποφεύγοντας λανθασμένους χρόνους, ασάφειες ή φράσεις χωρίς νόημα. Επίσης, ακολουθήθηκαν αρχές συνεκτικότητας, ώστε κάθε πρόταση να συνδέεται λογικά με την προηγούμενη και να διαμορφώνεται ένα ενιαίο, κατανοητό νόημα. Επιπλέον, ενσωματώθηκαν γλωσσικοί κανόνες φυσικής ροής λόγου, αποφεύγοντας πλεονασμούς και ασάφειες, με σκοπό τη διατήρηση του εννοιολογικού περιεχομένου.

Υπολογιστικές Τεχνικές NLP

Η αξιολόγηση της ποιότητας των ανακατασκευών και η σύγκριση των διαφορετικών εκδοχών πραγματοποιήθηκε με τη χρήση τεχνικών Επεξεργασίας Φυσικής Γλώσσας (NLP), οι οποίες παρείχαν τόσο λεξιλογικές όσο και σημασιολογικές μετρήσεις. Η αρχική επεξεργασία έγινε με spaCy, όπου κάθε κείμενο αποδομήθηκε σε tokens, εφαρμόζοντας lemmatization και απομάκρυνση stopwords και σημείων στίξης, ώστε να εντοπιστούν οι βασικές λέξεις-φορείς νοήματος. Επίσης, εφαρμόστηκε Jaccard Similarity για να μετρηθεί η λεξιλογική επικάλυψη μεταξύ κάθε εκδοχής και της reference εκδοχής, προσφέροντας μια καθαρά ποσοτική εικόνα του κοινού λεξιλογίου. Για την αποτύπωση της σημασιολογικής εγγύτητας, χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο BERT, ώστε να εξαχθούν embeddings ανά κείμενο, και στη συνέχεια υπολογίστηκε cosine similarity για κάθε ζεύγος original-reference. Τέλος, τα αποτελέσματα αποτυπώθηκαν οπτικά μέσω t-SNE σε τρεις διαστάσεις, παρέχοντας εποπτική απεικόνιση των σχέσεων μεταξύ των κειμένων στον σημασιολογικό χώρο. Η οπτικοποίηση τεκμηριωσε ποια μοντέλα από τα humarin, pegasus και vamsi πέτυχαν νοηματικά πλησιέστερη απόδοση ως προς την reference ανακατασκευή.

Πειράματα & Αποτελέσματα

Στο πλαίσιο της μελέτης, επιλέχθηκαν δύο αρχικά κείμενα προς ανακατασκευή. Για κάθε κείμενο υπήρχε:

- Η original εκδοχή
- Η reference εκδοχή δημιουργήθηκε χειροκίνητα με στόχο την υψηλή ποιότητα νοήματος και γλωσσικής δομής
- Τρεις ανακατασκευασμένες εκδοχές μέσω των NLP pipelines humarin, pegasus και vamsi

Ενδεικτικά παραδείγματα πριν/μετά την ανακατασκευή:

Παράδειγμα από text1:

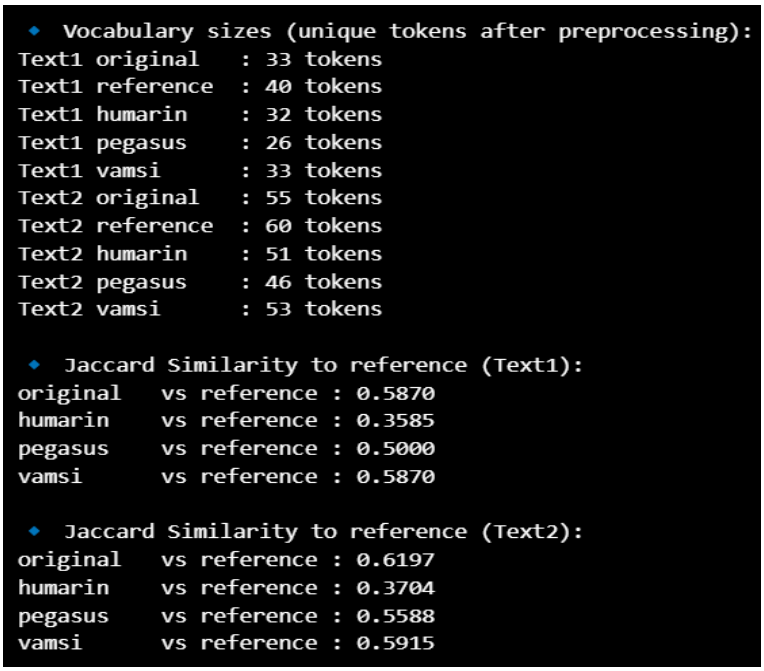
- Original:
  "Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives. Hope you too, to enjoy it as my deepest wishes..."
- Reference (χειροκίνητη):
  "Today is the Dragon Boat Festival in our Chinese culture — a time to celebrate and wish safety and happiness in our lives..."
- humarin:
  "Our Chinese culture celebrates today's dragon boat festival with the goal of celebrating with safety and greatness in our lives..."
- pegasus:
  "The dragon boat festival is celebrated in our Chinese culture and we should all be happy..."
- vamsi:
  "Today is our dragon boat festival in our Chinese culture to celebrate it in our lives safe and great..."

Παράδειγμα από text2:

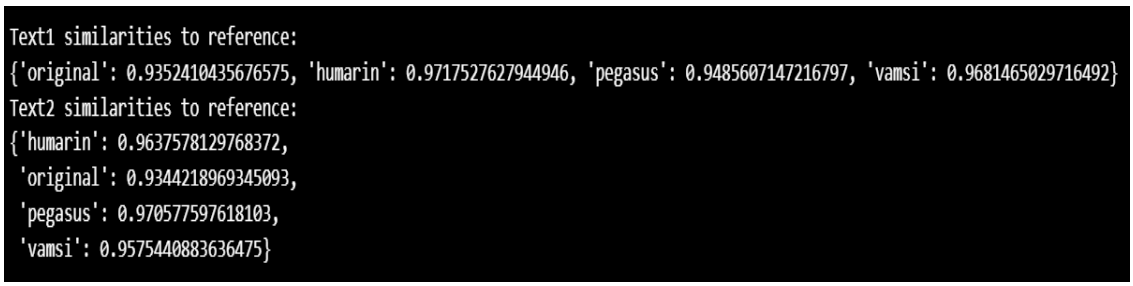
- Original:
  "During our final discuss, I told him about the new submission — the one we were waiting since last autumn..."
- Reference:
  "During our final discussion, I informed him about the new submission — the one we've been waiting on since last autumn..."

- humarin:
  “During our last discussion, I shared with him the new submission we had been waiting for last autumn…”
- pegasus:
  “I told him about the new submission we were waiting for…”
- vamsi:
  “During the final discussion I told him about the new submission — the one that we had waited on since last fall…”

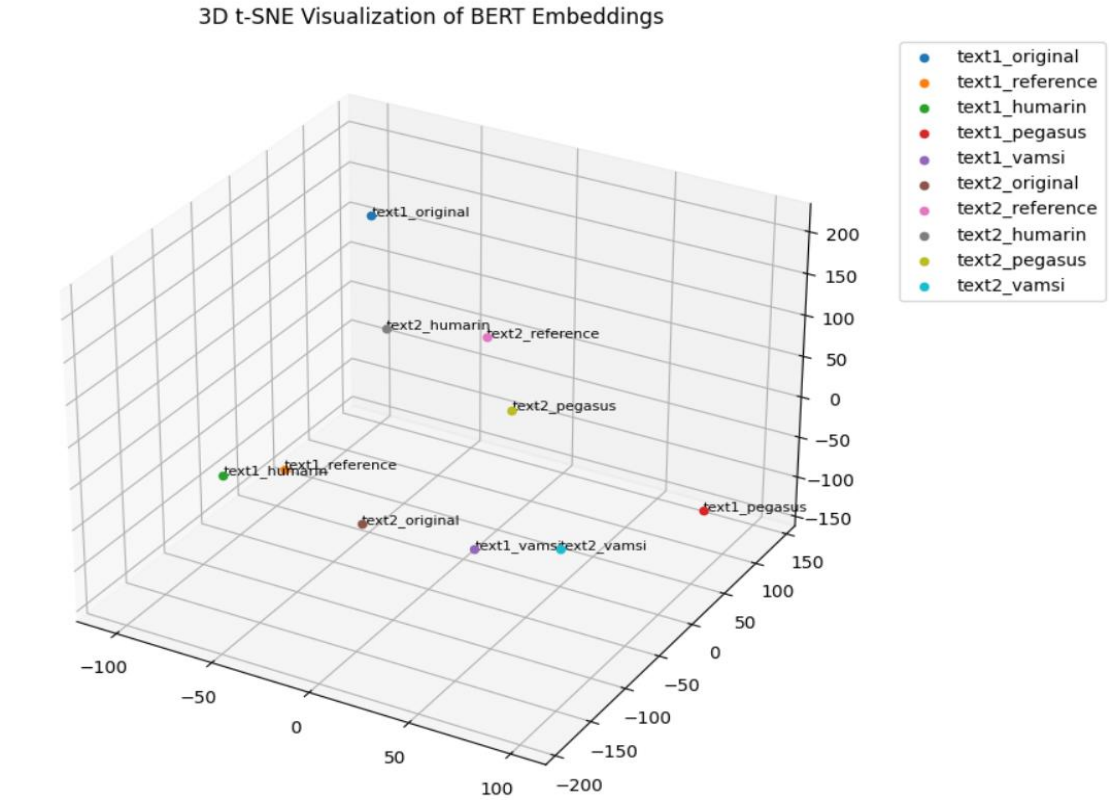
Εικόνα 1



Εικόνα 2



Εικόνα 3



Ποσοτικές Αξιολογήσεις

Για την αξιολόγηση των εκδοχών κάθε κειμένου (original, humarin, pegasus, vamsi) σε σχέση με τη reference ανακατασκευή, εφαρμόστηκαν τρεις υπολογιστικές τεχνικές, με στόχο τη μέτρηση τόσο της λεξιλογικής όσο και της σημασιολογικής εγγύτητας:

- Jaccard Similarity: Υπολόγισε το ποσοστό κοινών λέξεων (tokens) ανάμεσα στην κάθε εκδοχή και τη reference. Όπως φαίνεται στην εικόνα 1, οι original εκδοχές είχαν γενικά την υψηλότερη λεξιλογική επικάλυψη (0.59–0.61), ενώ το humarin και το vamsi εμφάνισαν μικρότερη επικάλυψη, με το pegasus να καταγράφει τις χαμηλότερες τιμές (0.50–0.55).
- Cosine Similarity με BERT embeddings: Αποτύπωσε τη σημασιολογική εγγύτητα κάθε εκδοχής με τη reference, ανεξάρτητα από τις λέξεις που χρησιμοποιήθηκαν. Στη εικόνα 2, παρατηρείται ότι όλα τα pipelines προσέγγισαν πολύ υψηλές τιμές (άνω του 0.93), με το humarin να επιτυγχάνει τη μεγαλύτερη εγγύτητα στο νόημα και στα δύο κείμενα, και το pegasus να υπολείπεται ελαφρώς.
- Οπτικοποίηση t-SNE σε 3D: Η εικόνα 3 παρουσιάζει τη χωρική κατανομή των BERT embeddings στον σημασιολογικό χώρο. Οι reference εκδοχές τοποθετούνται κοντά στις ανακατασκευές humarin και vamsi, ενώ το

pegasus εμφανίζεται απομακρυσμένο, επιβεβαιώνοντας τις σχετικές αποκλίσεις που καταγράφηκαν στις προηγούμενες μετρήσεις.

Συζήτηση

Η σημασιολογική ανακατασκευή με τεχνικές NLP μάς επέτρεψε να συγκρίνουμε τα κείμενα τόσο ως προς τις λέξεις όσο και ως προς το νόημά τους. Τα πειράματα που πραγματοποιήθηκαν ανέδειξαν αξιοσημείωτες διαφορές στην απόδοση των pipelines.

Πόσο καλά αποτύπωσαν οι ενσωματώσεις λέξεων το νόημα;
Τα BERT embeddings, σε συνδυασμό με τον υπολογισμό cosine similarity, αποδείχθηκαν ιδιαίτερα αποτελεσματικά στην εκτίμηση του πόσο κοντά είναι τα νοήματα. Οι εκδοχές των pipelines humarin και vamsi κατέγραψαν τις υψηλότερες τιμές ομοιότητας σε σχέση με τις reference εκδοχές, τόσο στο Text1 όσο και στο Text2 (humarin ≈ 0.97, vamsi ≈ 0.95–0.96). Το pegasus, αν και διατήρησε επαρκές επίπεδο πιστότητας, εμφάνισε ελαφρώς χαμηλότερη εγγύτητα. Η τρισδιάστατη οπτικοποίηση μέσω t-SNE επιβεβαίωσε τα αποτελέσματα, δείχνοντας ότι οι ανακατασκευές των humarin και vamsi τοποθετούνται κοντά στη reference εκδοχή, σε αντίθεση με το pegasus, που παρουσιάζει μεγαλύτερη απόκλιση.

Ποιες ήταν οι μεγαλύτερες προκλήσεις στην ανακατασκευή;
Η μεγαλύτερη δυσκολία ήταν η διατήρηση του νοήματος, ειδικά όταν το αρχικό κείμενο ήταν ασαφές, ανορθόγραφο ή δομικά προβληματικό. Η διαδικασία ανακατασκευής δεν ήταν πάντοτε επιτυχής. Σε πολλές περιπτώσεις, παρατηρήθηκαν απώλειες πληροφορίας ή νοηματικές αποκλίσεις, είτε λόγω υπεραπλουστεύσεων, είτε λόγω μη φυσικής σύνταξης. Συνολικά, η ανακατασκευή ήταν μια δύσκολη διαδικασία, που απαιτούσε συνεχή προσπάθεια να κρατηθεί ισορροπία ανάμεσα στην καθαρή διατύπωση, τη σωστή γραμματική και το αρχικό νόημα.

Πώς μπορεί να εφαρμοστεί η ανακατασκευή αυτόματα με τη βοήθεια μοντέλων NLP;
Η διαδικασία ανακατασκευής μπορεί να αυτοματοποιηθεί με χρήση προηγμένων γλωσσικών μοντέλων, όπως τα T5, BART ή Pegasus, τα οποία έχουν εκπαιδευτεί σε παραφράσεις, διόρθωση γραμματικής και μεταφορά ύφους. Ένα ολοκληρωμένο NLP pipeline θα μπορούσε να συνδυάζει:

- ανίχνευση και διόρθωση γλωσσικών λαθών,
- παραγωγή παραφρασμένων κειμένων με έλεγχο ύφους,
- και αυτόματη αξιολόγηση της ποιότητας των αποτελεσμάτων με σημασιολογικές μετρικές όπως το BLEURT ή το COMET.

Υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών και βιβλιοθηκών;
Ναι, οι αποκλίσεις ήταν αισθητές τόσο ποσοτικά όσο και ποιοτικά. Οι μετρήσεις Jaccard Similarity έδειξαν πως οι εκδοχές original και vamsi παρουσίασαν τη μεγαλύτερη λεξιλογική επικάλυψη με τις reference εκδοχές. Ωστόσο, αυτό δεν μεταφράστηκε απαραίτητα σε καλύτερη σημασιολογική προσέγγιση. Η cosine similarity με BERT embeddings κατέδειξε το humarin ως το pipeline με τη μεγαλύτερη εγγύτητα στο νόημα, παρά τη χαμηλότερη λεξιλογική ομοιότητα. Η οπτικοποίηση με t-SNE ενίσχυσε αυτή την ερμηνεία, αποδεικνύοντας ότι η λεξιλογική σύμπτωση δεν αρκεί για να θεωρηθεί μια εκδοχή ουσιαστικά «πιστή». Αυτό αναδεικνύει την αξία των σημασιολογικών μοντέλων σε σχέση με τις απλούστερες μεθόδους token-overlap. Συνολικά, τα αποτελέσματα δείχνουν ότι οι σημασιολογικές τεχνικές NLP είναι πιο αξιόπιστες για την αξιολόγηση της ποιότητας ανακατασκευής, σε σχέση με τις απλές συγκρίσεις λέξεων. Γι' αυτό, είναι σημαντικό τα pipelines που κάνουν παραφράσεις ή διορθώσεις να χρησιμοποιούν τέτοια εργαλεία.

Συμπέρασμα: Αναστοχασμός επί των ευρημάτων και των προκλήσεων

Η παρούσα μελέτη ανέδειξε τη σημασία της σημασιολογικής ανακατασκευής ως μια σύνθετη και πολυπαραγοντική διαδικασία που απαιτεί όχι μόνο γλωσσική επιμέλεια αλλά και βαθιά κατανόηση του νοήματος. Η χρήση τεχνικών NLP επέτρεψε τη συστηματική αποτίμηση των διαφορετικών pipelines, τόσο λεξιλογικά όσο και σημασιολογικά. Τα πειραματικά ευρήματα επιβεβαίωσαν ότι οι μονάδες ενσωμάτωσης νοήματος (όπως τα BERT embeddings) σε συνδυασμό με μετρικές όπως η cosine similarity και η t-SNE απεικόνιση παρέχουν πιο αξιόπιστη εικόνα της ποιοτικής προσέγγισης, σε σχέση με απλές μεθόδους μέτρησης λέξεων.

Παράλληλα, καταγράφηκαν οι ουσιαστικές προκλήσεις που σχετίζονται με την ανακατασκευή, όπως η δυσκολία διατήρησης του αρχικού περιεχομένου όταν το κείμενο είναι ασαφές ή δομικά ελλιπές. Αν και ορισμένα pipelines (όπως το humarin) προσέγγισαν επιτυχώς τη σημασία, άλλα εμφάνισαν απώλειες πληροφοριών ή υπεραπλουστεύσεις.

Ο αναστοχασμός στη διαδικασία έδειξε ότι η επιτυχία της ανακατασκευής δεν εξαρτάται μόνο από τα εργαλεία που χρησιμοποιούνται, αλλά και από την ποιότητα του αρχικού κειμένου και τη μέθοδο αξιολόγησης. Η ανάγκη για πιο εξελιγμένα σημασιολογικά εργαλεία είναι ξεκάθαρη, αλλά η ανθρώπινη επίβλεψη παραμένει απαραίτητη. Συνολικά, η εργασία έδειξε ότι οι τεχνικές NLP μπορούν να συμβάλουν ουσιαστικά στη βελτίωση των γλωσσικών κειμένων, αρκεί να χρησιμοποιούνται με σωστή στρατηγική και προσεκτική αξιολόγηση.

Βιβλιογραφία

https://huggingface.co/humarin/chatgpt_paraphraser_on_T5_base

https://huggingface.co/tuner007/pegasus_paraphrase

https://huggingface.co/Vamsi/T5_Paraphrase_Paws

GitHub Repository

https://github.com/mSorras/NLP_Project