ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

ΑΝΑΡΓΥΡΟΥ ΛΑΜΠΡΟΥ ΚΑΤΕΡΙΝΑ Π22009
ΔΙΖΟΥ ΧΡΙΣΤΙΝΑ Π22039
ΣΤΟΪΚΟΣ ΙΩΑΝΝΗΣ ΠΑΝΑΓΙΩΤΗΣ Π22164

Εισαγωγή

Η σημασιολογική ανακατασκευή κειμένου αποτελεί βασικό πρόβλημα της Επεξεργασίας Φυσικής Γλώσσας, καθώς συχνά πρέπει να μετατρέψουμε ασαφή και ασύντακτα κείμενα σε σαφείς και σωστά συντακτικά μορφές, διατηρώντας το αρχικό τους νόημα. Η αυτοματοποίηση αυτής της διαδικασίας έχει τεράστια σημασία για πλήθος εφαρμογών όπως η αυτόματη μετάφραση, η παραγωγή περιλήψεων, η διόρθωση λαθών και η ανάλυση δεδομένων σε φυσική γλώσσα. Η υιοθέτηση προηγμένων τεχνικών NLP, όπως οι σημασιολογικές ενσωματώσεις και τα μεγάλα γλωσσικά μοντέλα, επιτρέπει την αξιολόγηση και τη βελτιστοποίηση της ποιότητας των ανακατασκευών με αντικειμενικά κριτήρια.

ΠΑΡΑΔΟΤΕΟ 1

ΠΑΡΑΔΟΤΕΟ 1Α

Στο Παραδοτέο 1Α, αναπτύξαμε έναν αυτόματο μηχανισμό ανακατασκευής δύο επιλεγμένων προτάσεων με στόχο τη βελτίωση της σαφήνειας και της σύνταξης. Ο μηχανισμός βασίστηκε σε τρία διαδοχικά στάδια:

- Αυτόματο γραμματικό έλεγχο με language_tool_python
- Λεξική διόρθωση μέσω textBlob
- Εφαρμογή custom κανόνων, όπως για παράδειγμα αντικατάσταση κοινών λαθών όπως "Thank your message" γίνεται: "Thank you for your message"

Το τελικό αποτέλεσμα είναι μια σημασιολογικά πιστή αλλά γλωσσικά βελτιωμένη εκδοχή των αρχικών προτάσεων.

ΠΑΡΑΔΟΤΕΟ 1Β

Στο πλαίσιο του Παραδοτέου 1Β, ζητήθηκε η ανακατασκευή των δύο αρχικών κειμένων με χρήση τριών διαφορετικών αυτόματων pipelines επεξεργασίας φυσικής γλώσσας, με στόχο τη βελτίωση της γραμματικής, της σαφήνειας, διατηρώντας ταυτόχρονα όμως και το αρχικό τους νόημα.

Για καθένα από τα δύο κείμενα εφαρμόστηκε η παρακάτω διαδικασία:

1. Διαχωρισμός σε προτάσεις με τη βιβλιοθήκη nltk.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

2. Αυτόματη επεξεργασία κάθε πρότασης ξεχωριστά με τρία pipelines και, στη συνέχεια, επανένωση των προτάσεων σε ένα τελικό ανακατασκευασμένο κείμενο.

Συγκεκριμένα για τα pipelines που χρησιμοποιήθηκαν:

- **1: Grammar Correction με T5**
  - Για κάθε πρόταση, εφαρμόστηκε γραμματική διόρθωση χρησιμοποιώντας το μοντέλο "vennify/t5-base-grammar-correction" από το Huggingface Transformers.
  - Το μοντέλο αυτό είναι ειδικά εκπαιδευμένο για αυτόματη ανίχνευση και διόρθωση γραμματικών λαθών στην αγγλική γλώσσα.
- **2: LanguageTool + TextBlob**
  - Κάθε πρόταση περνάει πρώτα από αυτόματο γραμματικό έλεγχο με το language_tool_python, το οποίο εντοπίζει και διορθώνει συντακτικά/γραμματικά λάθη.
  - Στη συνέχεια, η διορθωμένη πρόταση απλοποιείται περαιτέρω με το TextBlob, που προσφέρει βασική ορθογραφική και λεξιλογική διόρθωση.
- **3: BART Paraphrasing**
  - Χρησιμοποιείται το παραφραστικό μοντέλο "eugenesiow/bart-paraphrase".
  - Για κάθε πρόταση, το BART επιχειρεί να δημιουργήσει μια εναλλακτική, παραφρασμένη εκδοχή, διατηρώντας το βασικό νόημα αλλά αλλάζοντας τη διατύπωση.

Εφαρμογή στα Κείμενα :

Κάθε pipeline εφαρμόστηκε ξεχωριστά σε όλα τα κείμενα και τα αποτελέσματα καταγράφηκαν για σύγκριση.

Παρατηρήσεις :

Συνοπτικά, τα pipelines που βασίζονται στη γραμματική διόρθωση (T5, LanguageTool+TextBlob) βελτιώνουν τα εμφανή λάθη, αλλά συχνά διατηρούν ή αφύσικη διατύπωση. Το BART Paraphrasing (Pipeline 3) προσφέρει παραφράσεις αλλά δεν εξαλείφει όλα τα ασαφή σημεία, ενώ κάποιες φορές αλλάζει περισσότερο το ύφος παρά τη σαφήνεια. Σε όλα τα pipelines, το αρχικό νόημα διατηρείται σε μεγάλο βαθμό, όμως η ποιότητα της τελικής ανακατασκευής διαφέρει. Γενικώς, παρατηρήσαμε ότι κανένα pipeline δεν διόρθωνα επαρκώς τα κείμενα, ακόμα και άλλα που δοκιμάσαμε σε προηγούμενες προσπάθειες. Τα κείμενα πέρα από κάποιες μικρές διορθώσεις ήταν πολύ παρόμοια με τα αρχικά.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

ΠΑΡΑΔΟΤΕΟ 1Γ

Σε αυτό το κομμάτι της εργασίας βάλαμε τα κείμενα στο ChatGPT ώστε να τα ανακατασκευάσει πλήρως και να καταφέρουμε μετά να κάνουμε τις συγκρίσεις αποτελεσματικά. Συγκεκριμένα :

Κείμενο 1

Pipeline 1: Grammar Correction (T5 model)

> Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safety and great in our lives. Hope you too, enjoy it as my deepest wishes. Thank you for your message to show our words to the doctor, as his next contract checking, to all of us. I got this message to see the approved message. In fact, I have received the message from the professor, to show me, this, a couple of days ago. I am very appreciated the full support of the professor, for our Springer proceedings publication.

Pipeline 2: LanguageTool + TextBlob

> 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. Thank your message to show our words to the doctor, as his next contract checking, to all of us. I got this message to see the approved message. In fact, I have received the message from the professor, to show me, this, a couple of days ago. I am very appreciated the full support of the professor, for our Springer proceedings publication.

Pipeline 3: BART Paraphrasing

> Today is our dragon boat festival in Chinese culture, to celebrate it with all safe and great in our lives. Hope you too enjoy it as my deepest wishes. Thank your message to show our words to the doctor, as his next contract checking, to all of us. I got this message to see the approved message. In fact, I have received the message from the professor to show me, this, a couple of days ago. I am very grateful for the full support of the professor for our Springer proceedings publication.

ChatGPT

> Today is the Dragon Boat Festival, an important celebration in Chinese culture. We observe this occasion wishing for safety and happiness in our lives. I hope you also enjoy the festival, as this is my sincere wish for you.
> Thank you for your message and for communicating our thoughts to the doctor in relation to his upcoming contract review. I have received the approved message a couple of days ago from the professor, who has provided valuable support. I am truly grateful for the professor’s continued assistance with our Springer proceedings publication.

Κείμενο 2

Pipeline 1: Grammar Correction (T5 model only)

> During our final discuss, 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 delayed and less communication at recent days, 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 finally came last week, I think. Also, please remind me, if the doctor still plans for the acknowledgments section edit before he's 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.

Pipeline 2: LanguageTool + TextBlob

During our final discuss, I told him about the new submission — the one we were waiting since last autumn, but the updates were confusing as it not included the full feedback from reviewer or maybe editor? Anyway, I believe the team, although a bit of 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 plans for the acknowledgments section edit before he's 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.

Pipeline 3: BART Paraphrasing

During our final discuss, I told him about the new submission — the one we were waiting since last autumn, but the updates were confusing as it not included the full feedback from the reviewer or maybe editor? Although a bit delay and less communication at recent days, I believe the team really tried best for paper and cooperation. We should be grateful, I mean all of us, for acceptance and efforts until the Springer link came finally last week, I think. Also, kindly remind me if the doctor still plans for the acknowledgments section edit before he's 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.

ChatGPT

During our final discussion, I informed him about the new submission—the one we have been waiting for since last autumn. However, the updates were confusing because they did not include the complete feedback from the reviewer or possibly the editor.
In any case, I believe that, despite some recent delays and reduced communication, the team has done their best regarding the paper and our collaboration. We should all be grateful for the acceptance and the efforts made until the Springer link finally arrived last week.
Additionally, please kindly remind me if the doctor still intends to edit the acknowledgments section before resubmitting. I have not yet seen the final version of that part, or perhaps I missed it—if so, I apologize.
Overall, let us ensure everyone is well and celebrate this achievement together—with strong coffee and new goals for the future.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Παρατηρήσεις :

- **ChatGPT:**
  Παράγει σαφές, φυσικό, ακαδημαϊκό αγγλικό κείμενο. Οι προτάσεις έχουν συνοχή, το ύφος είναι κατάλληλο για επιστημονική επικοινωνία και το νόημα διατηρείται απόλυτα, με βελτιωμένη δομή και ομαλότητα.
- **Pipeline 1 (Grammar Correction):**
  Διορθώνει βασικά γραμματικά λάθη και βελτιώνει ελαφρώς τη συνοχή, αλλά κρατά πολλές ασαφείς ή “σπασμένες” εκφράσεις. Το κείμενο παραμένει τεχνητό, με φράσεις όπως “Thank you for your message to show our words to the doctor, as his next contract checking, to all of us.” που δεν είναι φυσικές.
- **Pipeline 2 (LanguageTool + TextBlob):**
  Εστιάζει κυρίως σε ορθογραφικά/συντακτικά λάθη, αλλά **σχεδόν δεν βελτιώνει τη σαφήνεια** και αφήνει πολλές προτάσεις ακριβώς όπως ήταν στο αρχικό κείμενο.
- **Pipeline 3 (BART Paraphrasing):**
  Προσπαθεί να αλλάξει τη διατύπωση, αλλά σε μεγάλο βαθμό αναπαράγει το αρχικό νόημα με μικρές φραστικές αλλαγές, συχνά χωρίς ουσιαστική βελτίωση της σαφήνειας ή της δομής.

Συμπεράσματα:

Παρατηρούμε ότι κανένα από τα pipelines δεν πλησιάζει το επίπεδο του ChatGPT. Το Pipeline 1 είναι κάπως πιο κοντά, γιατί διορθώνει τα πιο φανερά γραμματικά λάθη, αλλά εξακολουθεί να αφήνει πολλές φράσεις που είναι συντακτικά λάθος και δεν βγάζουν νόημα. Τα Pipeline 2 και 3 προσφέρουν ελάχιστη διαφορά μεταξύ τους και σίγουρα πολύ χαμηλότερη ποιότητα σε σχέση με το ChatGPT.

Γενικά παρατηρούμε ότι τα αυτόματα pipelines είναι χρήσιμα για κάποιες πολυ βασικές διορθώσεις, αλλά δεν μπορούν να αντικαταστήσουν σε καμία περίπτωση το ChatGPT. Τελικά, για πραγματικά σημασιολογικά ακριβή και καλά δομημένη ανακατασκευή, η χρήση προηγμένων LLMs όπως το ChatGPT είναι απαραίτητη.

ΠΑΡΑΔΟΤΕΟ 2

Παραδοτέο 2Α:

Για την ποσοτική και οπτική ανάλυση της σημασιολογικής συνάφειας των προτάσεων που ανακατασκευάστηκαν στο Παραδοτέο 1Α, ακολουθήθηκαν τα εξής βήματα:

1. Επιλογή Προτάσεων
Επιλέχθηκαν δύο προτάσεις από τα αρχικά κείμενα και ανακατασκευάστηκαν με custom pipeline (συνδυασμός LanguageTool, TextBlob και custom κανόνων).

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Original Sentence 1:
Thank your message to show our words to the doctor, as his next contract checking, to all of us.

Reconstructed Sentence 1:
Thank you for your message to show our words to the doctor, as his next contract review, to all of us.

Original Sentence 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?

Reconstructed Sentence 2:
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 they did not include the full feedback from the reviewer or maybe editor.

2. Μετατροπή σε Ενσωματώσεις (Sentence Embeddings)
Για κάθε πρόταση δημιουργήθηκε το αντίστοιχο διάνυσμα-ενσωμάτωση μέσω του Sentence-BERT, ώστε να είναι δυνατή η ποσοτική σύγκριση στον σημασιολογικό χώρο.

3. Υπολογισμός Cosine Similarity
Υπολογίστηκε ο βαθμός ομοιότητας μεταξύ κάθε original και της ανακατασκευασμένης εκδοχής της.

Sentence 1: 0.974

Sentence 2: 0.984

Οι τιμές αυτές δείχνουν πολύ υψηλή συνάφεια, άρα η σημασιολογική διαφορά είναι ελάχιστη.

4. Οπτικοποίηση μέσω PCA
Τα embeddings των τεσσάρων προτάσεων (2 αρχικές + 2 ανακατασκευασμένες) απεικονίστηκαν σε δύο διαστάσεις μέσω Principal Component Analysis (PCA), ώστε να φανεί οπτικά η σχετική τους απόσταση στον σημασιολογικό χώρο.

Αποτελέσματα:

Cosine Similarity

- Οι τιμές cosine similarity για τις δύο προτάσεις (0.974 και 0.984) είναι εξαιρετικά υψηλές.
- Αυτό αποδεικνύει ότι το custom pipeline πέτυχε τη βελτίωση της διατύπωσης χωρίς να αλλάξει ουσιαστικά το νόημα.

Οπτικοποίηση (PCA plot)

- Στο PCA διάγραμμα, κάθε ζεύγος original–reconstructed προτάσεων εμφανίζεται πολύ κοντά.

- Η μικρή απόσταση μεταξύ τους δηλώνει ότι οι επεμβάσεις περιορίστηκαν κυρίως στη σύνταξη και τη σαφήνεια, χωρίς μετατόπιση του εννοιολογικού τους περιεχομένου.

Συμπεράσματα – Ερμηνεία

Η ανάλυση δείχνει ότι η χρήση custom αυτόματης ανακατασκευής βελτίωσε τις προτάσεις καθιστώντας τις πιο ορθές και κατανοητές, ενώ το αρχικό τους νόημα παρέμεινε αμετάβλητο. Αυτό τεκμηριώνεται τόσο από τις πολύ υψηλές τιμές cosine similarity, όσο και από τη πολύ μικρή απόσταση των προτάσεων στο σημασιολογικό χώρο, όπως φαίνεται στο PCA.

Συνολικά, η προσέγγιση αυτή αποτελεί ενδεικτικό παράδειγμα του πώς τα σύγχρονα εργαλεία NLP μπορούν να υποστηρίξουν τη δομημένη, μετρήσιμη σημασιολογική ανακατασκευή κειμένου.

Παραδοτέο 2Β

Για την υπολογιστική ανάλυση ακολουθήθηκαν τα εξής βήματα:

1. Ενσωμάτωση Κειμένων (Embeddings):
   Κάθε κείμενο (αρχικό και ανακατασκευασμένα με Pipeline1, Pipeline2, Pipeline3, ChatGPT) μετατράπηκε σε εννοιολογικό διάνυσμα μέσω του Sentence-BERT (all-MiniLM-L6-v2), που είναι κατάλληλο για σύγκριση προτάσεων και κειμένων με βάση το νόημά τους.
2. Υπολογισμός Ομοιότητας (Cosine Similarity):
   Υπολογίστηκε η συνημίτονος ομοιότητα ανάμεσα σε κάθε ανακατασκευασμένη εκδοχή και το αρχικό κείμενο. Η τιμή αυτή (μεταξύ 0 και 1) αποτυπώνει το βαθμό σημασιολογικής συνάφειας — όσο πιο κοντά στο 1, τόσο μικρότερη η αλλαγή στο νόημα.
3. Οπτικοποίηση με PCA:
   Για την οπτική ανάλυση, εφαρμόστηκε μείωση διαστάσεων με PCA ώστε τα embeddings να απεικονιστούν σε δύο διαστάσεις. Έτσι απεικονίζεται η σχετική “απόσταση” κάθε εκδοχής στον σημασιολογικό χώρο.

Αποτελέσματα

Cosine Similarity

Κείμενο	Pipeline1	Pipeline2	Pipeline3	ChatGPT
Text 1	0.998	1.000	0.993	0.953
Text 2	0.993	0.995	0.989	0.937

- Τα αυτόματα pipelines (1, 2, 3) εμφάνισαν πολύ υψηλή σημασιολογική ομοιότητα με το αρχικό κείμενο , γεγονός που δείχνει ότι περιορίστηκαν κυρίως σε διορθώσεις σύνταξης και ελαφριές παραφράσεις.
- Το ChatGPT εμφάνισε χαμηλότερες τιμές ομοιότητας. Αυτό δείχνει ότι η ανακατασκευή του άλλαξε περισσότερο το σημασιολογικό περιεχόμενο — όχι επειδή αλλοίωσε το νόημα, αλλά επειδή αναδιάρθρωσε το κείμενο πιο δραστικά, βελτιώνοντας τη σαφήνεια και τη συνοχή.

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

PCA Οπτικοποίηση

Η οπτική απεικόνιση των embeddings (με PCA) έδειξε ότι οι εκδοχές των pipelines τοποθετούνται πολύ κοντά μεταξύ τους και με το πρωτότυπο, ενώ η εκδοχή του ChatGPT εμφανίζεται απομακρυσμένη, υποδηλώνοντας σημαντική σημασιολογική ανακατασκευή.

Συμπεράσματα:

- Τα απλά αυτόματα pipelines (γραμ/συντακτικοί διορθωτές, παραφραστές) βελτίωσαν οριακά τη σαφήνεια χωρίς να αλλάξουν αισθητά το νόημα.
- Το ChatGPT κατάφερε να παραγάγει ένα κείμενο σαφέστερο, πιο δομημένο και ευανάγνωστο, όπως φάνηκε και στη χειροκίνητη αξιολόγηση, αλλάζοντας περισσότερο τον σημασιολογικό χώρο.
- Η ανάλυση αυτή δείχνει ότι τα embeddings και οι cosine similarity scores είναι αποτελεσματικά εργαλεία για την αντικειμενική σύγκριση εκδοχών κειμένων, αποκαλύπτοντας ποια τεχνική διατηρεί ή αλλάζει περισσότερο το νόημα.

Τελικά Συμπεράσματα

Η παρούσα εργασία ανέδειξε τη σημασία της σημασιολογικής ανακατασκευής και της αυτόματης βελτίωσης φυσικού λόγου, εφαρμόζοντας και συγκρίνοντας διαφορετικές τεχνικές επεξεργασίας κειμένου μέσω NLP. Διαπιστώθηκε ότι τα κλασικά αυτόματα pipelines (όπως γραμματικοί διορθωτές, λεξικολογικά εργαλεία και βασικά παραφραστικά μοντέλα) μπορούν να βελτιώσουν βασικά τη δομή, τη γραμματική και τη συνοχή του κειμένου, διατηρώντας σε μεγάλο βαθμό το αρχικό νόημα. Ωστόσο, υστερούν σε σύγκριση με τα σύγχρονα μεγάλα γλωσσικά μοντέλα (όπως το ChatGPT), τα οποία είναι ικανά να παράγουν όχι μόνο ορθά αλλά και πλήρως κατανοητά, φυσικά και ακαδημαϊκού ύφους κείμενα.

Η ανάλυση μέσω word embeddings, cosine similarity και οπτικοποιήσεων (PCA) προσέφερε αντικειμενικό τρόπο αξιολόγησης, αποδεικνύοντας τη μικρή σημασιολογική απόκλιση των pipelines αλλά και τη μεγαλύτερη, ποιοτικά βελτιωτική, αναδιατύπωση που προσφέρει το ChatGPT.

Είχαμε δυσκολία να βρούμε, παρότι προσπαθήσαμε python pipelines τα οποία θα έβγαζαν 100% σωστά αποτελέσματα. Αυτό μπορούμε να το δούμε και από τα ίδια τα αποτελέσματα αλλά και συγκριτικά με τα αποτελέσματα που μας έδωσε το ChatGPT για τα ίδια κείμενα, τα οποία ήταν εμφανώς ανώτερα από τα υπόλοιπα.

GitHub Clone Repository

https://github.com/pasharch/NLP_2025