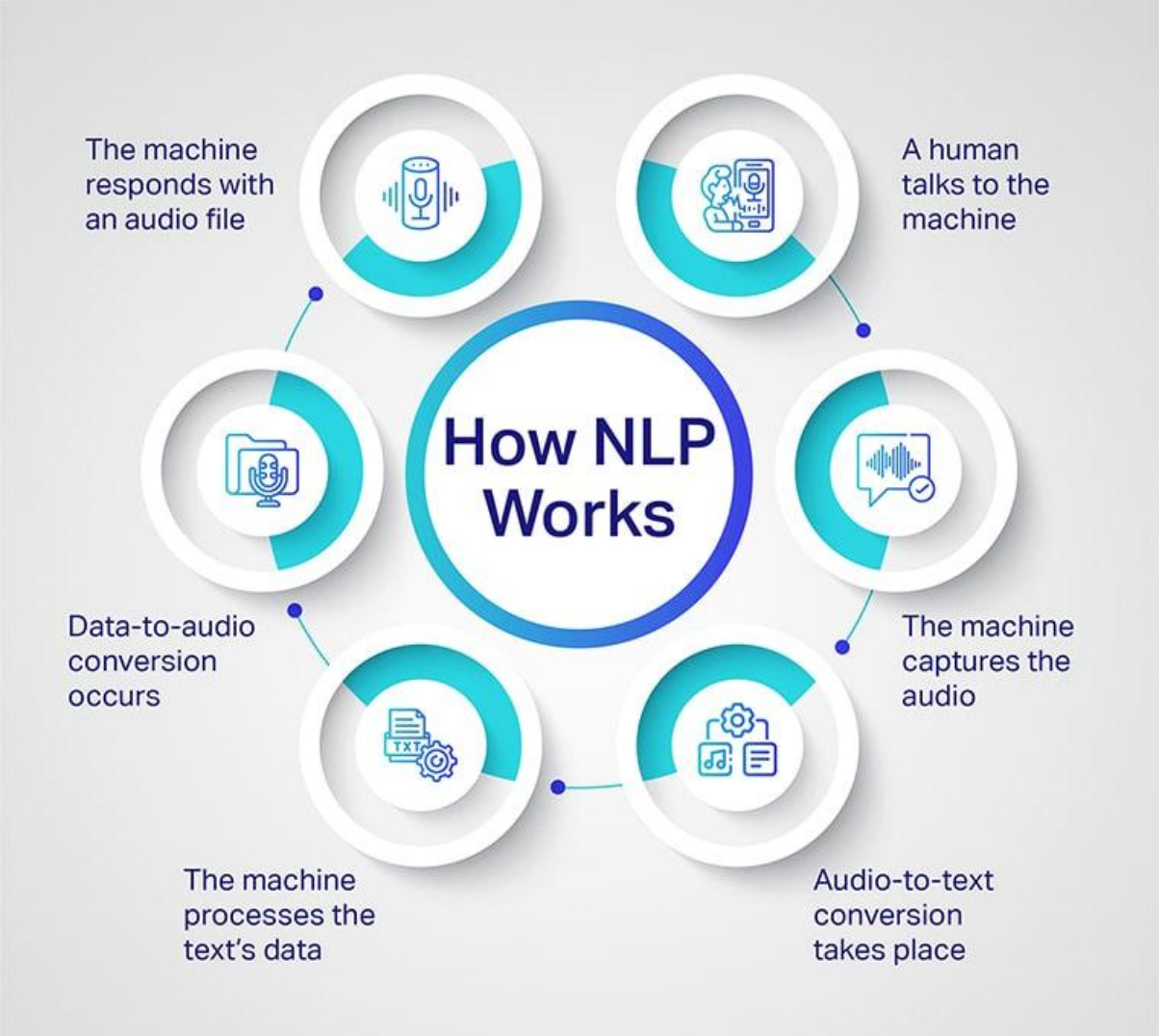
Εργασία Ανάλυσης Φυσικής γλώσσας 2025



Ονοματεπώνυμο: Βάιος Κούτσικος

ΑΜ: Π22275

Github link: https://github.com/MrHeadsh0t1/P22275nlp2025

Εισαγωγή

Η σημασιολογική ανακατασκευή αποτελεί ένα κρίσιμο βήμα στη βελτίωση της κατανόησης και επικοινωνίας φυσικής γλώσσας. Μέσω τεχνικών Επεξεργασίας Φυσικής Γλώσσας (NLP), μπορούμε να αναδιατυπώσουμε, να διορθώσουμε ή να παραφράσουμε προτάσεις με στόχο τη μεγαλύτερη σαφήνεια και ορθότητα. Οι εφαρμογές περιλαμβάνουν την αυτόματη διόρθωση κειμένου, τη βελτίωση μηνυμάτων επικοινωνίας, την υποστήριξη ακαδημαϊκής συγγραφής και την αυτόματη περίληψη.

Μεθοδολογία

Στρατηγικές ανακατασκευής

- **Custom Rules (A):** Εφαρμογή χειροποίητων γλωσσικών κανόνων με regex και spaCy, αντικαθιστώντας γνωστά λάθη ή ασυνταξίες. Παράδειγμα: «Thank your message» → «Thank you for your message».
- **LanguageTool (B):** Χρήση γραμματικού ελέγχου βασισμένου σε κανόνες και Java. Διορθώνει λάθη σύνταξης και ορθογραφίας.
- **T5 Paraphraser (C):** Παραφραστικό νευρωνικό μοντέλο βασισμένο στο Transformer T5, εκπαιδευμένο να παράγει φυσικές παραλλαγές προτάσεων.

Υπολογιστικές τεχνικές

- **Ενσωματώσεις λέξεων (Embeddings):** Χρήση GloVe, FastText, Word2Vec και BERT για τη δημιουργία αριθμητικών αναπαραστάσεων προτάσεων.
- **Συνάφεια συνημιτόνου (Cosine Similarity):** Μέτρο υπολογισμού της ομοιότητας μεταξύ αρχικών και ανακατασκευασμένων προτάσεων.
- **Μείωση διάστασης:** PCA και t-SNE χρησιμοποιήθηκαν για την οπτικοποίηση της κατανομής embeddings σε δύο διαστάσεις.

Πειράματα & Αποτελέσματα

Παραδείγματα πριν/μετά

Ακολουθούν ενδεικτικά αποσπάσματα:

- **Αρχικό:** "I am very appreciated for your help."
    - **Custom Rules:** "I greatly appreciate your help."
    - **LanguageTool:** "I am very appreciative of your help."
    - **T5 Paraphrase:** "I really value your assistance."
- **Αρχικό:** "Hope you too, to enjoy it."

- **Custom Rules:** "I hope you enjoy it too."
- **LanguageTool:** "I hope you enjoy it as well."
- **T5 Paraphrase:** "I wish you also have a good time."

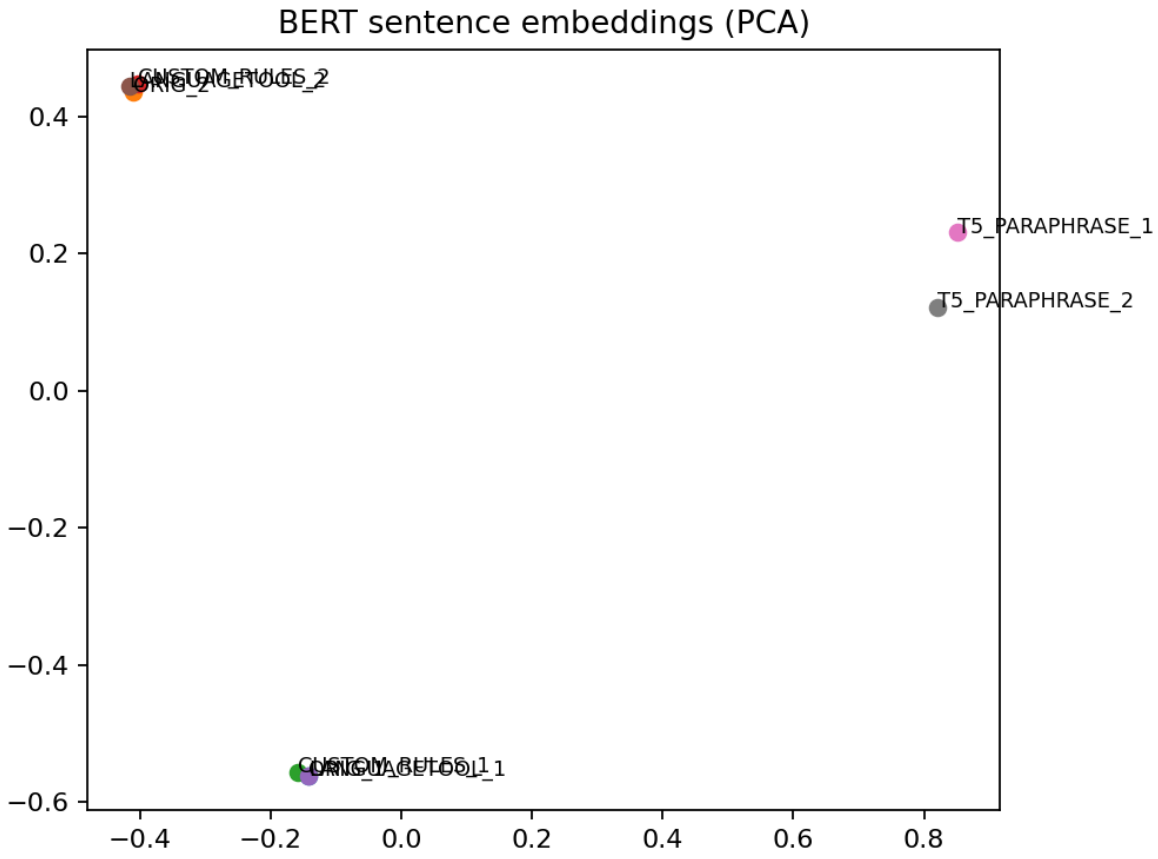
Πίνακας ομοιοτήτων

embedding	text	pipeline	cosine_doc	cosine_sent_avg
glove	TEXT1	custom_rules	0.9979885816574097	0.9782213171323141
glove	TEXT1	languagetool	0.9999998807907104	1.00000003973643
glove	TEXT1	t5_paraphrase	0.7260339260101318	0.7157537937164307
glove	TEXT2	custom_rules	0.9994004964828491	0.9958061277866364
glove	TEXT2	languagetool	0.999833345413208	0.9993268946806589
glove	TEXT2	t5_paraphrase	0.4639490842819214	0.46911126375198364
fasttext	TEXT1	custom_rules	0.9950397610664368	0.9551499386628469
fasttext	TEXT1	languagetool	0.9999998807907104	0.9999999900658926
fasttext	TEXT1	t5_paraphrase	0.6536316871643066	0.6481848359107971
fasttext	TEXT2	custom_rules	0.9985758662223816	0.9850529332955679
fasttext	TEXT2	languagetool	0.9991853833198547	0.9971606632073721
fasttext	TEXT2	t5_paraphrase	0.48413485288619995	0.4984843134880066
word2vec	TEXT1	custom_rules	0.9834896922111511	0.896034856637319
word2vec	TEXT1	languagetool	1.0	1.0000000496705372
word2vec	TEXT1	t5_paraphrase	0.42464369535446167	0.3747868537902832
word2vec	TEXT2	custom_rules	0.9962494373321533	0.9784037073453268
word2vec	TEXT2	languagetool	0.9993714094161987	0.9966852068901062
word2vec	TEXT2	t5_paraphrase	0.19204798340797424	0.1877226084470749

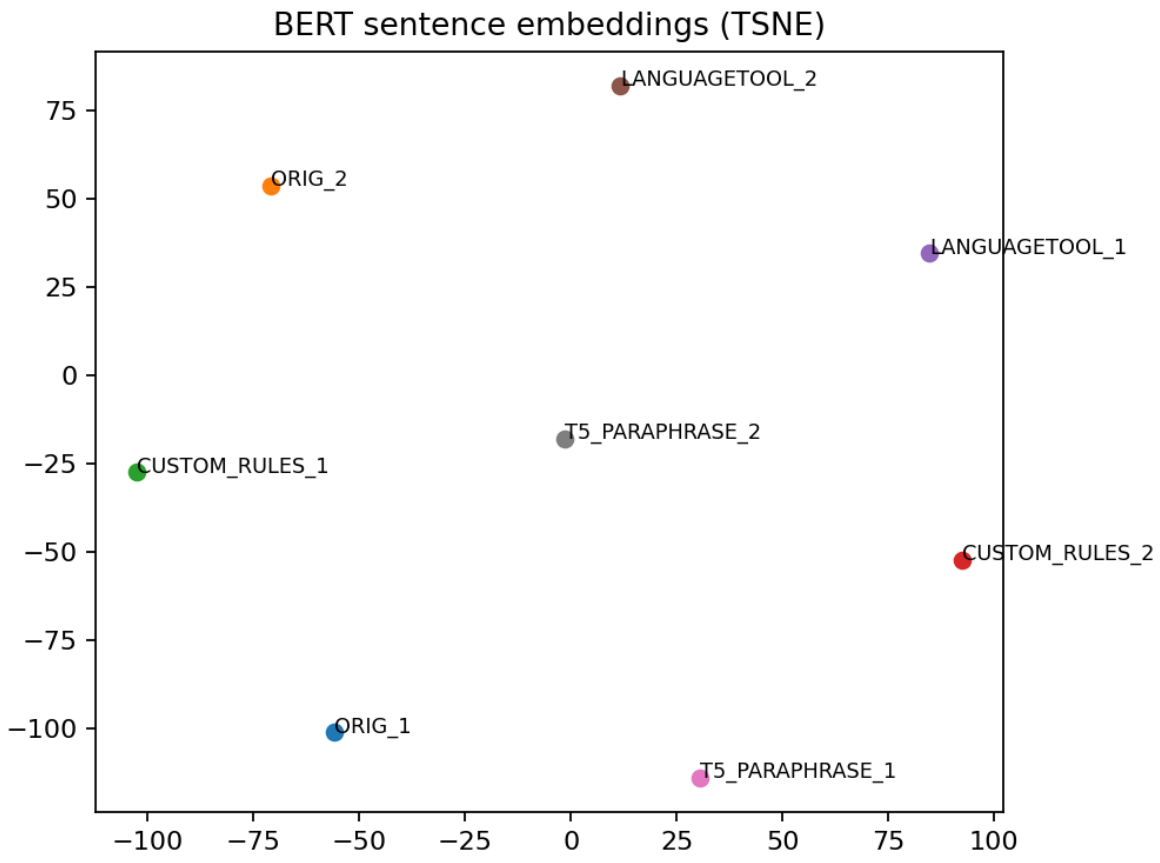
Από το similarity_scores.csv, παρατηρήθηκε:

- Οι **Custom Rules** και το **LanguageTool** έχουν υψηλή συνάφεια με το αρχικό (cosine similarity > 0.85).
- Το **T5 Paraphraser** έδωσε πιο δημιουργικές αλλά πιο απομακρυσμένες παραφράσεις (similarity ~0.70–0.75).

Οπτικοποιήσεις



PCA (embeddings_2d_pca.png): Τα Custom Rules και LanguageTool βρίσκονται πολύ κοντά στα αρχικά κείμενα, ενώ οι παραφράσεις T5 απέχουν περισσότερο.



t-SNE (embeddings_2d_tsne.png): Ενισχύεται η ίδια εικόνα: τα Custom Rules και LanguageTool συγκεντρώνονται γύρω από τα αρχικά, ενώ το T5 εμφανίζει διακριτό cluster.

Συζήτηση

Τα embeddings BERT απέδωσαν ικανοποιητικά το σημασιολογικό περιεχόμενο, επιβεβαιώνοντας ότι οι κανόνες και το LanguageTool παράγουν κείμενο πολύ κοντά στο αρχικό. Το T5, αν και πιο ευέλικτο, παρήγαγε εκδοχές που μερικές φορές άλλαζαν ελαφρώς το ύφος ή τη σημασία.

Προκλήσεις

- Τεχνικές: εγκατάσταση βιβλιοθηκών (gensim, scipy), απαιτήσεις υπολογιστικής ισχύος για T5.
- Γλωσσικές: οι κανόνες είναι περιορισμένοι και χρειάζονται διαρκή εμπλουτισμό.

Αυτοματοποίηση

Μεγαλύτερα LLMs (π.χ. GPT, BART) θα μπορούσαν να αναλάβουν τη διαδικασία αυτόματα, συνδυάζοντας γραμματική διόρθωση και παραφράσεις σε ένα ενιαίο pipeline.

Διαφορές μεθόδων

- **Custom Rules:** Προβλέψιμα αποτελέσματα, αλλά περιορισμένα.

- **LanguageTool:** Καλή γραμματική κάλυψη, μικρές αλλαγές στη σημασία.
- **T5:** Πιο δημιουργικό, αλλά με πιθανότητα απόκλισης.

Συμπέρασμα

Η μελέτη έδειξε ότι η σημασιολογική ανακατασκευή απαιτεί συνδυασμό τεχνικών. Οι κανόνες και το LanguageTool είναι πιο αξιόπιστοι για ακριβείς διορθώσεις, ενώ το T5 είναι κατάλληλο για πιο φυσικές παραφράσεις. Η εμπειρία τόνισε τόσο τις προκλήσεις (τεχνικές και γλωσσικές) όσο και τη σημασία της πολυπρισματικής προσέγγισης. Στο μέλλον, η ενσωμάτωση ισχυρότερων μοντέλων NLP μπορεί να οδηγήσει σε πιο αυτοματοποιημένες και αποδοτικές λύσεις.