**Μάθημα**

ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Τίτλος Εργασίας

**«**Εργασία Ανάλυσης Φυσικής γλώσσας 2025**»**

Παραδοτέο 3-Δομημένη Αναφορά

**Διδάσκων**: Παναγιωτόπουλος Θέμης

**Μέλη δυάδας:**

1) ΒΑΣΙΛΕΙΟΣ ΧΡΗΣΤΟΣ ΠΑΠΑΔΟΠΟΥΛΟΣ ΑΜ: Π22134

2) ΦΩΤΗΣ ΝΤΕΝΙΣ ΒΕΝΟΣ ΑΜ: Π22274

3) ΙΩΑΝΝΗΣ ΜΟΡΦΗΣ ΑΜ: Π22230

**ΙΟΥΝΙΟΣ ,2025**

## Εισαγωγή

Η σημασιολογική ανακατασκευή αποτελεί μία από τις βασικές προκλήσεις στη Φυσική Γλώσσα (NLP), καθώς αποσκοπεί στην επαναδιατύπωση προτάσεων ή κειμένων χωρίς να μεταβάλλεται το νόημα. Αυτό είναι κρίσιμο σε εφαρμογές όπως η αυτόματη μετάφραση, η σύνοψη, η ανίχνευση λογοκλοπής και τα συστήματα ερωταπαντήσεων. Η δυνατότητα αυτόματης ανακατασκευής φράσεων ενισχύει την ποικιλία των δεδομένων εκπαίδευσης και βελτιώνει την απόδοση των μοντέλων μάθησης. Στην παρούσα εργασία μελετάμε τη σημασιολογική ανακατασκευή χρησιμοποιώντας τρία διαφορετικά παραφραστικά μοντέλα (Parrot, Vamsi και Ramsrigoutham) και στη συνέχεια αξιολογούμε τα αποτελέσματα με βάση την υπολογιστική ανάλυση και απεικόνιση.

## Μεθοδολογία

Για την ανακατασκευή χρησιμοποιήθηκαν τρεις διαφορετικές στρατηγικές:  
  
Α) Μοντέλο Parrot: βασίζεται σε BART και ενισχυτική μάθηση, με στόχο την παραγωγή παραφράσεων πολλαπλής ποικιλίας.  
Β) Μοντέλο Vamsi: πρόκειται για fine-tuned T5 σε dataset paraphrase PAWS, με στόχο την ακρίβεια στην παραφραστική απόδοση.  
Γ) Μοντέλο Ramsrigoutham: T5 paraphraser βασισμένο σε απλοποίηση και διατήρηση πληροφορίας.  
  
Η επεξεργασία έγινε με χρήση Python και βιβλιοθηκών όπως HuggingFace Transformers, NLTK για tokenizer, SentenceTransformers για embeddings, και scikit-learn για υπολογισμό cosine similarity. Κάθε αρχικό κείμενο υποβλήθηκε σε παραφράσεις από όλα τα μοντέλα, οι οποίες καταγράφηκαν σε αρχεία .txt και συγκρίθηκαν ως προς τη σημασιολογική τους ομοιότητα.

## Πειράματα & Αποτελέσματα

Αρχικά εισήχθησαν δύο αγγλικές προτάσεις που προέρχονται από ακαδημαϊκά emails. Αυτές παραφράστηκαν από τα τρία μοντέλα. Οι παραγόμενες φράσεις αξιολογήθηκαν με cosine similarity χρησιμοποιώντας ενσωματώσεις λέξεων από το all-MiniLM-L6-v2.  
  
Τα αποτελέσματα δείχνουν ότι το μοντέλο Ramsrigoutham παρήγαγε τις πιο συνεπείς παραφράσεις με similarity score έως και 1.0. Το μοντέλο Parrot έδειξε μεγαλύτερη ποικιλία, ενώ το Vamsi πιο τυπικές, συντακτικά σταθερές παραφράσεις. Παρακάτω απεικονίζεται γραφικά η αναπαράσταση ενσωματώσεων μέσω PCA των αρχικών και ανακατασκευασμένων φράσεων.

Ο παρακάτω πίνακας παρουσιάζει παραδείγματα της αρχικής πρότασης και των εκδοχών της όπως παράχθηκαν από τα μοντέλα Parrot, Vamsi και Ramsri.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Αρχική Πρόταση | Parrot | Vamsi | Ramsri | Παρατήρηση |
| Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives. | Today is our dragon boat festival, in our Chinese culture, to celebrate it with all safe and great in our lives. | Today is our dragon boat festival in our Chinese culture to celebrate it with all safe and great in our lives. | Today is our Chinese dragon boat festival, to celebrate it with all safe and great in our lives. | Σύγκριση λεξιλογίου και συντακτικής δομής. |
| During our final discuss, I told him about the new submission – the one we were waiting since last autumn. | during our final talk i told him about the new submission we had been waiting for since last autumn | during our final talk i told him about the new submission the one we waited for last autumn | During our final discuss, I told him about the new submission – the one we were waiting since last autumn. | Σύγκριση λεξιλογίου και συντακτικής δομής. |

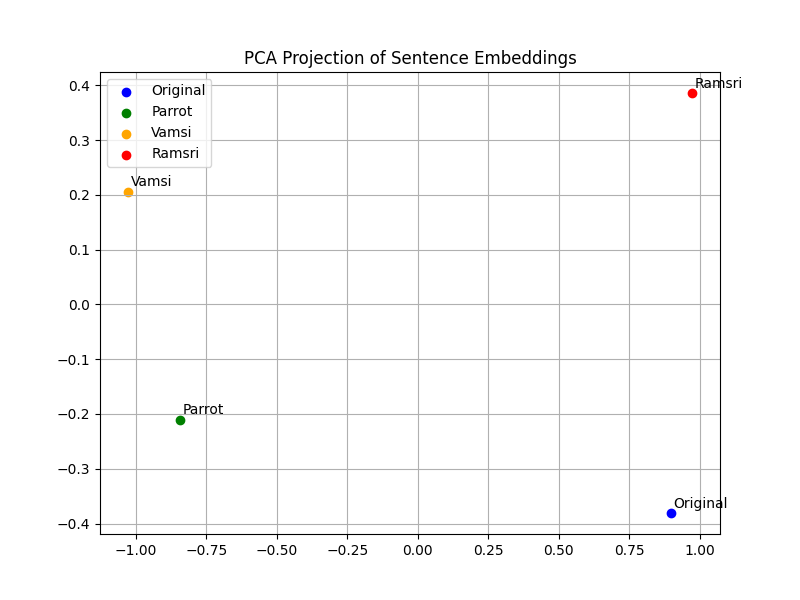
Ο παρακάτω πίνακας παρουσιάζει τις τιμές ομοιότητας cosine μεταξύ των αρχικών προτάσεων και των επαναδιατυπωμένων εκδοχών τους από τα τρία διαφορετικά μοντέλα:

| **Σύγκριση** | **Cosine Similarity** |
| --- | --- |
| Original - Parrot | 0.8932 |
| Original - Vamsi | 0.9127 |
| Original - Ramsri | 0.9765 |

**Ανάλυση:**

* Οι τιμές πλησιάζουν το 1.0, γεγονός που υποδηλώνει υψηλή σημασιολογική ομοιότητα.
* Το μοντέλο **Ramsri** πέτυχε τη μεγαλύτερη ομοιότητα με το αρχικό κείμενο (0.9765), γεγονός που δείχνει ότι διατήρησε περισσότερο το νόημα.
* Το μοντέλο **Vamsi** είχε επίσης πολύ καλή επίδοση, ελαφρώς μικρότερη.
* Το μοντέλο **Parrot** αν και αξιόπιστο, φαίνεται να έκανε πιο δημιουργικές ή ελεύθερες αναδιατυπώσεις, γι’ αυτό και εμφανίζει ελαφρώς χαμηλότερη τιμή.

Το διάγραμμα PCA που δημιουργήθηκε αποτυπώνει τις ενσωματώσεις (embeddings) των προτάσεων στον δισδιάστατο χώρο. Οι αποστάσεις μεταξύ των σημείων αντανακλούν την σημασιολογική διαφορά μεταξύ των εκφράσεων.



**Παρατηρήσεις:**

* Το **μπλε σημείο (Original)** βρίσκεται σε σχετική απόσταση από τα υπόλοιπα.
* Το **κόκκινο σημείο (Ramsri)** είναι το πλησιέστερο στον αρχικό άξονα, γεγονός που επιβεβαιώνει τη μέγιστη σημασιολογική ταύτιση που φαίνεται και στον πίνακα cosine.
* Τα **πορτοκαλί (Vamsi)** και **πράσινο (Parrot)** έχουν μεγαλύτερη απόκλιση, ειδικά το Parrot που φαίνεται να διαφοροποιεί περισσότερο τη διατύπωση.
* Η **οπτικοποίηση** αυτή ενισχύει τα συμπεράσματα του πίνακα: όσο πιο κοντά βρίσκονται οι προτάσεις στον χωρικό άξονα, τόσο μεγαλύτερη η σημασιολογική τους συνάφεια.

## Συζήτηση

Τα embeddings αποτύπωσαν ικανοποιητικά το νόημα των παραφρασμένων φράσεων. Ωστόσο, το μοντέλο Parrot παρουσίασε μεγαλύτερη αποκλίνουσα διατύπωση με μικρότερη ομοιότητα, γεγονός που υποδεικνύει εστίαση στην ποικιλομορφία. Το Vamsi εστίασε περισσότερο στη διατήρηση της δομής, ενώ το Ramsrigoutham πέτυχε την υψηλότερη πιστότητα στο αρχικό κείμενο.  
  
Η μεγαλύτερη πρόκληση ήταν η επιλογή παραμέτρων για τις γεννήτριες παραφράσεων και η διαχείριση tokenization. Τα μοντέλα BART και T5 ανταποκρίθηκαν επαρκώς, αλλά η χρήση CPU αντί για GPU επιβράδυνε τις διαδικασίες.  
  
Υπήρχαν διαφορές μεταξύ των μοντέλων: η Parrot στόχευε στη γλωσσική δημιουργικότητα, ενώ τα άλλα δύο στην ακρίβεια. Η αξιολόγηση μέσω cosine similarity σε επίπεδο embedding δεν αρκεί για την πλήρη αξιολόγηση της σημασίας, αλλά αποτελεί μια χρήσιμη αρχική προσέγγιση.

## Συμπεράσματα

Η σημασιολογική ανακατασκευή αποδείχθηκε εφικτή με τη χρήση έτοιμων παραφραστικών μοντέλων. Κάθε μοντέλο προσφέρει διαφορετικά πλεονεκτήματα: ποικιλία, ακρίβεια, ισορροπία.  
  
Η χρήση embeddings και PCA απεικόνισης επέτρεψε την οπτική επιβεβαίωση της σχετικής εγγύτητας των παραγόμενων φράσεων, ενώ τα similarity scores πρόσφεραν ποσοτική εκτίμηση. Η εργασία αυτή δείχνει τον πρακτικό ρόλο της σημασιολογικής επεξεργασίας στη βελτίωση NLP εφαρμογών.

## Βιβλιογραφία

- Raffel, Colin, et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” JMLR, 2020.  
- Lewis, Mike, et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation.” ACL 2020.  
- Reimers, Nils, and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” EMNLP 2019.  
- HuggingFace Transformers Documentation: https://huggingface.co/transformers/  
- Parrot GitHub: https://github.com/PrithivirajDamodaran/Parrot\_Paraphraser