**Στέφανος Αρναουτίδης Π22016**

**Μαυρέλλης Δημήτριος Π22098**

**ΑΝΑΦΟΡΑ**

**Εισαγωγή:**

Η σημασιολογική ανακατασκευή προτάσεων αφορά τη διαδικασία κατά την οποία ένα γλωσσικό σύστημα μετασχηματίζει εκφράσεις διατηρώντας το βασικό τους νόημα, αλλά βελτιώνοντας τη σύνταξη, τη ροή ή την κατανόηση. Στην σημερινή εποχή, η ανάγκη για αυτόματη επεξεργασία φυσικής γλώσσας (NLP) είναι επιτακτική, ειδικά σε περιβάλλοντα επικοινωνίας όπου η σαφήνεια του νοήματος είναι κρίσιμη.

Το παρόν έργο εξετάζει πώς παραφραστικά μοντέλα και τεχνικές embeddings μπορούν να εφαρμοστούν για την ανακατασκευή ασαφών ή ελλειπτικών αγγλικών προτάσεων. Η μελέτη αφορά την υπολογιστική σύγκριση αρχικών και παραφρασμένων εκδοχών, τόσο ποσοτικά όσο και οπτικά.

**ΠΑΡΑΔΟΤΕΟ 1**

**ΜΕΡΟΣ Α**

Ο ανακατασκευαστής χρησιμοποιεί custom φράσεις και λεξικό συνωνύμων, βασισμένο σε γλωσσικά μοτίβα. Δεν γίνεται χρήση transfomers ή NLP pipelines.

**ΜΕΡΟΣ Β**

Tο παραδοτέο αφορά την ανακατασκευή δύο δομημένων αγγλικών κειμένων με τη χρήση τριών διαφορετικών παραφραστικών μοντέλων. Στόχος είναι η σύγκριση των αποτελεσμάτων και η αξιολόγηση της αποτελεσματικότητας καθενός μοντέλου ως προς τη διατήρηση της σημασιολογίας και τη βελτίωση της εκφραστικής ποιότητας.

Μοντέλα:

* Το Vamsi επιφέρει τις πιο ουσιαστικές, γραμματικά βελτιωμένες αλλαγές.
* Το Ramsri παραμένει σχεδόν ταυτόσημο με το αρχικό κείμενο.
* Το BART διατηρεί τη σημασιολογία αλλά δίνει ελαφρώς πιο ρέουσα διατύπωση.

Η ανακατασκευή έγινε με Python, χρησιμοποιώντας τη βιβλιοθήκη transformers και spaCy. Οι προτάσεις περνούσαν διαδοχικά από τα 3 μοντέλα και εκτυπώνονταν στην κονσόλα.

**ΜΕΡΟΣ Γ**

Χρησιμοποιήθηκε το all-MiniLM-L6-v2 (Sentence Transformers) για να συγκριθούν οι αρχικές προτάσεις με τις ανακατασκευασμένες.

**Κείμενο 1:**

| **Πρόταση** | **Vamsi** | **Ramsri** | **BART** |
| --- | --- | --- | --- |
| 1 | 0.996 | 1.000 | 0.997 |
| 2 | 1.000 | 0.975 | 0.949 |
| 3 | 0.994 | 0.994 | 0.996 |
| 4 | 1.000 | 1.000 | 1.000 |
| 5 | 0.946 | 1.000 | 0.996 |
| 6 | 0.966 | 1.000 | 0.966 |
|  |  |  |  |

**Κείμενο 2:**

| **Πρόταση** | **Vamsi** | **Ramsri** | **BART** |
| --- | --- | --- | --- |
| 1 | 0.985 | 1.000 | 0.999 |
| 2 | 0.956 | 1.000 | 0.893 |
| 3 | 0.999 | 1.000 | 0.998 |
| 4 | 0.980 | 1.000 | 0.993 |
| 5 | 0.994 | 0.998 | 0.998 |
| 6 | 0.975 | 1.000 | 1.000 |
|  |  |  |  |
|  |  |  |  |

**Τα σκορ ομοιότητας ήταν γενικά πολύ υψηλά** (>0.95 στις περισσότερες περιπτώσεις), γεγονός που υποδηλώνει ότι τα μοντέλα **δεν τροποποίησαν σημαντικά το περιεχόμενο.** Ειδικά το μοντέλο ramsrigouthamg είχε τα περισσότερα σκορ **1.000,** γεγονός που δείχνει μηδενική αλλαγή. Τέλος το BART είχε ελαφρώς μεγαλύτερη διαφοροποίηση σε σχέση με τα υπόλοιπα μοντέλα, αλλά δεν ήταν αρκετό.

**ΠΑΡΑΔΟΤΕΟ 2**

Σκοπός του παραδοτέου είναι η μελέτη της σημασιολογικής μεταβολής που προκύπτει μετά την αυτόματη ανακατασκευή φυσικών γλωσσικών προτάσεων. Η ανακατασκευή αφορά δύο κείμενα (Text 1 και text 2) με χρήση των τριών διαφορετικών μοντέλων παραφραστικής επεξεργασίας που αξιοποιήθηκαν στο παραδοτέο 1Β. Η ανάλυση περιλαμβάνει τον υπολογισμό ενσωματώσεων λέξεων (word embeddings), τη μέτρηση σημασιολογικής συνάφειας και τη γραφική απεικόνιση των αποτελεσμάτων στον εννοιολογικό χώρο.

**Μεθοδολογία:**

Χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο all-MiniLM-L6-v2 από τη βιβλιοθήκη sentence-transformers για την απόκτηση σημασιολογικών διανυσμάτων των προτάσεων (sentence embeddings). Τα κείμενα υποβλήθηκαν σε παραφραστική επεξεργασία μέσω τριών pipelines που αξιοποιήθηκαν στο μέρος 1Β:

* Vamsi (Vamsi/T5\_Paraphrase\_Paws)
* Ramsri (ramsrigouthamg/t5\_paraphraser)
* BART (μέσω facebook/bart-large-cnn)

Οι ενσωματώσεις που παρήχθησαν συγκρίθηκαν με τα αρχικά κείμενα μέσω του συνημίτονου της γωνίας (cosine similarity), το οποίο μετρά τη σχετική εγγύτητα στο σημασιολογικό χώρο. Επιπλέον, εφαρμόστηκε PCA για την οπτικοποίηση των εννοιών σε δύο διαστάσεις.

**Πειράματα και Αποτελέσματα:**

Οι τιμές cosine similarity ανά πρόταση και μοντέλο κυμαίνονται ως επί το πλείστον άνω του 0.95, υποδεικνύοντας ότι το αρχικό νόημα διατηρείται σε σημαντικό βαθμό. Η τιμή 1.0 εμφανίζεται συχνά στο Ramsri, γεγονός που οφείλεται στη μη τροποποίηση συγκεκριμένων προτάσεων. Αντιθέτως, το BART παρουσιάζει ελαφρώς μεγαλύτερες σημασιολογικές μεταβολές σε επιλεγμένες περιπτώσεις.

Η PCA οπτικοποίηση δείχνει ότι οι παραφρασμένες εκδοχές κάθε πρότασης τοποθετούνται κοντά στο αρχικό σημείο στον σημασιολογικό χώρο, επιβεβαιώνοντας ποσοτικά τα αποτελέσματα cosine similarity.

**Συζήτηση:**

Η ποσοτική ανάλυση των ανακατασκευασμένων προτάσεων μέσω cosine similarity έδειξε ότι οι σημασιολογικές μεταβολές που επήλθαν από τα παραφραστικά μοντέλα ήταν ελάχιστες. Στις περισσότερες περιπτώσεις, οι τιμές συνάφειας ήταν μεγαλύτερες από 0.95, ενώ το μοντέλο ramsrigouthamg παρήγαγε πλήθος προτάσεων ταυτόσημων με τις αρχικές (similarity = 1.0), υποδηλώνοντας μηδενική παρέμβαση στο περιεχόμενο.

Ακόμη και το μοντέλο **BART**, το οποίο εμφάνισε τη μεγαλύτερη διακύμανση, περιορίστηκε σε επιφανειακές αλλαγές γραμματικού ή συντακτικού χαρακτήρα, χωρίς να επηρεάσει τη βαθύτερη σημασιολογική δομή των προτάσεων. Το **μοντέλο Vamsi**, αν και είχε κάποιες ουσιαστικές βελτιώσεις γραμματικής, δεν επέφερε σημαντικές διαφοροποιήσεις στο εννοιολογικό φορτίο.

Η PCA οπτικοποίηση των ενσωματώσεων (embeddings) επιβεβαίωσε τα παραπάνω: οι ανακατασκευασμένες εκδοχές κάθε πρότασης εντοπίζονται πολύ κοντά στο αρχικό τους σημείο στον σημασιολογικό χώρο, υποδεικνύοντας ελάχιστη μετατόπιση νοήματος.

Συνολικά, τα εργαλεία NLP που αξιοποιήθηκαν στο πλαίσιο της εργασίας **απέτυχαν να προσφέρουν ουσιαστική σημασιολογική ανακατασκευή**, περιοριζόμενα κυρίως σε λεκτικές ή συντακτικές παραλλαγές.

**Συμπέρασμα:**

Η παρούσα μελέτη ανέδειξε τις **περιορισμένες δυνατότητες των παραφραστικών μοντέλων** που αξιοποιήθηκαν, σε ό,τι αφορά την ουσιαστική σημασιολογική μεταβολή φυσικών γλωσσικών προτάσεων. Παρότι τα εργαλεία λειτούργησαν τεχνικά σωστά, το τελικό αποτέλεσμα δεν παρουσίασε αξιοσημείωτες αποκλίσεις από το αρχικό περιεχόμενο, γεγονός που υπονομεύει τη χρήση τους σε σενάρια όπου ζητείται ενεργή ανακατασκευή.

Οι προκλήσεις που εντοπίστηκαν σχετίζονται με:

* την υπερβολική συντηρητικότητα των μοντέλων (ιδίως του Ramsri),
* τον περιορισμένο βαθμό παραμετροποίησης λόγω έτοιμων pipelines,
* και τη γενικότερη αδυναμία τους να εκτελέσουν επαναδιατύπωση πέρα από επιφανειακό επίπεδο

# Βιβλιογραφία/Δικτυογραφία

1. Raffel, C., et al. (2020). **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. Journal of Machine Learning Research (JMLR), 21(140), 1-67.  
   <https://arxiv.org/abs/1910.10683>  
   (Για το μοντέλο T5 που χρησιμοποιείται στα Vamsi και Ramsri)
2. Lewis, M., et al. (2020). **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**. ACL 2020.  
   <https://arxiv.org/abs/1910.13461>  
   (Για το BART model)
3. Reimers, N., & Gurevych, I. (2019). **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. EMNLP 2019.  
   <https://arxiv.org/abs/1908.10084>  
   (Για το μοντέλο all-MiniLM-L6-v2 των Sentence Transformers)
4. Van der Maaten, L., & Hinton, G. (2008). **Visualizing Data using t-SNE**. Journal of Machine Learning Research, 9(Nov), 2579–2605.  
   http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf  
   (Για το t-SNE dimensionality reduction)
5. Pedregosa, F., et al. (2011). **Scikit-learn: Machine Learning in Python**. Journal of Machine Learning Research, 12, 2825–2830.  
   https://jmlr.org/papers/v12/pedregosa11a.html  
   (Για την υλοποίηση PCA και cosine similarity)
6. Hugging Face Transformers Documentation  
   https://huggingface.co/docs/transformers  
   (Για χρήση και τεκμηρίωση των pipelines NLP και μοντέλων T5, BART)
7. Sentence-Transformers Documentation  
   <https://www.sbert.net>  
   (Για την υλοποίηση sentence embeddings με pre-trained μοντέλα)

**.**

**GitHub Repository:** <https://github.com/Stefanos-Aranoutidis/assignmentNLP2025>