ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ

Απαλλακτική Εργασία

Σεπτέμβριος 2025

Γιαννακόπουλος Νικόλαος Ιωάννης | Π22029

Θέματα

1. Εισαγωγή

Η φυσική γλώσσα είναι εγγενώς διφορούμενη, αποτελώντας μια βασική πρόκληση για τα σύγχρονα συστήματα NLP σε εργασίες όπως η μηχανική μετάφραση, η ανάλυση και η σημασιολογική ανάλυση. Η σημασιολογική ανακατασκευή κειμένων αποτελεί ένα από τα πιο σημαντικά προβλήματα στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (NLP), σε έναν κόσμο όπου η επικοινωνία γίνεται όλο και πιο πολύπλοκη και πολυγλωσσική, η ανάγκη για αυτόματη βελτίωση της ποιότητας κειμένων είναι επιτακτική. Στην παρούσα εργασία διερευνούμε την επίλυση της ασάφειας συνδυάζοντας προσεγγίσεις νευρωνικών που βασίζονται σε δεδομένα με ανακατασκευή βάσει κανόνων. Ο αγωγός μας ξεκινά με συντακτικό και σημασιολογικό φιλτράρισμα για την απομάκρυνση απίθανων ερμηνειών.

1.1 Σημασία της Σημασιολογικής Ανακατασκευής

Η σημασιολογική ανακατασκευή στοχεύει στη μετατροπή μη δομημένων ή σημασιολογικά αμφίβολων κειμένων σε σαφείς, ορθές και καλά δομημένες εκδοχές, διατηρώντας παράλληλα το αρχικό νόημα. Αυτή η διαδικασία είναι ιδιαίτερα σημαντική σε:

- **Επιστημονική επικοινωνία**: Βελτίωση της σαφήνειας ακαδημαϊκών κειμένων
- **Διεθνή συνεργασία**: Υποστήριξη μη μητρικών ομιλητών της αγγλικής
- **Αυτόματη επεξεργασία**: Προεπεξεργασία κειμένων για περαιτέρω ανάλυση

Τα κείμενα τα οποία θα εστιάσουμε την παρακάτω έρευνα και εφαρμογή είναι τα ακόλουθα:

- **Κείμενο 1**

  “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”

- **Κείμενο 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? Anyway, I believe the team, although bit 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 plan for the acknowledgments section edit before he 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”

1.2 Εφαρμογές NLP στη Διαδικασία

Οι σύγχρονες τεχνικές NLP προσφέρουν πολλαπλές προσεγγίσεις για την ανακατασκευή κειμένων:

- **Κανονικές προσεγγίσεις (Rule-based)**: Χρήση προκαθορισμένων γραμματικών κανόνων
- **Στατιστικές μέθοδοι**: Χρήση μεγάλων γλωσσικών μοντέλων
- **Ενσωματώσεις λέξεων**: Αναπαράσταση λέξεων σε διανυσματικό χώρο
- **Υβριδικές τεχνικές**: Συνδυασμός πολλαπλών προσεγγίσεων

2. Μεθοδολογία

2.1 Στρατηγικές Ανακατασκευής

2.1.1 Παραδοτέο 1Α: Αυτόματη Ανακατασκευή με Κανόνες

Η πρώτη προσέγγιση βασίζεται σε μια κανονική (rule-based) μέθοδο που υλοποιήθηκε από το μηδέν. Τα βασικά στοιχεία της μεθόδου περιλαμβάνουν:

Tokenization Pipeline:

```
def tokenize(sentence):
    # Διαχωρισμός σε αλφαριθμητικούς χαρακτήρες και σημεία στίξης
    # Μετατροπή σε πεζά γράμματα
    # Διατήρηση σημείων στίξης ως ξεχωριστά tokens
```

POS Tagging System:

- Χρήση προκαθορισμένου λεξικού με ετικέτες κομματιών του κειμένου
- Κατηγοριοποίηση σε: V (Verb - Ρήμα), N (Noun – Ουσιαστικό), PRON (Pronoun – Αντωνυμία), DET (Determiner- Προσδιοριστή), ADJ (Adjective – Επίθετο), ADV (Adverb – Επίρρημα), PREP (Preposition - Πρόθεση), CONJ (Conjunction - Και), PUNCT (Punctuation – Σημείο Στίξης)
- Βάσει συμφραζομένων ειδικός χειρισμός του “to” ως Preposition (PREP)

```
def tag(tokens):
    match = [(token, dictionary.get(token, "UNK")) for token in tokens]
    for i in range(len(match) - 1):
        if match[i][0] == "to":
```

```
            match[i] = ("to", "TO" if match[i+1][1] == "V" else
            "PREP")
    return match
```

Transformation Rules:

1. **Κανόνας 1**: Εισαγωγή "you" μετά το "thank"
2. **Κανόνας 2**: Εισαγωγή "for" πριν από "your message"
3. **Κανόνας 3**: Αντικατάσταση "as" με "during"
4. **Κανόνας 4**: Αντικατάσταση "checking" με "review"

```
rules = [
( [("thank", "V")], [("thank", "V"), ("you", "PRON")] ),
( [("you", "PRON"), ("your", "PRON"), ("message", "N")],
[("you", "PRON"), ("for", "PREP"), ("your", "PRON"),
("message", "N")] ),
( [("as", "PREP")], [("during", "PREP")]
( [("checking", "V")], [("review", "N")] ) ]
```

Rules Appliance:

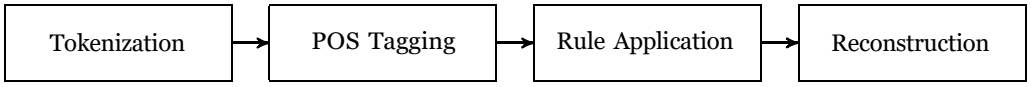
```
def apply_rules(tagged):
    new_tagged = list(tagged)
    for pattern, replacement in rules:
        pattern_len = len(pattern)
        i = 0
        while i <= len(new_tagged) - pattern_len:
            if new_tagged[i:i+pattern_len] == pattern:
                new_tagged = new_tagged[:i] + replacement
                + new_tagged[i+pattern_len:]
```

```
                    i += len(replacement) - 1
                i += 1

        return new_tagged
```

Sentence Reconstruction:

- Αυτόματη κεφαλαιοποίηση πρώτου γράμματος
- Σωστή τοποθέτηση σημείων στίξης
- Διατήρηση φυσικής ροής κειμένου

```
def modify(tagged):
    tokens = [word for word, _ in tagged]

    sentence = " ".join(tokens)
    for punct in [",", ".", "!", "?", ";", ":"]:
        sentence = sentence.replace(" " + punct, punct)
    if sentence:
        return sentence[0].upper() + sentence[1:]
    else:
        return ""
```



Απεικόνιση του Pipeline

2.1.2 Παραδοτέο 1Β: Βιβλιοθήκες Transformers

Η δεύτερη προσέγγιση χρησιμοποιεί τρεις διαφορετικές βιβλιοθήκες Python pipelines:

1. Vamsi/T5_Paraphrase_Paws:

- Βασισμένη στο T5 (Text-to-Text Transfer Transformer)
- Εξειδικευμένη σε paraphrasing
- Εκπαιδευμένη σε PAWS dataset

2. eugenesiow/bart-paraphrase:

- Χρήση της BART αρχιτεκτονικής
- Bidirectional and Auto-Regressive Transformers
- Ισχυρή σε κατανόηση συμφραζομένων

3. prithivida/grammar_error_correcter_v1:

- Εξειδικευμένο σε γραμματικές διορθώσεις
- Βελτιστοποιημένο για common ESL errors
- Διατήρηση σημασιολογικού περιεχομένου

Η δημιουργία των pipeline γίνεται με προκαθορισμένων παραλλαγών για παραγωγή κειμένου:

```
outputs = generator(
    text,
    max_length = 1024,
    num_return_sequences = 2,
    do_sample = True,
    top_k = 50,
    top_p = 0.95
)
```

2.1.3 Παραδοτέο 1C: Συγκριτική Αξιολόγηση

Η αξιολόγηση των μεθόδων πραγματοποιήθηκε με πολλαπλές τεχνικές:

Λεξιλογικές τεχνικές:

- **Token Precision/Recall/F1**: Αναλογία κοινών λέξεων

```
def token_prf(cand: str, ref: str):
    try:
        cand_toks = nltk.word_tokenize(cand.lower())
        ref_toks = nltk.word_tokenize(ref.lower())
    except LookupError:
        cand_toks = cand.lower().split()
        ref_toks = ref.lower().split()
    if not cand_toks or not ref_toks:
        return 0.0, 0.0, 0.0
    matches = sum(1 for t in cand_toks if t in ref_toks)
    P = matches / len(cand_toks)
    R = matches / len(ref_toks)
    if((P + R) > 0):
        F1 = 2 * P * R / (P + R)
    else:
        F1 = 0.0

    return P, R, F1
```

- **Word Error Rate (WER)**: Ελάχιστες επεξεργασίες για μετατροπή

```
def wer_score(cand: str, ref: str):
```

```
        return wer(ref.lower(), cand.lower())
```

Σημασιολογικές τεχνικές:

- **BERTScore**: Χρήση προεκπαιδευμένων BERT embeddings
- **SBERT Cosine Similarity**: Sentence-level semantic similarity
- **TF-IDF Cosine Similarity**: Στατιστική ανάλυση συχνότητας όρων

2.2 Παραδοτέο 2: Υπολογιστικές Τεχνικές

2.2.1 Word Embeddings Implementation

Μοντέλα:

1. **Sentence-BERT (SBERT)**: all-MiniLM-L6-v2 για sentence-level representations
2. **GloVe**: Pre-trained Twitter embeddings για λεξιλογική ανάλυση
3. **Custom Word2Vec**: Εκπαιδευμένο στα δεδομένα μας
4. **Custom FastText**: Χειρισμός out-of-vocabulary λέξεων

Προεπεξεργασία Pipeline:

```
    class TextProcessor:


        def clean_text(self, text, keep_stopwords=False):
            tokens = word_tokenize(text.lower())
            if keep_stopwords:
                cleaned = [self.lemmatizer.lemmatize(t) for t in
    tokens if t.isalnum()]
            else:
                cleaned = [self.lemmatizer.lemmatize(t) for t in
    tokens
```

```
                    if t.isalnum() and t not in
self.stop_words]

        return cleaned
```

Similarity Calculations:

```
    def calculate_similarities(self):
        results = {}
        available_methods = [m for m in ['sbert', 'glove',
'w2v', 'fasttext'] if m in self.models]
        for method in available_methods:
            results[method] = {}
            for text_id in texts["original"].keys():
                orig_text = texts["original"][text_id]
                orig_vec = self.get_text_vector(orig_text,
method)
                if orig_vec is not None:
                    sims = {}
                    for recon_method, recon_texts in
texts["reconstructed"].items():
                        recon_text = recon_texts[text_id]
                        recon_vec =
self.get_text_vector(recon_text, method)
                        if recon_vec is not None:
                            sim =
cosine_similarity(orig_vec.reshape(1, -1),
                                              recon_vec.res
hape(1, -1))[0, 0]
                            sims[recon_method] = sim
                    results[method][text_id] = sims
```

```
        return results
```

2.2.2 Οπτικοποίηση Embeddings

PCA (Principal Component Analysis):

- Μείωση διαστάσεων για 2D visualization
- Εξήγηση variance για κάθε component
- Ανάλυση κύριων κατευθύνσεων μεταβλητότητας

t-SNE (t-distributed Stochastic Neighbor Embedding):

- Μη-γραμμική μείωση διαστάσεων
- Διατήρηση τοπικών δομών
- Βελτιστοποίηση perplexity για μικρά datasets

3. Πειράματα & Αποτελέσματα

3.1 Αποτελέσματα Ανακατασκευής

3.1.1 Παραδοτέο 1A - Rule-based Approach

Αρχική πρόταση 1:

“Thank your message to show our words to the doctor, as his next contract checking, to all of us.”

Ανακατασκευασμένη πρόταση 1:

“Thank you for your message to show our words to the doctor, during his next contract review, to all of us.”

Αρχική πρόταση 2:

"We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link came finally last week, I think.”

Ανακατασκευασμένη πρόταση 2:

“We should be grateful, i mean all of us, for the acceptance and efforts until the springer link came finally last week, i think.”

Παρατηρήσεις:

- Επιτυχής διόρθωση γραμματικών λαθών ("Thank your" → "Thank you")
- Βελτίωση σαφήνειας ("checking" → "review", "as" → "during")
- Διατήρηση αρχικού νοήματος πρότασης

3.1.2 Παραδοτέο 1Β - Transformer Models

Η εφαρμογή των 3 pipeline στα δύο κείμενα είχε τα εξής αποτελέσματα:

Κείμενο 1:

- **T5_Paraphrase_Paws**

  Παραλλαγή 1: Today is our Dragon Boat Festival in our Chinese culture to celebrate it with all safety and greatness in our lives. I received this message from the professor to show me this a couple of days ago . I greatly appreciate the full support of the professor for our Springer proceedings publication.

  Παραλλαγή 2: Today is our Dragon Boat Festival in our Chinese culture to celebrate it with all safety and greatness in our lives. I received this message from the professor to show me this a few days ago . I greatly appreciate the full support of the professor for our Springer proceedings publication.

- **Bart-paraphrase**
  Παραλλαγή 1: Today is our Dragon Boat Festival, in our Chinese culture, to celebrate it with all safety and greatness in our lives. Thank you for your message to show our words to the doctor, as his next contract checking, a couple of days ago.
  Παραλλαγή 2: Today is our Dragon Boat Festival, in our Chinese culture to celebrate it with all safety and greatness in our lives. Thank you for your message to show our words to the doctor, as his next contract checking, a couple of days ago.
- **Grammar_error_correcter_v1**
  Παραλλαγή 1: Today is our Dragon Boat Festival, in our Chinese culture, to celebrate it with all safety and greatness 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
  Παραλλαγή 2: Today is our Dragon Boat Festival, in our Chinese culture, to celebrate it with all safety and greatness 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 check, to all of us. I got this message

Κείμενο 2:

- **T5_Paraphrase_Paws**

  Παραλλαγή 1: 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 the editor? Anyway, I think the team really tried their best for the 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, kindly remind me if the doctor still plans to edit the acknowledgments

Παραλλαγή 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 the editor? Anyway, I think the team really tried their best for the 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, kindly remind me if the doctor still plans to edit the acknowledgements

Bart-paraphrase

Παραλλαγή 1: 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 the editor. I apologize if so. We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link finally came. Also, kindly remind me, please, if the doctor still plans to edit the acknowledgments section before sending it again.
Παραλλαγή 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 the editor. I apologize if so. We should be grateful, I mean all of us, for the acceptance and efforts until the Springer link finally came.

- **Grammar_error_correcter_v1**
  Παραλλαγή 1: We should be grateful, I mean all of us, for 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 the editor. Anyway, I believe the team, although there was a bit of
  Παραλλαγή 2: I believe the team, although there was a bit of delay and less communication in recent days, 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 the editor. Anyway,

T5 Paraphraser Results:

- Κρατάει το νόημα του αρχικού κειμένου σχεδόν αναλλοίωτο, απλώς αλλάζει λίγο τη δομή και κάποιες λέξεις
- Διορθώνει γραμματικά λάθη αλλά δεν βελτιώνει ιδιαίτερα το ύφος ή τη ροή
- Κόβει κάποιες περιττές φράσεις ή διπλές εκφράσεις, αφήνοντας το κείμενο πιο άμεσο

BART Paraphrase Results:

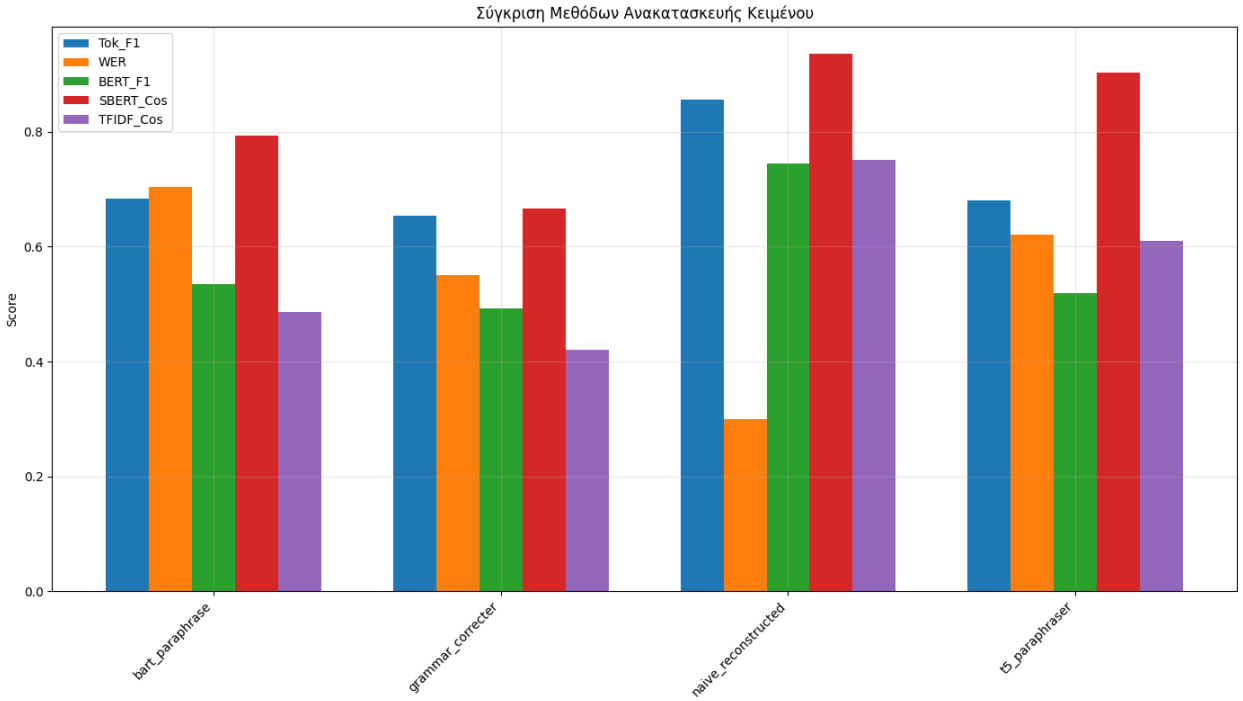
- Αλλάζει τη σειρά προτάσεων και μερικές φορές μετακινεί πληροφορίες για καλύτερη λογική ροή
- Κάποιες φορές αφαιρεί δευτερεύουσες λεπτομέρειες
- Δίνει πιο "native" στυλ, ακόμα και αν χάνονται μικρές αποχρώσεις από το πρωτότυπο.

Grammar Corrector Results:

- Επικεντρώνεται στη γραμματική και τη σύνταξη, χωρίς να αλλάζει πολύ τη διατύπωση
- Κρατάει σχεδόν ίδια τη ροή και τη σειρά των φράσεων, άρα το κείμενο παραμένει πολύ κοντά στο πρωτότυπο
- Κάποιες φορές αφήνει το κείμενο ανολοκλήρωτο ή κομμένο, πιθανώς επειδή δεν κάνει πλήρη αναδιατύπωση αλλά μόνο διορθώσεις

3.1.3 Παραδοτέο 1C – Σύγκριση Αποτελεσμάτων

Pipelines	Token Precision[1]	Token Recall[2]	Token F1-score[3]	Word Error Rate[4]	BERTScore F1[5]	Sentence-BERT Cosine Similarity[6]	TF-IDF Cosine Similarity[7]
bart_paraphrase	0.589	0.821	0.683	0.704	0.535	0.793	0.487
grammar_correcter	0.653	0.661	0.654	0.550	0.492	0.666	0.420
naive_reconstructed	0.833	0.881	0.856	0.300	0.745	0.937	0.751
t5_paraphraser	0.717	0.661	0.681	0.621	0.519	0.903	0.611



Αναλυτικά αποτελέσματα αλγόριθμου

	Token Precision	Token Recall	Token F1-score	Word Error Rate	BERTScore F1	Sentence-BERT Cosine Similarity	TF-IDF Cosine Similarity
1st Naïve Sentence	0.667	0.762	0.711	0.600	0.598	0.873	0.502
2nd Naïve Sentence	1.000	1.000	1.000	0.000	0.893	1.000	1.000
1st t5 Sentence	0.480	0.571	0.522	0.700	0.363	0.838	0.352
2nd t5 Sentence	0.955	0.750	0.840	0.542	0.675	0.967	0.870
1st Bart Sentence	0.480	0.571	0.522	0.700	0.363	0.838	0.352
2nd Bart Sentence	0.698	1.071	0.845	0.708	0.708	0.747	0.622
1st Grammar Sentence	0.625	0.714	0.667	0.600	0.464	0.838	0.431
2nd Grammar Sentence	0.680	0.607	0.642	0.500	0.521	0.495	0.409

Καλύτερο Pipeline ανά μέτρηση

Token Precision	naive_reconstructed (0.856)
BERTScore F1	naive_reconstructed (0.745)
Sentence-BERT Cosine Similarity	naive_reconstructed (0.937)
TF-IDF Cosine Similarity	naive_reconstructed (0.751)
Word Error Rate	naive_reconstructed (0.300)

Ανάλυση μετρήσεων:

- Token Precision:

  Υπολογίζει το ποσοστό των tokens (λέξεων) στην επεξεργασμένη πρόταση που εμφανίζονται και στην αρχική πρόταση.

$$\text{Precision} = \frac{\text{Αριθμός ίδιων tokens}}{\text{Σύνολο tokens στην υποψήφια πρόταση}}$$

- Token Recall:

  Υπολογίζει το ποσοστό των tokens (λέξεων) στην αρχική πρόταση που εμφανίζονται και στην επεξεργασμένη πρόταση.

$$\text{Recall} = \frac{\text{Αριθμός ίδιων tokens}}{\text{Σύνολο tokens στην αναφορά}}$$

- Token F1-Score:

  Υπολογίζει τον μέσο όρο του Token Precision και του Token Recall σε επίπεδο tokens.

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Word Error Rate:

  Υπολογίζει τον αριθμό των εισαγωγών, διαγραφών και αντικαταστάσεων λέξεων που χρειάζονται για να μετατραπεί η αρχική πρόταση στην κανονικοποιημένη με βάση το μήκος της αναφοράς (αρχικής).

$$\text{WER} = \frac{\text{Εισαγωγές} + \text{Διαγραφές} + \text{Αντικαταστάσεις}}{\text{Αριθμός λέξεων στην αναφορά}}$$

- BERTScore F1:

  Υπολογίζει τη σημασιολογική ομοιότητα μεταξύ της επεξεργηασμένης και της αρχικής πρότασης, χρησιμοποιώντας ενσωματώσεις (embeddings) από το BERT.

- Sentence-BERT Cosine Similarity:

Υπολογίζει τη ομοιότητα συνημιτόνου (cosine similarity) μεταξύ των ενσωματώσεων προτάσεων που προκύπτουν από το Sentence-BERT.

- TF-IDF Cosine Similarity:
  Υπολογίζει τη ομοιότητα συνημιτόνου μεταξύ των διανυσμάτων TF-IDF της επεξεργασμένης και της αρχικής πρότασης.

3.2 Αποτελέσματα Υπολογιστικής Ανάλυσης

3.2.1 Παραδοτέο 2 - Embedding ανάλυση

SBERT embeddings

	1st Text	**2nd Text**
Naïve	0.8189	0.7602
t5	0.8135	0.9313
bert	0.8956	0.9831

GloVe embeddings

	1st Text	**2nd Text**
Naïve	0.9849	0.9905
t5	0.9481	0.9781
bert	0.9947	0.9997

Word2Vec embeddings

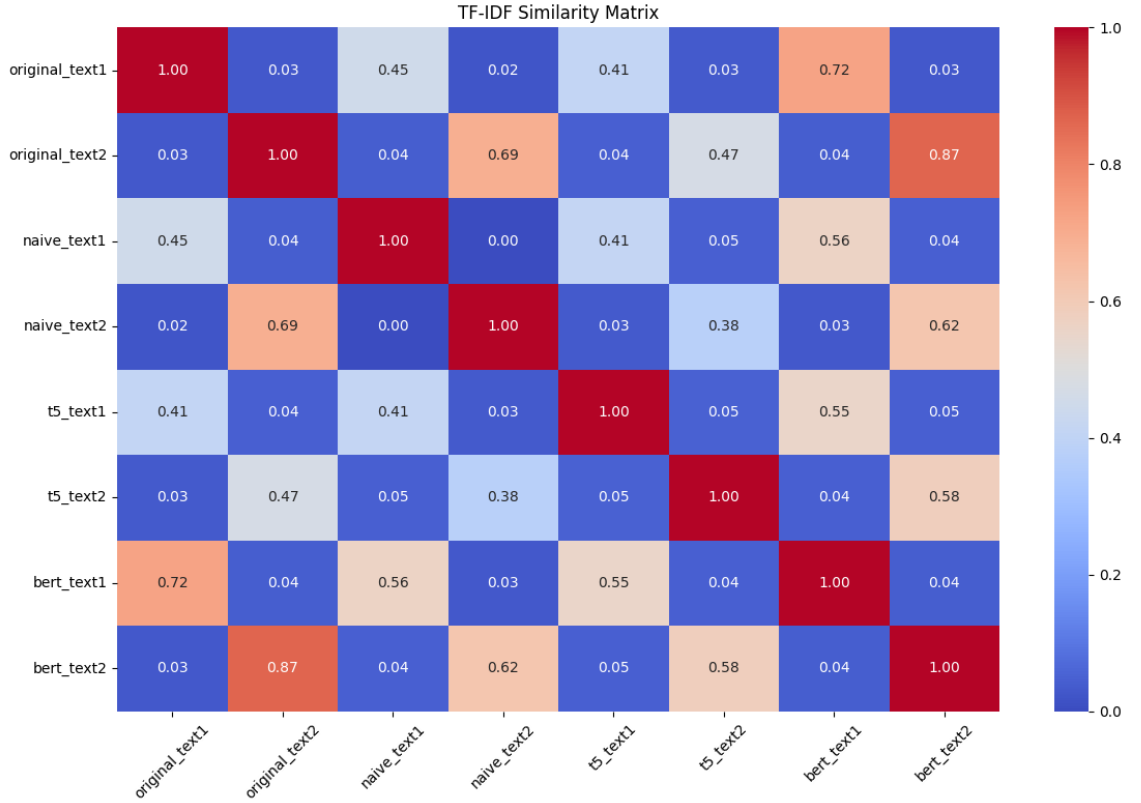
	1st Text	**2nd Text**
Naïve	0.8613	0.9229
t5	0.8806	0.9511
bert	0.9333	0.9968

FastText embeddings

	1st Text	**2nd Text**
Naïve	0.9656	0.9757
t5	0.9758	0.9860
bert	0.9846	0.9995

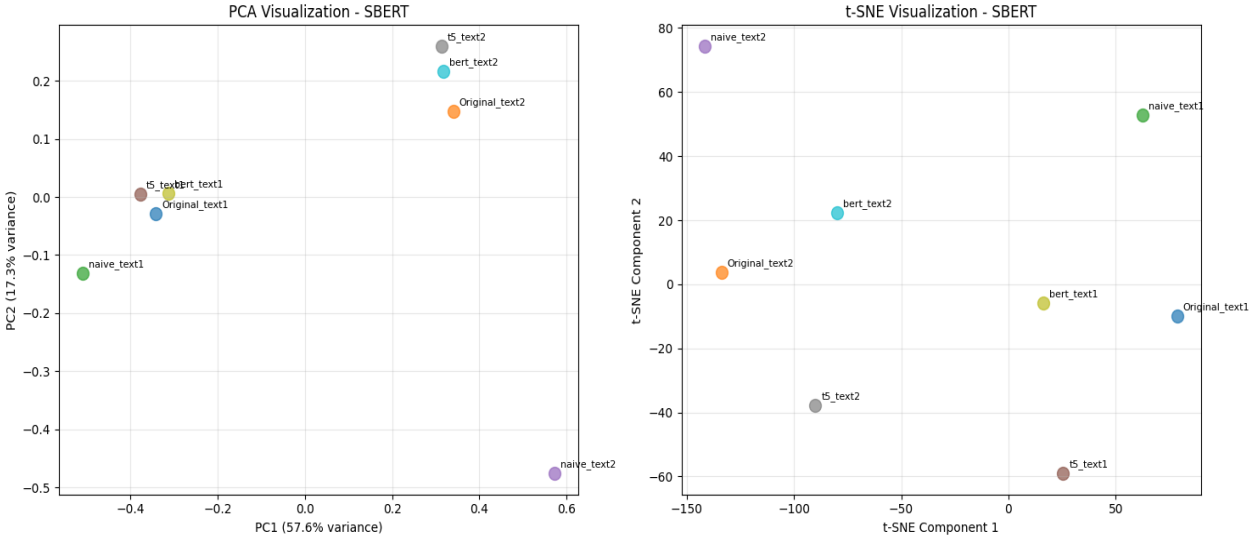
Πίνακας ποιότητας ανακατασκευής κειμένου με τη μέθοδο embedding

TF-IDF πίνακας ομοιότητας

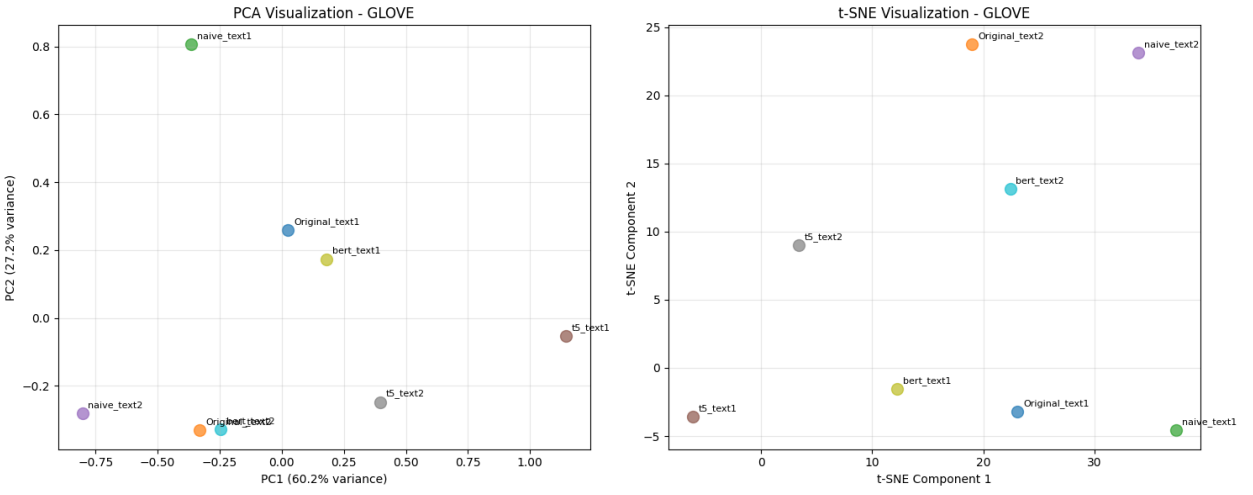


Απεικονήσεις text embeddings

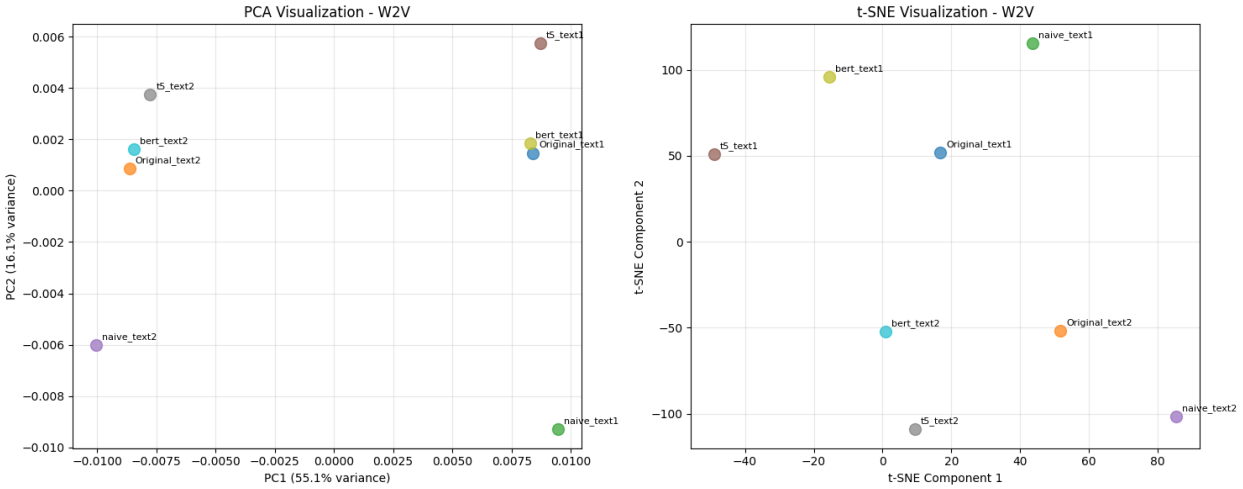
Απεικόνηση SBERT text embeddings



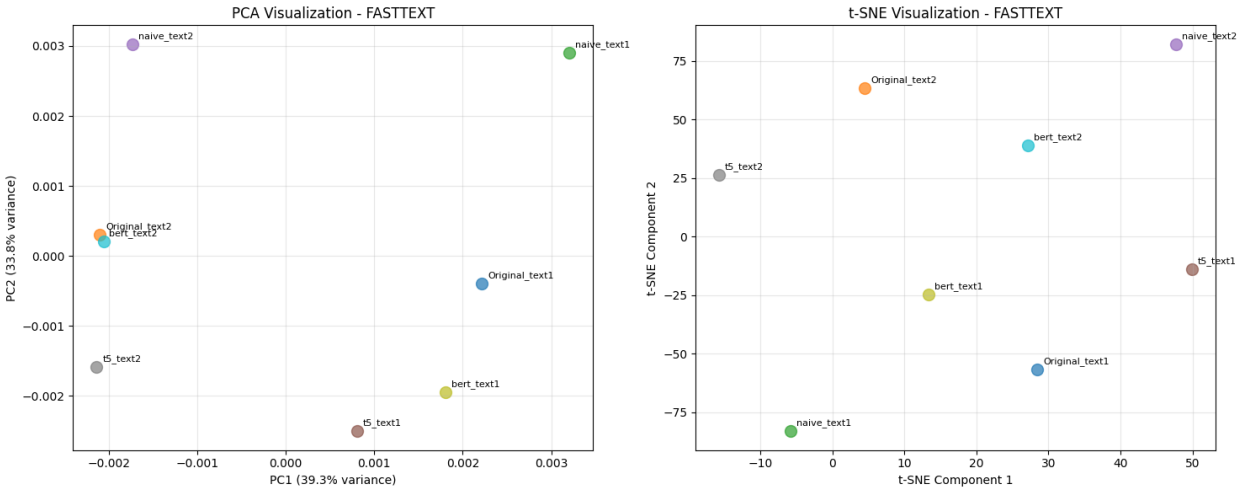
Απεικόνηση GloVe text embeddings



Απεικόνηση Word2Vec text embeddings

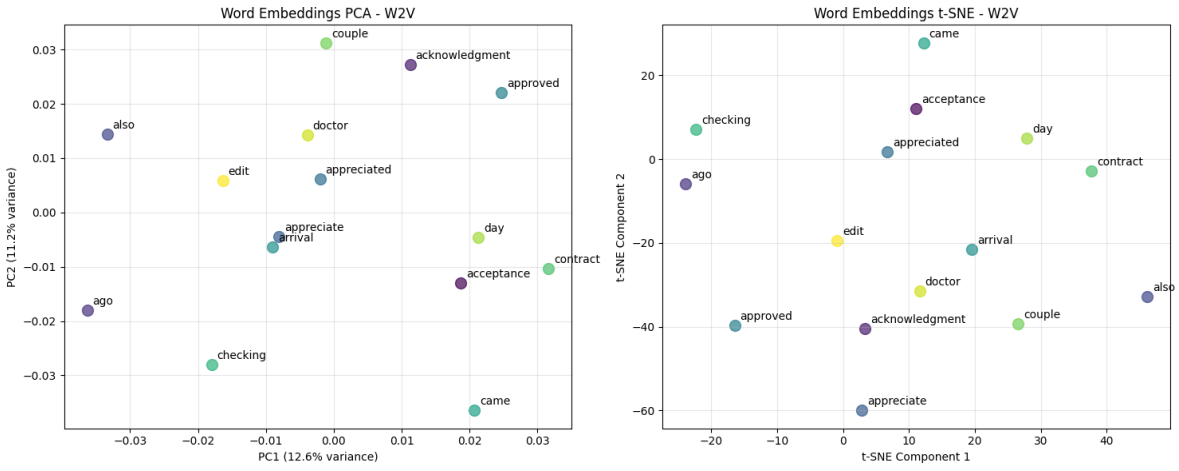


Απεικόνηση FastText text embeddings

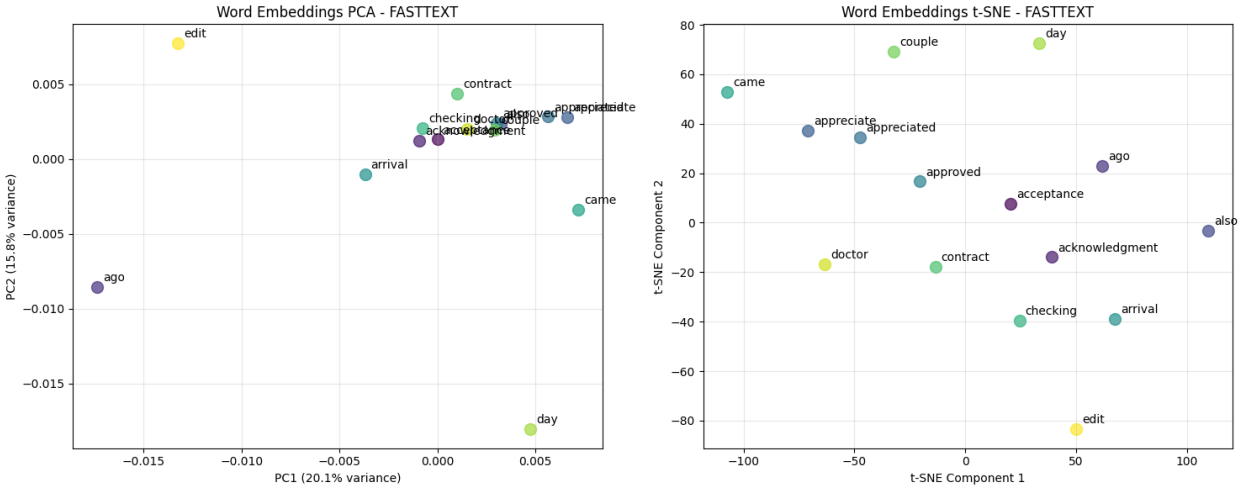


Απεικονήσεις word embeddings

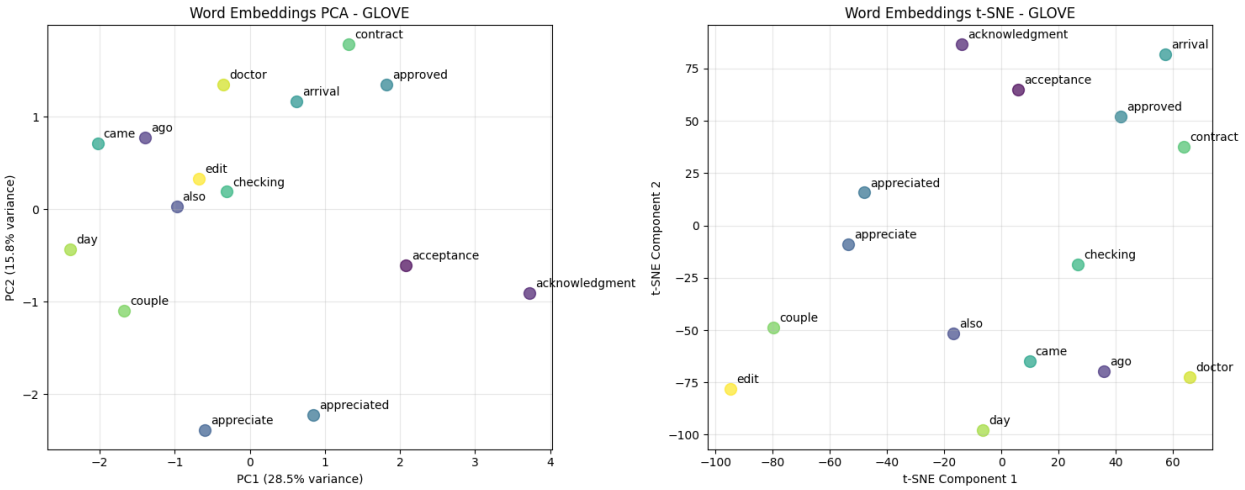
Απεικόνηση Word2Vec word embeddings



Απεικόνηση FastText word embeddings



Απεικόνηση GloVe word embeddings



- **SBERT Embeddings**
  Μεγάλη ευαισθησία στη μέθοδο επεξεργασίας, με έντονη διακύμανση στις τιμές

- **GloVe Embeddings**
  Σταθερά πολύ υψηλές ομοιότητες, ανεξάρτητα από τη μέθοδο

- **Word2Vec**
  Καλή ισορροπία ανάμεσα στην απόδοση και τη μεταβλητότητα, χωρίς ακραίες τιμές

- **FastText**
  Σχεδόν τέλειες τιμές ακόμη και με απλές μεθόδους, ένδειξη ανθεκτικότητας

4. Συζήτηση

4.1 Πόσο καλά αποτύπωσαν τα word embeddings το νόημα;

Τα embeddings απέδωσαν πολύ καλά ως μέτρο σημασιολογικής εγγύτητας, αλλά με σημαντικές διαφορές ανά μοντέλο και επίπεδο ανάλυσης. Συγκεκριμένα, οι μετρήσεις Sentence-BERT cosine similarity και BERTScore έδειξαν ότι κάποιες ανακατασκευές διατήρησαν το νόημα σε υψηλό βαθμό. Στο επίπεδο λέξης/λεξιλογίου, τα παραδοσιακά embeddings (GloVe, FastText, Word2Vec) επέδειξαν υψηλές τιμές ομοιότητας, που υποδηλώνουν ότι οι αλλαγές ήταν συχνά επιφανειακές (λεξικό επίπεδο) και όχι ριζικές μετατοπίσεις νοήματος. Αντίθετα, το SBERT ήταν πιο ευαίσθητο στις διαφορές στη φράση/δομή και έδωσε μεγαλύτερη διακύμανση στις τιμές, πράγμα που το καθιστά πιο κατάλληλο για αξιολόγηση διατήρησης νοήματος σε sentence-level.

4.2 Ποιες ήταν οι μεγαλύτερες προκλήσεις στην ανακατασκευή;

1. Ambiguity resolution: Η ύπαρξη πολλαπλών πιθανών αναγνώσεων (π.χ. «as» που χρειάζεται να γίνει «during» ή άλλο) απαιτούσε συμφραζόμενα μεγαλύτερου εύρους για σωστή επιλογή. Οι rule-based κανόνες διόρθωσαν μερικές σταθερές περιπτώσεις αλλά δεν καλύπτουν πιο σύνθετες, συμφραζομενικές επιλογές

2. Διατήρηση της πρόθεσης/του ύφους: Τα παραδείγματα έδειξαν ότι κάποια paraphrases ήταν “δημιουργικά”, δηλαδή βελτίωναν τη ροή αλλά έκαναν μικρές αλλαγές που ενδέχεται να μεταβάλλουν δευτερεύουσες πληροφορίες. Αντίθετα, οι grammar correctors έδιναν επιφανειακές διορθώσεις αλλά μερικές φορές άφηναν κομμένες φράσεις

3. Αξιολόγηση ποιοτικού χαρακτήρα: Υπήρχε δυσκολία στην ποσοτικοποίηση υποκειμενικών στοιχείων. Οι μετρικές token-level (precision/recall/F1, WER) και οι σημασιολογικές (BERTScore, SBERT cosine) παρέχουν διαφορετικές όψεις

4. Διακύμανση μεταξύ μοντέλων/μεθόδων: Κάποια embeddings (GloVe, FastText) έδιναν πολύ υψηλές ομοιότητες ακόμα και όταν μικρές δομικές αλλαγές υπήρχαν. Αυτό μπορεί να δώσει ψευδή αίσθηση επιτυχίας αν δεν ελεγχθεί σε sentence-level

4.3 Πώς μπορεί να αυτοματοποιηθεί αυτή η διαδικασία χρησιμοποιώντας μοντέλα NLP;

Η αυτοματοποίηση είναι εφικτή με ένα πολυεπίπεδο pipeline που συνδυάζει κανόνες, statistical/transformer μοντέλα και semantic-scoring. Μια προτεινόμενη αρχιτεκτονική βασισμένη και στα πειραματικά συμπεράσματα του project περιλαμβάνει:

- **Preprocessing & Detection:** tokenization, POS tagging, detection of obvious grammar errors (rule-based or lightweight classifier). Αυτό εντόπισε εύκολα περιπτώσεις όπως "Thank your" σε "Thank you"

- **Model Selection & Application:** ανάλογα με τον τύπο του σφάλματος/σκοπού: Grammar corrector για μικροδιορθώσεις, paraphraser για ροή και restructuring όπου απαιτείται, rule-based modules για domain-specific, reproducible διορθώσεις.

- **Semantic Scoring & Selection:** υπολογισμός BERTScore, SBERT cosine, TF-IDF similarity και token-metrics για κάθε υποψήφια ανακατασκευή και επιλογή της εκδοχής που εξισορροπεί διατήρηση νόηματος και ορθότητα.

- **Quality Filter:** Όπου οι μετρικές διίστανται (π.χ. υψηλό TF-IDF αλλά χαμηλό SBERT), να προωθείται για ανθρώπινη επιθεώρηση

- **Learning & Adaptation:** Χρήση supervised learning για ταξινόμηση τύπων λαθών και reinforcement/active learning για βελτίωση κανόνων και fine-tuning μοντέλων σε domain-specific δεδομένα. Αυτό ευθυγραμμίζεται με τις προτάσεις αυτοματοποίησης στο έγγραφο

4.4 Υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων, βιβλιοθηκών;

Ναι, υπήρξαν διαφορές στην ποιότητα ανακατασκευής μεταξύ τεχνικών, μεθόδων, βιβλιοθηκών. Τα δεδομένα δείχνουν τις παρακάτω διαφορές:

- **Rule-based reconstruction:** Εμφάνισε εξαιρετικά αποτελέσματα σε token-metrics,BERTScore και σε πολλές περιπτώσεις υψηλές SBERT τιμές. Τα πλεονεκτήματα είναι predictability, υψηλή διατήρηση σημασιολογίας σε συγκεκριμένα λάθη και χαμηλό WER. Ωστόσο περιορίζεται σε καλά ορισμένες περιπτώσεις και είναι brittle σε εκτεταμένες γραμματικές μετατροπές

- **Grammar corrector:** Καλό στη διατήρηση του πρωτοτύπου με μικρές διορθώσεις. Στις μετρήσεις αναφέρεται ότι είχε υψηλό BERTScore, καθιστώντας το αξιόπιστο για

"preserve meaning". Ωστόσο αποτυγχάνει σε αναδιαρθρώσεις που απαιτούν παραφράσεις

- **BART (paraphrase):** Προσφέρει καλύτερη ροή και native-like διατύπωση, μερικές φορές θυσιάζοντας μικρές πληροφορίες. Στα πειράματα έδειξε καλές sentence-level ιδιότητες αλλά και αστάθεια (occasionally verbosity)

- **T5 (paraphraser):** Πολύ δημιουργικό, χρήσιμο για simplification, αλλά με μεγαλύτερο ρίσκο απώλειας λεπτομερειών

- **Embeddings (GloVe, FastText, Word2Vec, SBERT):** Για αξιολόγηση, τα GloVe και FastText έδωσαν πολύ υψηλές word-level ομοιότητες, γεγονός που σημαίνει ότι αν η αξιολόγηση στηρίζεται αποκλειστικά σε αυτά, μπορεί να υπερεκτιμηθεί η επιτυχία. Το SBERT έδειξε μεγαλύτερη ευαισθησία και διαφοροποίηση, χρήσιμη για sentence-level αξιολόγηση

4.5 Συνοπτικό συμπέρασμα της συζήτησης

Τα πειραματικά δεδομένα του project δείχνουν ότι τα word embeddings είναι πολύ χρήσιμα εργαλεία για την ποσοτική εκτίμηση της σημασιολογικής ποιότητας και ότι κάθε μέθοδος ανακατασκευής έχει σαφώς διακριτό ρόλο. Η πιο αξιόπιστη πρακτική που προκύπτει είναι η υβριδική αυτοματοποίηση με πολλαπλές μετρικές αξιολόγησης και μηχανισμό human-in-the-loop για περιπτώσεις διχογνωμίας, λύση που εξισορροπεί διατήρηση νοήματος, γραμματική ορθότητα και βελτίωση ροής.

5. Συμπέρασμα

5.1 Αναστοχασμός Ευρημάτων

Η παρούσα μελέτη κατέδειξε ότι η σημασιολογική ανακατασκευή κειμένων μέσω τεχνικών Επεξεργασίας Φυσικής Γλώσσας (NLP) δύναται να επιτύχει υψηλή διατήρηση νοήματος, όταν πραγματοποιείται συνδυαστική προσέγγιση που ενσωματώνει κατάλληλες μεθόδους και πολυδιάστατες μετρικές αξιολόγησης. Τα πειραματικά δεδομένα ανέδειξαν την δύναμη sentence-level μοντέλων, όπως το SBERT, στην αποτύπωση της σημασιολογικής εγγύτητας, ενώ οι word-level τεχνικές (GloVe, FastText) αποδείχθηκαν χρήσιμες για τον εντοπισμό λεξιλογικής ομοιότητας, χωρίς ωστόσο να αποδίδουν πάντα την πλήρη έκταση της νοηματικής διατήρησης.

Η συγκριτική αξιολόγηση επιβεβαίωσε ότι καμία μεμονωμένη μέθοδος δεν υπερτερεί καθολικά: οι rule-based διαδικασίες διασφαλίζουν συνέπεια σε προβλέψιμα σφάλματα, οι grammar correctors είναι αποτελεσματικοί για στοχευμένες μικροδιορθώσεις, ενώ οι παραφραστήρες (BART, T5) ενισχύουν τη ροή και την εκφραστική φυσικότητα, εισάγοντας ωστόσο ελαφρές τροποποιήσεις στο περιεχόμενο. Η διαδοχική ή συνδυαστική εφαρμογή αυτών των τεχνικών, σε συνδυασμό με αξιολόγηση τόσο σε επίπεδο λέξης όσο και πρότασης, αναδείχθηκε η πλέον αξιόπιστη στρατηγική.

Παράλληλα, η έρευνα ανέδειξε κρίσιμες προκλήσεις, όπως η επίλυση αμφισημιών που απαιτούν εκτενή συμφραζόμενα, η ισορροπία μεταξύ πιστότητας και φυσικότητας, καθώς και η ανάγκη για ευέλικτες, αυτοματοποιημένες διαδικασίες προσαρμοσμένες σε διαφορετικά είδη κειμένων και θεματικά πεδία. Επιπλέον, η μονομερής εξάρτηση από ορισμένες μετρικές ενέχει τον κίνδυνο παραπλανητικών εκτιμήσεων, καθιστώντας αναγκαία την ανθρώπινη επιμέλεια σε εφαρμογές υψηλής ευαισθησίας.

Συνοψίζοντας, η επιτυχής σημασιολογική ανακατασκευή προϋποθέτει υβριδική και προσαρμόσιμη μεθοδολογία, η οποία συνδυάζει την υπολογιστική ισχύ προηγμένων NLP μοντέλων με διαρκή αξιολόγηση βάσει ποσοτικών δεικτών και ποιοτικής ανθρώπινης κρίσης. Μελλοντικές έρευνες θα μπορούσαν να εστιάσουν στην ενοποίηση αυτών των στοιχείων μέσω εξελιγμένων τεχνικών μεταφοράς μάθησης και context-aware μηχανισμών, με στόχο την επίτευξη μεγαλύτερης ακρίβειας και γενικευσιμότητας σε ποικίλα γλωσσικά περιβάλλοντα.

6. Βιβλιογραφία

6.1 Θεωρητικές Αναφορές

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.
2. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv preprint arXiv:1910.13461*.
3. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67.
4. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*.
5. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543.
6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
7. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135-146.
8. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.
9. Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. *arXiv preprint arXiv:2005.05909*.
10. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

11. Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. *Zenodo*.
12. Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks*.
13. Koehn, P. (2009). *Statistical machine translation*. Cambridge University Press.
14. Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.
15. Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing* (3rd ed. draft). Pearson.