ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία
Μεταπτυχιακού Διπλώματος Ειδίκευσης

«*Εξαγωγή και αξιολόγηση κανόνων παράφρασης*»

Γεωργία Κωνσταντίνου

Επιβλέπων: Ίων Ανδρουτσόπουλος

ΑΘΗΝΑ, ΙΟΥΝΙΟΣ 2011

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή 3
1.1 Το θέμα της εργασίας 3
1.2 Διάρθρωση του υπόλοιπου κειμένου 4
Κεφάλαιο 2: Μέθοδοι εξαγωγής παραφράσεων 5
2.1 Η μέθοδος των Bannard κ.ά. 5
2.2 Η μέθοδος του Callison–Burch 7
2.3 Η μέθοδος των Zhao κ.ά. 8
2.4 Η μέθοδος των Kok κ.ά. 12
Κεφάλαιο 3:Αξιολόγηση κανόνων παράφρασης 16
3.1 Χρησιμότητα και είδη μεθόδων αξιολόγησης κανόνων παράφρασης 16
3.2 Χειρωνακτική αξιολόγηση κανόνων παράφρασης 17
3.3 Αυτόματη αξιολόγηση κανόνων παράφρασης 20
Κεφάλαιο 4: Μια νέα μέθοδος αξιολόγησης κανόνων παράφρασης 22
4.1 Δεδομένα που χρησιμοποιήθηκαν στην εργασία 22
4.2 Αλγόριθμοι μηχανικής μάθησης 23
4.3 Γνωρίσματα 23
4.3.1 Γνωρίσματα γλωσσικού μοντέλου 24
4.3.2 Γνωρίσματα σημειακής αμοιβαίας πληροφορίας 26
4.3.3 Γνωρίσματα ανάλυσης λανθάνουσας σημασίας 30
Κεφάλαιο 5: Πειραματικά δεδομένα και αποτελέσματα 34
5.1 Περισσότερες πληροφορίες για τα πειραματικά δεδομένα 34
5.2 Πειράματα με γραμμικό διαχωριστή 35
5.2.1 Σύστημα σύγκρισης 35
5.2.2 Αποτελέσματα ταξινομητή μεγίστης εντροπίας 36
5.3 Πειράματα με παλινδρόμηση διανυσμάτων υποστήριξης 39
Κεφάλαιο 6: Συμπεράσματα και μελλοντική εργασία 47
Βιβλιογραφία 49

Κεφάλαιο 1: Εισαγωγή

1.1 Το θέμα της εργασίας

Ως αυτόματη εξαγωγή παραφράσεων εννοούμε την αυτόματη εξαγωγή ζευγών φράσεων, προτάσεων ή προτύπων (patterns) φράσεων από σώματα κειμένων, ώστε τα μέλη του κάθε ζεύγους να έχουν το ίδιο ή πολύ παρόμοιο νόημα (1). Τα παραγόμενα ζεύγη καλούνται συχνά και *κανόνες παράφρασης*. Παραδείγματα τέτοιων κανόνων είναι οι παρακάτω:

maintaining NN_1 ⇔ upholding NN_1
unrelated to NNP_1⇔ not relevant to NNP_1

Στους παραπάνω δύο κανόνες, έχουμε ζεύγη προτύπων φράσεων. Τα πρότυπα αποτελούνται από λέξεις και υποδοχές (slots) λέξεων (NN_1, NNP_1) που αντιστοιχούν σε συντακτικές κατηγορίες. Οι υποδοχές ταιριάζουν (μπορούν να γεμίσουν) με εκφράσεις των αντίστοιχων κατηγοριών. Για παράδειγμα, το αριστερό μέρος του δεύτερου από τους παραπάνω κανόνες ταιριάζει με την παρακάτω πρόταση:

*His comment was **unrelated to** the topic of the thread and was deleted.*

Χρησιμοποιώντας το δεξί μέρος του κανόνα, μπορούμε να δημιουργήσουμε την ακόλουθη παράφραση:

*This topic was **not relevant to** the topic of the thread and was deleted.*

Έχουν προταθεί πολλές μέθοδοι για την αυτόματη εξαγωγή παραφράσεων από σώματα κειμένων. Ένα πρόβλημα, όμως, που δεν έχει ακόμη διερευνηθεί επαρκώς είναι ότι οι παραγόμενοι κανόνες παράφρασης ενδέχεται να είναι ή να μην είναι σωστοί, ανάλογα με τα συμφραζόμενα. Για παράδειγμα, οι Szpektor κ.ά. (2) παρατηρούν ότι ένας κανόνας της παρακάτω μορφής:

X acquire Y ⇔ X buy Y

είναι σωστός σε πολλές περιπτώσεις, αλλά δεν πρέπει να εφαρμοστεί σε μια πρόταση όπως:

Children acquire language quickly.

Παρομοίως, οι Μαλακασιώτης και Ανδρουτσόπουλος (3) σχολιάζουν πως ο παρακάτω κανόνας δεν πρέπει να εφαρμοστεί σε προτάσεις για μπαταρίες.

X charged Y with ⇔ X accused Y of

Η παρούσα εργασία (α) μελετά μεθόδους εξαγωγής κανόνων παράφρασης που έχουν προταθεί στη βιβλιογραφία και (β) διερευνά τρόπους με τους οποίους είναι δυνατόν η απόφαση εφαρμογής ή όχι ενός κανόνα παράφρασης να λαμβάνεται αυτόματα, εξετάζοντας τα συμφραζόμενα και χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης.

1.2 Διάρθρωση του υπόλοιπου κειμένου

Η διάρθρωση του υπολοίπου κειμένου της εργασίας έχει ως εξής:

- Στο Κεφάλαιο 2 παρουσιάζονται προηγούμενες μέθοδοι εξαγωγής παραφράσεων.
- Στο Κεφάλαιο 3 παρουσιάζονται μέθοδοι που έχουν προταθεί για την αυτόματη ή χειρωνακτική αξιολόγηση κανόνων παράφρασης.
- Το Κεφάλαιο 4 περιγράφει μια μέθοδο που αναπτύχθηκε στη διάρκεια της εργασίας, προκειμένου να είναι δυνατόν η απόφαση εφαρμογής ή όχι ενός κανόνα παράφρασης να λαμβάνεται αυτόματα και λαμβάνοντας υπόψη τα συμφραζόμενα..
- Στο Κεφάλαιο 5 παρουσιάζονται πειραματικά αποτελέσματα της μεθόδου της εργασίας.
- Στο Κεφάλαιο 6 συνοψίζονται τα συμπεράσματα της εργασίας και προτείνονται μελλοντικές κατευθύνσεις έρευνας.

Κεφάλαιο 2: Μέθοδοι εξαγωγής παραφράσεων

Σε αυτό το κεφάλαιο παρουσιάζονται προηγούμενες μέθοδοι εξαγωγής παραφράσεων από σώματα κειμένων.

2.1 Η μέθοδος των Bannard κ.ά.

Η μέθοδος των Bannard κ.ά. (4) ήταν μια από τις πρώτες μεθόδους εξαγωγής παραφράσεων που αξιοποίησαν τεχνικές και σώματα κειμένων από το χώρο της στατιστικής μηχανικής μετάφρασης (5). Χρησιμοποιεί δίγλωσσα παράλληλα σώματα κειμένων και τεχνικές ευθυγράμμισης (alignment) φράσεων των δύο γλωσσών.

Ένα δίγλωσσο παράλληλο σώμα κειμένων αποτελείται από ζεύγη κειμένων (t_1, t_2), όπου το κάθε t_1 είναι κείμενο της μιας γλώσσας, κάθε t_2 είναι κείμενο της άλλης γλώσσας, και το ένα κείμενο είναι μετάφραση του άλλου. Στην περίπτωσή μας, η μια γλώσσα του παράλληλου σώματος είναι αυτή για την οποία θέλουμε να εξαγάγουμε παραφράσεις (αγγλικά στα πειράματα των Bannard κ.ά.) και η δεύτερη μπορεί να είναι οποιαδήποτε άλλη γλώσσα, την οποία καλούμε *γλώσσα-άξονα* (pivot language, γερμανικά στα πειράματα των Bannard κ.ά.).

Έχουν αναπτυχθεί στο χώρο της στατιστικής μηχανικής μετάφρασης τεχνικές που ευθυγραμμίζουν τις προτάσεις του κάθε t_1 με τις αντίστοιχες προτάσεις του t_2 (sentence alignment), καθώς και τεχνικές που ευθυγραμμίζουν τις αντίστοιχες λέξεις ή φράσεις δύο προτάσεων (word, phrase alignment) (6) (7) (8). Οι Bannard κ.ά. θεωρούν ότι αν μια φράση e_1 της αρχικής γλώσσας ευθυγραμμίζεται συχνά με μια φράση f της γλώσσας-άξονα και η f ευθυγραμμίζεται συχνά με μια φράση e_2 της αρχικής γλώσσας (διαφορετική της e_1), τότε είναι πολύ πιθανό οι e_1 και e_2 να αποτελούν η μία παράφραση της άλλης.

Οι μέθοδοι ευθυγράμμισης φράσεων που χρησιμοποιούν οι Bannard κ.ά. παράγουν έναν πίνακα της παρακάτω μορφής. Ο πίνακας περιέχει ζεύγη φράσεων e της αρχικής γλώσσας και φράσεων f της γλώσσας-άξονα, καθώς και μια στήλη $count(e, f)$ που δείχνει πόσες φορές ευθυγραμμίστηκαν οι e και f.

$\boldsymbol{f}$	$\boldsymbol{e}$	$count(e, f)$
ενέργεια	*energy*	$count(ενέργεια, energy)$
.	.	.
.	.	.
.	.	.
η επιτροπή έθεσε	*the commission set*	$count(η\ επιτροπή\ έθεσε, the\ commission\ se$

Πίνακας 1: Ευθυγράμμιση φράσεων.

Οι Bannard κ.ά. εκτιμούν την πιθανότητα $P(f|e)$ να αποτελεί η f μετάφρασή της e ως το πλήθος των ευθυγραμμίσεων της f με την e προς το πλήθος των ευθυγραμμίσεων της e με όλες τις άλλες φράσεις f' της γλώσσας-άξονα. Εντελώς αντίστοιχα εκτιμάται και η πιθανότητα $P(e|f)$.

$$P(f|e) = \frac{count(e, f)}{\sum_{f'} count(e, f')}$$

Η πιθανότητα $P(e_2|e_1)$ να αποτελεί η e_2 παράφραση της e_1 εκτιμάται με τον παρακάτω τύπο:

$$P(e_2|e_1) = \sum_f P(f|e_1)\, P(e_2|f, e_1) \approx \sum_f P(f|e_1)\, P(e_2|f)$$

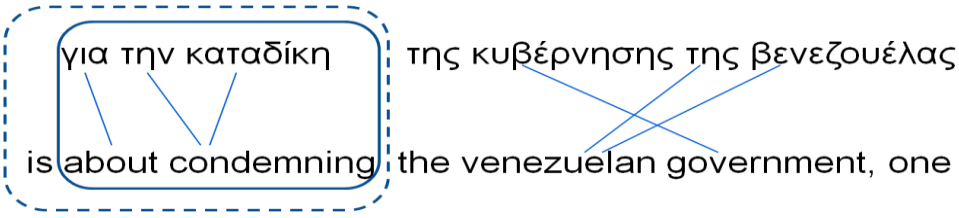
Ωστόσο στη παραπάνω διαδικασία ο υπολογισμός των πιθανοτήτων γίνεται αγνοώντας τα συμφραζόμενα των φράσεων e_2 και e_1. Προκειμένου να λαμβάνονται υπόψη τα συμφραζόμενα, σε μια παραλλαγή της μεθόδου τους οι Bannard κ.ά., εξετάζουν και την πιθανότητα που επιστρέφει ένα γλωσσικό μοντέλο (βλ. ενότητα 4.3.1) για την πρόταση *S* στην οποία εμφανίζεται η φράση e_1, αφού αντικατασταθεί η e_1 με την e_2.

Μια περαιτέρω βελτίωση χρησιμοποιεί πολλές γλώσσες – άξονες *F*. Στην περίπτωση αυτή, η καλύτερη παράφραση $\widehat{e_2}$ της e_1 είναι:

$$\widehat{e_2} = arg \max_{e_2 : e_2 \neq e_1} \sum_F \sum_{f \in F} P(f|e_1)\, P(e_2|f)$$

2.2 Η μέθοδος του Callison–Burch

Στη μέθοδο των Bannard κ.ά., που περιγράφηκε στην προηγούμενη ενότητα, δεν αξιοποιούνται καθόλου συντακτικοί αναλυτές (parsers), παρά μόνο (σε μια από τις επεκτάσεις της μεθόδου) ένα γλωσσικό μοντέλο *n*-γραμμάτων. Υπάρχουν περιπτώσεις, όμως, όπου ένα απλό γλωσσικό μοντέλο *n*-γραμμάτων δεν αρκεί. Το βασικό πρόβλημα είναι ότι η ευθυγράμμιση φράσεων εξάγει συχνά παραφράσεις που δεν ανήκουν στην ίδια συντακτική κατηγορία. Ένα παράδειγμα παρουσιάζεται στην Εικόνα 1, όπου η γλώσσα – άξονα είναι τα ελληνικά και η φράση «για την καταδίκη» ευθυγραμμίζεται τόσο με την «about condemning» όσο και με την «is about condemning». (Οι ευρετικές που χρησιμοποιεί η ευθυγράμμιση φράσεων θα επέτρεπαν στη λέξη «is» να συμπεριληφθεί στη φράση «is about condemning», που είναι μία από τις φράσεις με τις οποίες ευθυγραμμίζεται η «για την καταδίκη», έστω κι αν η «is» δεν ευθυγραμμίζεται άμεσα με καμία λέξη της «για την καταδίκη».)



Εικόνα 1

Αυτό έχει ως αποτέλεσμα να θεωρηθούν παραφράσεις οι «about condemning» και «is about condemning», συμπέρασμα που είναι λανθασμένο, με την έννοια ότι αν σε μια πρόταση αντικαταστήσουμε τη φράση «about condemning», που είναι εμπρόθετος προσδιορισμός, με την «is about condemning», η οποία είναι ρηματική φράση, τότε θα προκύψει μια πρόταση που δεν είναι συντακτικά ορθή, όπως φαίνεται στο παρακάτω παράδειγμα. Σημειώνουμε με αστερίσκο τη συντακτικά λανθασμένη πρόταση.

*Are there any Bible verses **about condemning** others?*

** Are there any Bible verses **is about condemning** others?*

Ακριβώς αυτό το πρόβλημα είναι που προσπαθεί να αντιμετωπίσει η μέθοδος του Callison–Burch (9), που λαμβάνει υπόψη της και τα συντακτικά δέντρα των φράσεων. Η καλύτερη παράφραση $\widehat{e_2}$ μιας φράσης e_1 γίνεται πλέον:

$$\widehat{e_2} = \arg \max_{e_2:\, e_2 \neq e_1 \wedge s(e_2)=s(e_1)} \sum_{F} \frac{\sum_{f \in F} P\left(f \middle| e_1, s(e_1)\right) P(e_2 | f, s(e_2))}{|C|}$$

όπου $s(e)$ η συντακτική ετικέτα (η κατηγορία της ρίζας του συντακτικού δέντρου) της φράσης e στη συγκεκριμένη πρόταση που εξετάζουμε και:

$$P\left(f \middle| e_1, s(e_1)\right) = \frac{count(f, e_1, s(e_1))}{\sum_{f'} count(f', e_1, s(e_1))}$$

$$P\left(e_2 \middle| f, s(e_2)\right) = \frac{count(f, e_2, s(e_2))}{\sum_{f'} count(f', e_2, s(e_2))}$$

Ουσιαστικά εξετάζεται πόσο συχνά οι e_1 και e_2 ευθυγραμμίζονται στην ίδια φράση f της γλώσσας-άξονα, όταν οι e_1 και e_2 έχουν τη συντακτική ετικέτα με την οποία χρησιμοποιείται η e_1 στην πρόταση που εξετάζουμε.

2.3 Η μέθοδος των Zhao κ.ά.

Οι Zhao κ.ά. (10) πρότειναν και αυτοί μια μέθοδο εξαγωγής κανόνων παράφρασης από παράλληλα κείμενα που χρησιμοποιεί γλώσσες-άξονες. Αντί όμως να εξάγουν ζεύγη φράσεων εξάγουν ζεύγη προτύπων φράσεων (patterns). Τα πρότυπα αποτελούνται από λέξεις και υποδοχές (slots) λέξεων, όπως φαίνεται στο παρακάτω παράδειγμα κανόνα παράφρασης.

NN_2 **is considered by** *NN_1* ⇔ *NN_1* **consider** *NN_2*

Σύμφωνα με τον παραπάνω κανόνα, αν σε κάποια πρόταση συναντήσουμε μια ονοματική φράση με κύριο ουσιαστικό NN_2 ακολουθούμενη από τη φράση «is considered by» και μια άλλη ονοματική φράση με κύριο ουσιαστικό NN_1, τότε μπορούμε να αντιστρέψουμε τις δύο ονοματικές φράσεις και να αντικαταστήσουμε την «is considered by» με τη λέξη «consider», όπως φαίνεται παρακάτω.

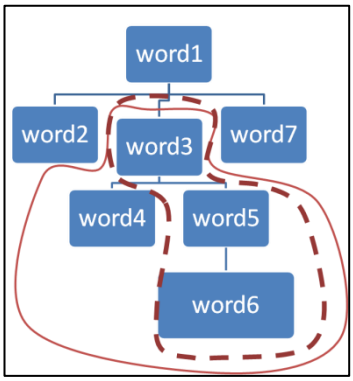
This ***is considered by*** *most collectors to be the best pen ever made.*

↓

Most collectors ***consider*** *this to be the best pen ever made.*

Οι Zhao κ.ά., χρησιμοποιούν στα πειράματά τους τα κινέζικα ως γλώσσα – άξονα και τα αγγλικά ως γλώσσα στην οποία θέλουμε παραφράσεις. Ένα στάδιο προεπεξεργασίας της μεθόδου τους, ευθυγραμμίζει τις προτάσεις και τις λέξεις των κειμένων και δημιουργεί τα συντακτικά δέντρα εξαρτήσεων (dependency trees) των αγγλικών προτάσεων.

Οι Zhao κ.ά. εξάγουν πρότυπα φράσεων εξετάζοντας όλα τα μερικά υποδέντρα του δέντρου εξαρτήσεων κάθε αγγλικής πρότασης, όπως αυτό της εικόνας 2. Ένα μερικό υποδέντρο ορίζεται ως ένα υποδέντρο που δεν περιέχει αναγκαστικά όλους τους απογόνους της ρίζας του. Από κάθε μερικό υποδέντρο προκύπτει και ένα πρότυπο. Στο παράδειγμα του σχήματος εξετάζεται το υποδέντρο word3 →word5 →word6, που περικλείεται από τη διακεκομμένη γραμμή. Για όλους τους απογόνους που δεν περιέχονται στο μερικό υποδέντρο, εντάσσεται στο πρότυπο φράσης η ετικέτα του μέρους του λόγου τους (part of speech tag, POS). Συνεπώς, θεωρώντας ότι η σειρά των λέξεων της αγγλικής πρότασης είναι …word3 word4 word5 word6…, το πρότυπο που προκύπτει από το παράδειγμα είναι το:



Εικόνα 2: Παράδειγμα δέντρου εξαρτήσεων

word3 POS(word4) word5 word6

Προς αποφυγή πολύπλοκων προτύπων, αν στο δέντρο του προτύπου που προκύπτει εμφανίζεται κάποια ετικέτα μέρους του λόγου ως απόγονος μιας άλλης ετικέτας, τότε η ετικέτα-απόγονος διαγράφεται.

Ολοκληρώνοντας αυτή τη διαδικασία, οι Zhao κ.ά. έχουν δημιουργήσει ένα σύνολο από αγγλικά πρότυπα φράσεων. Από κάθε αγγλικό πρότυπο δημιουργείται ένα αντίστοιχο κινέζικο εξετάζοντας την αγγλική πρόταση από την οποία προήλθε το αγγλικό πρότυπο και την ευθυγραμμισμένη με αυτήν κινέζικη. Οι λέξεις του αγγλικού προτύπου αντικαθίστανται από τις αντίστοιχες (ευθυγραμμισμένες) κινέζικες, ενώ οι ετικέτες μερών του λόγου διατηρούνται όπως στο αγγλικό πρότυπο. Το παραγόμενο κινέζικο πρότυπο θεωρείται ότι ευθυγραμμίστηκε

με το αγγλικό. Αν δύο αγγλικά πρότυπα φράσεων ευθυγραμμίζονται συχνά με το ίδιο κινέζικο, τότε θεωρούνται παραφράσεις.

Για τον υπολογισμό της πιθανότητας ευθυγράμμισης των αγγλικών προτύπων με τα κινέζικα πρότυπα-άξονα, οι Zhao κ.ά. προτείνουν τρία μοντέλα, με τα δύο τελευταία να αποτελούν επεκτάσεις των προηγουμένων. Παρακάτω περιγράφουμε το κάθε μοντέλο ξεχωριστά.

i. Μοντέλο 1

Στο πρώτο μοντέλο προτείνεται ένα λογαριθμικό – γραμμικό (log-linear) μοντέλο για τον υπολογισμό της πιθανότητας ευθυγράμμισης, διότι χειρίζεται το πρόβλημα αραιών δεδομένων (data sparseness) καλύτερα από τον αντίστοιχο υπολογισμό των Bannard κ.ά. Ο υπολογισμός της πιθανότητας παράφρασης ακολουθεί:

$$score_1(e_2|e_1) = \sum_{\mathrm{f}} exp\left[\sum_{i=1}^{4} \lambda_i h_i(e_1, e_2, f)\right]$$

Όπου:

f: πρότυπο της γλώσσας – άξονα,

$\boldsymbol{e_1, e_2}$**:** πρότυπα της γλώσσας στην οποία παράγουμε παραφράσεις,

$\boldsymbol{\lambda_i}$**:** βάρη των ιδιοτήτων (attributes)h_i,

$\boldsymbol{h_i}$**:** ιδιότητες που υπολογίζονται ως εξής:

$$h_1(e_1, e_2, f) = score_{MLE}(f|e_1)$$
$$h_2(e_1, e_2, f) = score_{MLE}(e_2|f)$$
$$h_3(e_1, e_2, f) = score_{LW}(f|e_1)$$
$$h_4(e_1, e_2, f) = score_{LW}(e_2|f)$$

Τα $score_{MLE}(e|f)$ και $score_{MLE}(f|e)$ είναι οι λογάριθμοι των αντίστοιχων πιθανοτήτων που υπολογίζουν οι Bannard κ.ά. (ενότητα 2.1). Τα $score_{LW}(f|e)$ και $score_{LW}(e|f)$, εξετάζουν πόσο καλά ευθυγραμμίζονται οι λέξεις των e_1 και e_2 με τις λέξεις της f και υπολογίζονται ως εξής:

$$score_{LW}(f|e) = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{1}{|\{j|(i,j) \in \alpha\}|}\sum_{\forall(i,j)\in\alpha} w(f_i|e_j)\right)$$

όπου:

$\boldsymbol{f_i}$**:** οι λέξεις της φράσης f,

$\boldsymbol{e_i}$**:** οι λέξεις της φράσης e,

$\boldsymbol{n}$**:** το πλήθος των λέξεων της f,

$\boldsymbol{\alpha}$**:** οι ευθυγραμμίσεις (alignments) μεταξύ των λέξεων της φράσης e και της φράσης f,

$\boldsymbol{w(f_i|e_j)}$**:** η πιθανότητα ευθυγράμμισης της f_iμε την e_i, που υπολογίζεται ως εξής:

$$w(f_i|e_j) = \frac{count(f_i|e_j)}{\sum_{f'_i} count(f'_i|e_j)}$$

Οι Zhao κ.ά. ορίζουν ένα κατώφλι T. Αν η πιθανότητα παράφρασης ανάμεσα σε δύο πρότυπα της αγγλικής γλώσσας ξεπερνά το T, τότε σχηματίζουν έναν κανόνα παράφρασης που περιλαμβάνει τα δύο πρότυπα.

ii. Μοντέλο 2

Οι Zhao κ.ά. παρατήρησαν ότι συχνά δύο πρότυπα της αρχικής γλώσσας (αγγλικά) αντιστοιχίζονται σε δύο διαφορετικά πρότυπα-άξονα, που έχουν όμως το ίδιο νόημα, είναι δηλαδή παραφράσεις. Για να συνυπολογίσουν και αυτή την περίπτωση, πρόσθεσαν στις ιδιότητες του προηγούμενου μοντέλου μια ιδιότητα που υπολογίζει την πιθανότητα δύο πρότυπα-άξονα f_1 και f_2 να αποτελούν παραφράσεις. Επομένως έχουμε τους παρακάτω υπολογισμούς για το μοντέλο 2 κατά αναλογία με εκείνους του μοντέλου 1:

$$score_2(e_2|e_1) = \sum_{f_1,f_2} exp\left[\sum_{i=1}^{5} \lambda_i h_i(e_1, e_2, f_1, f_2)\right]$$

Οι ιδιότητες $h_1 - h_4$ υπολογίζονται όπως και στο μοντέλο 1 (για τους συνδυασμούς $e_1,\ f_1$ και $e_2,\ f_2$), ενώ η h_5 υπολογίζεται ως εξής:

$$h_5(e_1, e_2, f_1, f_2) = score_1(f_2|f_1)$$

όπου $score_1(f_2|f_1)$ η πιθανότητα παράφρασης του μοντέλου 1, αντιστρέφοντας τη γλώσσα-άξονα με την αρχική (δηλαδή τώρα χρησιμοποιείται η αγγλική ως γλώσσα άξονα, για να βρούμε κινέζικες παραφράσεις).

iii. Μοντέλο 3

Σε αυτό το μοντέλο χρησιμοποιούνται όλες οι ιδιότητες του μοντέλου 2, αλλά προστίθενται δύο επιπλέον ιδιότητες που υπολογίζονται θεωρώντας όλους τους παραγόμενους κανόνες του μοντέλου 2 (ζεύγη αγγλικών προτύπων) ως ένα νέο σώμα παράλληλων κειμένων, στο οποίο εφαρμόζεται ευθυγράμμιση λέξεων. Οι δύο νέες ιδιότητες υπολογίζουν πόσο καλά ευθυγραμμίζονται οι λέξεις των e_1, e_2 στο νέο αυτό σώμα.

$$h_6(e_1, e_2, c_1, c_2) = score_{MWA}(e_2|e_1)$$
$$h_7(e_1, e_2, c_1, c_2) = score_{MWA}(e_1|e_2)$$

Όπου:

$\boldsymbol{score_{MWA}(e_2|e_1)}$**:** είναι η πιθανότητα ευθυγράμμισης των λέξεων των e_1, e_2 , όπως επιστρέφεται από τον αλγόριθμο ευθυγράμμισης λέξεων.

Οι Zhao κ.ά. ρύθμισαν τις παραμέτρους λ_i των ιδιοτήτων όλων των μοντέλων με ανάβαση κλίσης (11), μεγιστοποιώντας το μέτρο F_1, δηλαδή τον αρμονικό μέσο της ανάκλησης (recall) και της ακρίβειας (precision) των παραγόμενων ζευγών.

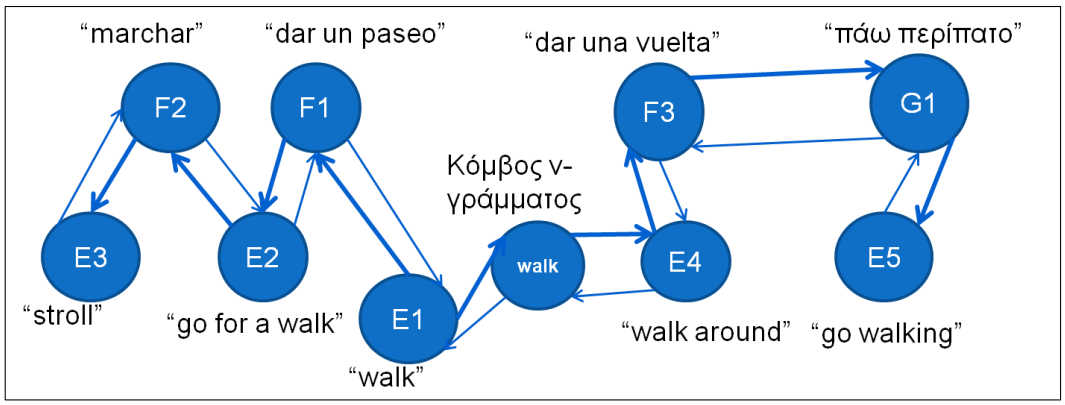
2.4 Η μέθοδος των Kok κ.ά.

Η μέθοδος των Kok κ.ά. (12) χρησιμοποιεί πίνακες ευθυγραμμισμένων φράσεων διαφορετικών γλωσσών, που έχουν εξαχθεί εκ των προτέρων από παράλληλα σώματα κειμένων. Για κάθε ζεύγος γλωσσών, υπάρχει ένας πίνακας που περιέχει ζεύγη ευθυγραμμισμένων φράσεων των δύο γλωσσών, όπως στη μέθοδο των Bannard κ.ά. (ενότητα 2.1). Οι Kok κ.ά. χρησιμοποίησαν στα πειράματά τους τρία σώματα παράλληλων κειμένων, ένα αγγλο-γαλλικό, ένα αγγλο-γερμανικό και ένα γαλλο-γερμανικό, οπότε υπήρχαν τρεις πίνακες με ζεύγη φράσεων.

Οι φράσεις όλων των πινάκων αναπαρίστανται ως κόμβοι ενός γράφου. Οι ακμές του γράφου παριστάνουν ευθυγραμμίσεις μεταξύ φράσεων, δηλαδή αντιστοιχούν σε γραμμές των

πινάκων. Το βάρος κάθε ακμής είναι η πιθανότητα ευθυγράμμισης της μιας φράσης με την άλλη, πιθανότητες οι οποίες υπολογίζονται επίσης εύκολα από τους πίνακες.

Στην πράξη, κάθε φορά που θέλουμε να παραφράσουμε μια φράση, ο γράφος πριονίζεται (ή δημιουργείται έτσι) ώστε να περιέχει κόμβους μέχρι ένα μέγιστο μήκος μονοπατιού από τον κόμβο της φράσης που θέλουμε να παραφράσουμε. Ένα παράδειγμα γράφου φαίνεται στην Εικόνα 3.



Εικόνα 3: Γράφος της μεθόδου των Kok κ.ά.

Σε μια βελτιωμένη μορφή της μεθόδου των Kok κ.ά., ο γράφος περιλαμβάνει και πρόσθετους κόμβους (feature nodes), οι οποίοι βοηθούν στην καλύτερη σύνδεση των υπολοίπων κόμβων μεταξύ τους και συνεπώς στη δημιουργία περισσότερων μονοπατιών. Οι κόμβοι αυτοί ανήκουν σε τρεις κατηγορίες και συνδέονται μόνο με κόμβους φράσεων της αρχικής γλώσσας (της γλώσσας στην οποία θέλουμε να δημιουργήσουμε παραφράσεις):

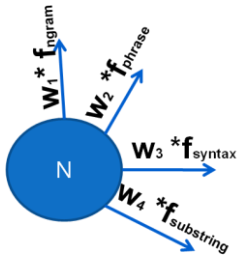
a) Κόμβοι ν-γραμμάτων: Για όλα τα ν-γράμματα ($1 \leq \nu \leq 4$) που περιέχονται στις φράσεις της αρχικής γλώσσας, προστίθενται κόμβοι ν-γραμμάτων στο γράφο. Κάθε κόμβος φράσης που περιέχει ένα ν-γραμμα συνδέεται με δύο κατευθυνόμενες ακμές με τον αντίστοιχο κόμβο ν-γραμμάτων (μία ακμή από τον κόμβο φράσης προς τον κόμβο ν-γράμματος και μία αντίθετη). Έτσι δημιουργούνται πιο σύντομα μονοπάτια μεταξύ κόμβων με κοινές λέξεις. Ως παράδειγμα, στην Εικόνα 3 φαίνεται ο κόμβος του 1-γράμματος «walk».

b) Συντακτικοί κόμβοι: Υπάρχουν δύο είδη πρόσθετων συντακτικών κόμβων, οι οποίοι δηλώνουν αν μια φράση ξεκινάει ή τελειώνει αντίστοιχα με μια λέξη που ανήκει σε μια

συγκεκριμένη συντακτική κατηγορία (π.χ. μια κατηγορία είναι τα άρθρα, μια άλλη οι ερωτηματικές λέξεις κτλ.). Κάθε κόμβος φράσης που ξεκινάει ή τελειώνει με μια λέξη που ανήκει σε κάποια συντακτική κατηγορία συνδέεται με δύο κατευθυνόμενες ακμές με τον αντίστοιχο συντακτικό κόμβο. Χάρη στους πρόσθετους αυτούς κόμβους δημιουργούνται πιο σύντομα μονοπάτια μεταξύ κόμβων φράσεων που ξεκινούν ή τελειώνουν με λέξεις των ίδιων συντακτικών κατηγοριών.

c) Κόμβος «όχι υπο-φράση / υπερ-φράση»: Αυτός ο (μοναδικός) κόμβος συνδέεται πάντα με τον κεντρικό κόμβο (τον κόμβο της φράσης που θέλουμε να παραφράσουμε) με δύο κατευθυνόμενες ακμές. Κάθε κόμβος φράσης που δεν είναι υπο-φράση ή υπερ-φράση της φράσης που θέλουμε να παραφράσουμε συνδέεται και αυτός με τον κόμβο «όχι υπο-φράση / υπερ-φράση» με δύο κατευθυνόμενες ακμές. Ο κόμβος «όχι υπο-φράση / υπερ-φράση» προέκυψε από την παρατήρηση ότι φράσεις που έχουν σχέση υπο-φράσης ή υπερ-φράσης έχουν υψηλή πιθανότητα ευθυγράμμισης, χωρίς όμως να είναι παραφράσεις. Έτσι αυτός ο κόμβος δημιουργήθηκε για να ευνοεί τις φράσεις που δεν είναι υπο-φράσεις ή υπερ-φράσεις της φράσης που θέλουμε να παραφράσουμε.

Έτσι ο κάθε κόμβος του γράφου έχει πλέον τέσσερα είδη εξερχόμενων ακμών, που φαίνονται στην Εικόνα 4.



Εικόνα 4

Σε κάθε ένα από τα τέσσερα είδη εξερχόμενων ακμών αντιστοιχεί ένα βάρος f_{phrase}, f_{ngram}, f_{syntax} και $f_{substring}$ με:

$$f_{phrase} + f_{ngram} + f_{syntax} + f_{substring} = 1$$

Οι τιμές των παραπάνω βαρών καθορίζονται με δοκιμές σε δεδομένα ρύθμισης.

Για κάθε ακμή μεταξύ κόμβων φράσεων, το βάρος της ακμής υπολογίζεται πλέον ως το γινόμενο της πιθανότητας ευθυγράμμισης των φράσεων με το βάρος f_{phrase}.

Για κάθε μια από τις k εξερχόμενες ακμές ενός κόμβου φράσης που οδηγούν σε κόμβους ν-γραμμάτων, το βάρος της ακμής είναι $^{\mathrm{f_{ngram}}}/_{\mathrm{k}}$. Αντίστοιχα για τις ακμές ενός κόμβου φράσης που οδηγούν σε συντακτικούς κόμβους και στον κόμβο «όχι υπο-φράση / υπερ-φράση».

Τέλος, για κάθε μια από τις k εξερχόμενες ακμές ενός πρόσθετου κόμβου (feature node), το βάρος είναι $^{1}/_{\mathrm{k}}$.

Ο αλγόριθμος των Kok κ.ά. στη συνέχεια υπολογίζει το μέσο χρόνο εύρεσης (hitting time) κάθε κόμβου φράσης j, όταν εκτελούνται τυχαίοι περίπατοι που ξεκινούν πάντα από τον κόμβο της φράσης που θέλουμε να παραφράσουμε. Ως τυχαίος περίπατος (13) ορίζεται μια διάσχιση του γράφου, κατά την οποία σε κάθε κόμβο επιλέγουμε την ακμή που θα ακολουθήσουμε με πιθανότητα ανάλογη προς το βάρος της. Ο μέσος χρόνος εύρεσης ενός κόμβου j ορίζεται ως το μέσο πλήθος βημάτων που απαιτούνται για να συναντήσουμε τον j για πρώτη φορά, ξεκινώντας από τον κόμβο της φράσης που θέλουμε να παραφράσουμε. Επειδή ο χρόνος ολοκλήρωσης μπορεί να τείνει στο άπειρο, οι Kok κ.ά. χρησιμοποιούν μια παραλλαγή του αλγορίθμου (14) (truncated hitting time), η οποία θέτει ένα άνω όριο βημάτων, έστω T, σε κάθε περίπατο. Οι καλύτερες πιθανές παραφράσεις της φράσης που θέλουμε να παραφράσουμε είναι εκείνες των κόμβων φράσεων της αρχικής γλώσσας που έχουν τους μικρότερους μέσους χρόνους εύρεσης.

Κεφάλαιο 3: Αξιολόγηση κανόνων παράφρασης

Στο κεφάλαιο αυτό παρουσιάζονται τρόποι που έχουν προταθεί για την αυτόματη ή χειρωνακτική αξιολόγηση των κανόνων παράφρασης που παράγουν οι μέθοδοι εξαγωγής παραφράσεων.

3.1 Χρησιμότητα και είδη μεθόδων αξιολόγησης κανόνων παράφρασης

Οι μέθοδοι του προηγουμένου κεφαλαίου παράγουν πολύ μεγάλο πλήθος κανόνων παράφρασης. Έτσι, αν θέλουμε κατόπιν να χρησιμοποιήσουμε τους κανόνες για να παραφράσουμε μια πρόταση, ενδέχεται να μπορούν να εφαρμοστούν πολλοί διαφορετικοί κανόνες ή συνδυασμοί τους, οδηγώντας σε μια πληθώρα από υποψήφιες παραφράσεις της πρότασης. Δεν είναι, όμως, συνήθως όλες οι υποψήφιες παραφράσεις εξίσου καλές, γιατί δεν είναι όλοι οι κανόνες παράφρασης εξίσου καλοί ή ενδέχεται να μην είναι όλοι κατάλληλοι αν ληφθούν υπόψη τα συμφραζόμενα. Χρειάζονται, επομένως, μέθοδοι που να αξιολογούν τους κανόνες παράφρασης. Οι μέθοδοι αυτές είναι είτε χειρωνακτικές (με ανθρώπους-κριτές που αξιολογούν τους κανόνες) είτε αυτόματες. Επίσης, ενδέχεται να λαμβάνουν ή όχι υπόψη τους τα συμφραζόμενα. Ο κεντρικός στόχος της παρούσας εργασίας ήταν η ανάπτυξη μιας μεθόδου αξιολόγησης κανόνων παράφρασης που να είναι αυτόματη και να λαμβάνει υπόψη της τα συμφραζόμενα. Η μέθοδος που αναπτύχθηκε τελικά στη διάρκεια της εργασίας και οι διάφορες παραλλαγές της περιγράφονται στο επόμενο κεφάλαιο. Σε αυτό το κεφάλαιο περιγράφουμε πρώτα προηγούμενες σχετικές μεθόδους αξιολόγησης κανόνων παράφρασης.

Σημειώνουμε ότι η αξιολόγηση κανόνων παράφρασης μπορεί να γίνει και έμμεσα, για παράδειγμα αξιολογώντας τις επιδόσεις συστημάτων που χρησιμοποιούν ή όχι τους κανόνες ή που χρησιμοποιούν κανόνες που έχουν παραχθεί από διαφορετικές μεθόδους εξαγωγής παραφράσεων (15) (16) (17). Στην παρούσα εργασία ασχολούμαστε κυρίως με την άμεση αξιολόγηση κανόνων παράφρασης, δηλαδή εξετάζουμε τους ίδιους τους κανόνες, όχι το πώς επηρεάζουν τις επιδόσεις συστημάτων στα οποία χρησιμοποιούνται. Λαμβάνουμε, όμως, υπόψη μας ότι οι κανόνες παράφρασης ενδέχεται να πρέπει να χρησιμοποιηθούν σε εφαρμογές με διαφορετικές απαιτήσεις, για παράδειγμα εφαρμογές όπου η ακριβής διατήρηση του νοήματος της αρχικής πρότασης ή η διατήρηση της συντακτικής ορθότητας είναι περισσότερο ή λιγότερο σημαντικές. Επιστρέφουμε σε αυτό το θέμα στο επόμενο κεφάλαιο. Τέλος, η αξιολόγηση των κανόνων παράφρασης είναι προφανώς χρήσιμη και για

τη βελτίωση ή σύγκριση των μεθόδων που τους παράγουν.

3.2 Χειρωνακτική αξιολόγηση κανόνων παράφρασης

Η χειρωνακτική αξιολόγηση των κανόνων παράφρασης από ανθρώπους-κριτές (17) (18) είναι η πιο αξιόπιστη μέθοδος, αλλά είναι δύσκολο να επαναλαμβάνεται συχνά. Επίσης, δεν είναι πάντα εύκολο για τους κριτές να καταλάβουν (ή να συμφωνήσουν) πώς πρέπει να αξιολογούν τους κανόνες.

Οι Szpector κ.ά. (18) πρότειναν μια χειρωνακτική προσέγγιση αξιολόγησης κανόνων κειμενικής συνεπαγωγής (textual entailment rules). Οι κανόνες κειμενικής συνεπαγωγής είναι ζεύγη προτύπων φράσεων, όπου το ένα μέλος του ζεύγους έπεται από το άλλο, ενώ το αντίθετο δεν συμβαίνει απαραίτητα. Ένα παράδειγμα είναι το παρακάτω:

$$X \textbf{ steal } Y \Rightarrow X \textbf{ get hold of } Y$$

Αν θέσουμε X = «Jonathan» και Y = «the phone», τότε αληθεύει σίγουρα η συνεπαγωγή:

$$Jonathan \textbf{ stole } the\ phone \Rightarrow Jonathan \textbf{ got hold of } the\ phone$$

Όπως προαναφέραμε, το αντίστροφο δεν αληθεύει απαραίτητα.

$$Jonathan \textbf{ got hold of } the\ phone \not\Rightarrow Jonathan \textbf{ stole } the\ phone$$

Με διαφορετικά συμφραζόμενα, ωστόσο, η συνεπαγωγή ενδέχεται να μην ισχύει, όπως στο ακόλουθο παράδειγμα.

$$Mary \textbf{ stole } his\ heart \Rightarrow Mary \textbf{ got hold of } his\ heart$$

Μπορούμε να θεωρήσουμε ότι οι παραφράσεις, με τις οποίες κυρίως ασχολούμαστε στην παρούσα εργασία, είναι αμφίδρομοι κανόνες κειμενικής συνεπαγωγής.

Στην πιο απλή περίπτωση, δίνεται στους ανθρώπους-κριτές ένα σύνολο κανόνων κειμενικής συνεπαγωγής και οι κριτές καλούνται να αξιολογήσουν κάθε κανόνα ως ορθό ή λανθασμένο.

Η συμφωνία, όμως, μεταξύ των κριτών (inter-annotator agreement) για τους ίδιους κανόνες έχει αναφερθεί πως είναι σχετικά μικρή σε αυτή την περίπτωση (19) (20) (21). Εν μέρει αυτό φαίνεται πως οφείλεται στο ότι διαφορετικοί κριτές υποθέτουν διαφορετικά συμφραζόμενα, με αποτέλεσμα να διαφέρουν οι αποφάσεις τους ως προς την ορθότητα των κανόνων.

Για να αυξηθεί η συμφωνία των κριτών, οι Szpector κ.ά. προτείνουν ο κάθε κανόνας κειμενικής συνεπαγωγής να συνοδεύεται από παραδείγματα εφαρμογής του, ώστε να αξιολογείται με συγκεκριμένα συμφραζόμενα. Τα παραδείγματα για κάθε κανόνα παράγονται αυτόματα, εξάγοντας από ένα σώμα κειμένων προτάσεις στις οποίες αυτός εφαρμόζεται. Οι Szpector κ.ά. θεωρούν ότι ένας κανόνας εφαρμόζεται σε μια πρόταση αν κάποιο υποδέντρο του συντακτικού δέντρου της πρότασης ταιριάζει με το δέντρο του αριστερού μέλους του κανόνα. Για παράδειγμα, ας θεωρήσουμε τον παρακάτω κανόνα:

$$X \text{ arrive } Y \Rightarrow X \text{ get } Y$$

Μια πρόταση στην οποία θα μπορούσε να εφαρμοστεί ο κανόνας είναι η παρακάτω:

"*He* ***arrived*** *there on time*"

από την οποία θα παράγονταν οι ακόλουθες δύο φράσεις:

he ***arrived*** *there* ⇒ *he* ***got*** *there*

Έχοντας μπροστά τους την πρόταση και τις παραγόμενες φράσεις, οι κριτές καλούνται να απαντήσουν στις παρακάτω ερωτήσεις:

1. «Έπεται η αριστερή φράση από την πρόταση;»
2. «Η δεξιά φράση είναι μια πιθανή φράση στα αγγλικά;»
3. «Έπεται η δεξιά φράση από την πρόταση;»

Οι ερωτήσεις υποβάλλονται με αυτή τη σειρά στους κριτές. Μόνο αν απαντήσουν θετικά στην πρώτη καλούνται να απαντήσουν και στις άλλες δύο. Επίσης, αν απαντήσουν αρνητικά στη δεύτερη ερώτηση, η απάντησή τους στην τρίτη δεν λαμβάνεται πάντα υπόψη (βλ. παρακάτω).

Στην πρώτη ερώτηση, αν η αριστερή φράση δεν έπεται από την πρόταση, αυτό είναι λάθος της αυτόματης παραγωγής του παραδείγματος και όχι του κανόνα, οπότε αυτό το παράδειγμα πρέπει να αγνοηθεί κατά την αξιολόγηση.

Η δεύτερη ερώτηση έχει στόχο να εντοπίσει περιπτώσεις στις οποίες το δεξί μέλος του κανόνα δεν έχει καμία σχέση με τα συμφραζόμενα. Ένα παράδειγμα θα ήταν ο κανόνας:

$$X \textbf{ shoot } Y \Rightarrow X \textbf{ kill } Y$$

Και η πρόταση:

The photographer often ***shot*** *sunsets.*

Όπου θα προέκυπτε η δεξιά φράση «The photographer often killed sunsets», που είναι πολύ απίθανη να εμφανιστεί σε κείμενα.

Η τελευταία ερώτηση είναι η κυριότερη και έχει στόχο να εντοπίσει αν ισχύει η κειμενική συνεπαγωγή στο συγκεκριμένο παράδειγμα.

Στη συνέχεια, οι Szpector κ.ά. υπολογίζουν ένα άνω και ένα κάτω φράγμα για την ακρίβεια του κάθε κανόνα. Ορίζουν το άνω φράγμα ακρίβειας ως το πλήθος των παραδειγμάτων στα οποία ισχύει η συνεπαγωγή του κανόνα (θετική απάντηση στο ερώτημα 3) προς το πλήθος των παραδειγμάτων στα οποία ο κανόνας παρήγαγε πιθανή δεξιά φράση (θετική απάντηση στο ερώτημα 2). Το κάτω φράγμα προκύπτει ομοίως, αλλά διαιρώντας με το πλήθος όλων των παραδειγμάτων στα οποία η αριστερή φράση έπεται από την πρόταση (θετική απάντηση στο ερώτημα 1, ανεξαρτήτως της απάντησης στο ερώτημα 2). Επομένως:

$$ακρίβεια\ (άνω\ φράγμα) = \frac{\#"ναι\ στο\ ερώτημα\ 3"}{\#"ναι\ στο\ ερώτημα\ 2"}$$

$$ακρίβεια\ (κάτω\ φράγμα) = \frac{\#"ναι\ στο\ ερώτημα\ 3"}{\#"ναι\ στο\ ερώτημα\ 1"}$$

Τέλος, οι Szpector κ.ά. ορίζουν ένα κατώφλι (στα πειράματά τους 80%) που όταν το ξεπερνάει η τιμή της ακρίβειας (με το κάτω ή το άνω φράγμα, ανάλογα με την αυστηρότητα

της αξιολόγησης), ο αντίστοιχος κανόνας θεωρείται ορθός.

3.3 Αυτόματη αξιολόγηση κανόνων παράφρασης

Σε άλλη εργασία, οι Szpector κ.ά. (2) πρότειναν μια μέθοδο αυτόματης αξιολόγησης κανόνων κειμενικής συνεπαγωγής, η οποία λαμβάνει υπόψη της τα συμφραζόμενα. Πιο συγκεκριμένα, η μέθοδος εξετάζει κατά πόσον μια αρχική πρόταση (ή άλλο κείμενο) *t* συνεπάγεται μια άλλη πρόταση *h* που προκύπτει εφαρμόζοντας στην *t* έναν κανόνα παράφρασης *r*. Η *h* επιτρέπεται να είναι και πρότυπο (template) πρότασης· τότε κατά την εφαρμογή του κανόνα παράφρασης, συμπληρώνονται και οι υποδοχές (slots) του προτύπου.

Στη μέθοδο αυτή, οι Szpector κ.ά. ορίζουν ως *cp(r), cp(t)* και *cp*(*h*) τις προτιμήσεις συμφραζομένων (contextual preferences) ενός κανόνα *r*, μιας αρχικής πρότασης *t* ή μιας τελικής πρότασης *h* αντίστοιχα. Ονομάζουν γενικά *στοιχείο* έναν κανόνα *r*, μια αρχική πρόταση *t* ή μια τελική πρόταση *h*. Οι πληροφορίες συμφραζομένων ενός στοιχείου περιέχουν δύο συνιστώσες, τα γενικά συμφραζόμενα (cp_g) και τα τοπικά συμφραζόμενα (cp_v).

Τα γενικά συμφραζόμενα cp_g ενός στοιχείου *z* αναπαριστούν τα συμφραζόμενα με τα οποία αυτό συναντάται συνήθως. Πιο συγκεκριμένα, $cp_g(z)$ είναι το άθροισμα των διανυσμάτων ανάλυσης λανθάνουσας σημασίας (ΑΛΣ, Latent Semantic Analysis, βλ. και ενότητα 4.3.3 παρακάτω) των χαρακτηριστικών όρων του *z*. Αν το *z* είναι κείμενο (*t* ή *h*), οι χαρακτηριστικοί όροι του είναι τα ουσιαστικά και ρήματα του κειμένου, πιο συγκεκριμένα οι ρίζες τους (stems) μαζί με τις ετικέτες μερών του λόγου τους (POS tags). Αν το *z* είναι κανόνας (*r*) της μορφής ΑΜ ⇒ ΔΜ, τότε οι χαρακτηριστικοί όροι του είναι τα ουσιαστικά και ρήματα (οι ρίζες τους και οι ετικέτες μερών του λόγου τους) που εμφανίζονται στα ΑΜ και ΔΜ.

Τα τοπικά συμφραζόμενα cp_v ορίζονται μόνο για πρότυπα *h* και κανόνες *r* και αναπαριστούν συγκεκριμένες προτιμήσεις λέξεων για τις μεταβλητές τους. Για παράδειγμα, αν έχουμε τον κανόνα:

X arrive Y ⇒ X get Y

η μεταβλητή Y «προτιμά» να παίρνει ως τιμές εκφράσεις όπως «here», «there» κτλ. για να έχει νόημα ο κανόνας. Τα cp_v χωρίζονται σε δύο λίστες, τη $cp_{v:e}$, η οποία περιλαμβάνει

πιθανές τιμές-λέξεις για τις μεταβλητές, και τη $cp_{v:\mathrm{n}}$, η οποία περιέχει τύπους ονομάτων οντοτήτων (Person, Location, Organization) που προτιμούν οι μεταβλητές. Η λίστα $cp_{v:e}$ κατασκευάζεται από χειρωνακτικά επισημειωμένα κείμενα, ενώ η $cp_{v:n}$ χρησιμοποιώντας έναν αναγνωριστή ονομάτων οντοτήτων (named entity recognizer).

Για την σύγκριση μεταξύ των γενικών συμφραζομένων cp_g δύο στοιχείων *α, β,* απλά υπολογίζεται η ομοιότητα συνημιτόνου των δύο αντίστοιχων διανυσμάτων ΑΛΣ. Οι Szpector κ.ά. ονομάζουν αυτό το μέτρο $m_g(\alpha, \beta)$. Ορίζουν, επίσης, ένα μέτρο ομοιότητας που συγκρίνει τα τοπικά συμφραζόμενα cp_v δύο στοιχείων, το οποίο καλούν $m_v(\alpha, \beta)$· δεν περιγράφουμε αυτό το μέτρο εδώ, επειδή είναι αρκετά περίπλοκο.

Τέλος, με βάση όλα τα παραπάνω, οι Szpector κ.ά. υπολογίζουν ένα συνολικό βαθμό *allCP* που δείχνει κατά πόσο η αρχική πρόταση *t* συνεπάγεται την τελική *h*, όταν η *h* προκύπτει από την *t* εφαρμόζοντας έναν κανόνα παράφρασης *r*. Ο βαθμός αυτός υπολογίζεται ως το παρακάτω γινόμενο· αν κάποιο από τα εμπλεκόμενα μέτρα δεν ορίζεται, η τιμή του θεωρείται ίση με τη μονάδα.

$$allCP = m_g(h,t) \cdot m_v(h,t) \cdot m_g(h,r) \cdot m_v(h,r) \cdot m_g(r,t) \cdot m_v(r,t)$$

Στη συνέχεια πολλαπλασιάζουν αυτό το βαθμό με το σκορ του κανόνα παράφρασης, όπως αυτό προκύπτει από τον αλγόριθμο εξαγωγής παραφράσεων.

Κεφάλαιο 4: Μια νέα μέθοδος αξιολόγησης κανόνων παράφρασης

Στο κεφάλαιο αυτό παρουσιάζουμε μια μέθοδο αξιολόγησης κανόνων παράφρασης που αναπτύχθηκε στη διάρκεια της παρούσας εργασίας. Η μέθοδος χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση. Δοκιμάσαμε διάφορα σύνολα γνωρισμάτων (features), με κάποια από τα οποία η μέθοδος λαμβάνει υπόψη της και τα συμφραζόμενα, δηλαδή την πρόταση στην οποία εφαρμόζεται ο κανόνας παράφρασης.

4.1 Δεδομένα που χρησιμοποιήθηκαν στην εργασία

Τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία προέρχονται από τη χειρωνακτική αξιολόγηση προτάσεων στις οποίες έχουν εφαρμοστεί ένας ή περισσότεροι κανόνες παράφρασης. Οι κανόνες έχουν εξαχθεί με τη μέθοδο των Zhao κ.ά. (10) (βλ. Ενότητα 2.3) και συνοδεύονται (ο καθένας) από τρεις βαθμούς (score) που έχουν προκύψει αυτόματα από τη διαδικασία εξαγωγής παραφράσεων. Πιο συγκεκριμένα, οι τρεις βαθμοί είναι η βαθμολογία παράφρασης από το μοντέλο 1 των Zhao κ.ά. και δύο βαθμολογίες παράφρασης από το μοντέλο 3 οι οποίες είναι οι πιθανότητες ευθυγράμμισης προς τις δύο κατευθύνσεις. Στους τρεις αυτούς βαθμούς προσθέτουμε και το μέσο όρο τους, ως τέταρτο βαθμό. Κανονικοποιούμε και τους τέσσερις βαθμούς στο [0, 1].

Το σύνολο των δεδομένων, το οποίο διατίθεται δημόσια, περιλαμβάνει τα ακόλουθα:[1]

- Ζεύγη προτάσεων της μορφής αρχική πρόταση – παραχθείσα πρόταση.
- Τους κανόνες παράφρασης οι οποίοι εφαρμόστηκαν στην αρχική πρόταση κάθε ζεύγους ώστε να προκύψει η παραχθείσα.
- Τους τέσσερις βαθμούς (βλ. παραπάνω) του κανόνα που εφαρμόστηκε σε κάθε ζεύγος ή τους μέσους όρους τους, αν εφαρμόστηκαν περισσότεροι από έναν κανόνες.
- Οι βαθμοί που έδωσαν άνθρωποι-κριτές σε κάθε ζεύγος. Κάθε κριτής έδωσε σε κάθε ζεύγος τρεις βαθμούς (όλοι από το 1 ως το 4). Ο πρώτος βαθμός αξιολογεί τη διατήρηση του νοήματος της αρχικής πρότασης στην τελική. Ο δεύτερος τη διατήρηση της γραμματικής ορθότητας. Ο τρίτος αξιολογεί τη συνολική ποιότητα της παράφρασης.

[1]Το σύνολο δεδομένων θα διατίθεται από την ιστοσελίδα: http://nlp.cs.aueb.gr/software_gr.html.

4.2 Αλγόριθμοι μηχανικής μάθησης

Στόχος μας ήταν να μπορούμε τελικά να αποφασίζουμε αυτόματα αν ένας κανόνας παράφρασης πρέπει ή όχι να εφαρμοστεί σε μια πρόταση. Έχοντας μια συλλογή από ήδη βαθμολογημένες (από ανθρώπους-κριτές) παραφράσεις που έχουν προκύψει εφαρμόζοντας κανόνες παράφρασης, αποφασίσαμε να χρησιμοποιήσουμε αλγορίθμους επιβλεπόμενης μάθησης και γνωρίσματα (features) που παρέχουν πληροφορίες τόσο για τον εξεταζόμενο κανόνα παράφρασης, όσο και για την πρόταση στην οποία εξετάζεται αν πρέπει να εφαρμοστεί. Πειραματιστήκαμε με έναν ταξινομητή μεγίστης εντροπίας (maximum entropy classifier) (22) και έναν αλγόριθμο παλινδρόμησης διανυσμάτων υποστήριξης (support vector regression, SVR).[2]

4.3 Γνωρίσματα

Για κάθε ζεύγος αρχικής-τελικής πρότασης (και τον ή τους κανόνες που μετέτρεψαν την αρχική πρόταση στην τελική), τα γνωρίσματα (features) που δοκιμάσαμε είναι τριών ειδών:

i. Γνωρίσματα που βασίζονται σε ένα γλωσσικό μοντέλο *n*-γραμμάτων (*n*-gram language model) (23) (24).
ii. Γνωρίσματα που βασίζονται στη σημειακή αμοιβαία πληροφορία (ΣΑΠ – Pointwise Mutual Information) (25).
iii. Γνωρίσματα που βασίζονται στην ανάλυση λανθάνουσας σημασίας (ΑΛΣ – Latent Semantic Analysis) (26) (27).

Επιπλέον, χρησιμοποιούνται ως γνωρίσματα και οι τέσσερις βαθμοί των κανόνων παράφρασης των Zhao κ.α. (10), στους οποίους αναφερθήκαμε στην προηγούμενη ενότητα. Επειδή σε ένα ζεύγος αρχικής – τελικής πρότασης μπορεί να έχουν χρησιμοποιηθεί περισσότεροι από έναν κανόνες, θεωρούμε ως γνωρίσματα:

iv. το μέσο όρο του κάθε βαθμού στους κανόνες που εφαρμόστηκαν (4 γνωρίσματα),

[2]Η υλοποίηση του ταξινομητή μεγίστης εντροπίας που χρησιμοποιήσαμε είναι του Πανεπιστημίου του Stanford και διατίθεται από τη διεύθυνση http://nlp.stanford.edu/software/classifier.shtml. Η υλοποίηση SVR που χρησιμοποιήσαμε περιλαμβάνεται στο πακέτο LIBSVM, που επίσης διατίθεται ελεύθερα από τη διεύθυνση http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3]Βλ. http://www.speech.sri.com/projects/srilm/και http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31/.

v. το μέγιστο κάθε βαθμού στους κανόνες που εφαρμόστηκαν (4 γνωρίσματα),
vi. τον ελάχιστο κάθε βαθμού στους κανόνες που εφαρμόστηκαν (4 γνωρίσματα).

Ακολουθεί αναλυτικότερη παρουσίαση των γνωρισμάτων (i) – (iii).

4.3.1 Γνωρίσματα γλωσσικού μοντέλου

Ένα γλωσσικό μοντέλο *n*-γραμμάτων υπολογίζει την πιθανότητα εμφάνισης μιας συγκεκριμένης ακολουθίας λέξεων σε (ορθά) κείμενα μιας φυσικής γλώσσας. Όσο μεγαλύτερη είναι η πιθανότητα, τόσο πιθανότερο είναι η ακολουθία λέξεων να αποτελεί γραμματικά σωστή έκφραση της φυσικής γλώσσας.

Ένα μοντέλο *n*-γραμμάτων εκπαιδεύεται σε ένα σώμα κειμένων της γλώσσας, από το οποίο εκτιμά τις πιθανότητες εμφάνισης όλων των *n*-γραμμάτων (ακολουθιών n λέξεων) της γλώσσας. Η πιθανότητα να συναντήσουμε μια ακολουθία λέξεων $w_1, w_2, \dots, w_m$ εκτιμάται κατόπιν με τον παρακάτω τύπο, στον οποίο γίνεται η παραδοχή ότι η πιθανότητα εμφάνισης κάθε λέξης εξαρτάται μόνο από τις προηγούμενες $n-1$ λέξεις της ακολουθίας· για να απλουστευθεί ο τύπος, θεωρούμε επίσης ότι κάθε ακολουθία ξεκινά με δύο ψευδο-λέξεις w_{-1}, w_0.

$$P(w_1, w_2, \dots, w_m) \cong \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \dots, w_{i-1})$$

Στην πιο απλή περίπτωση, οι πιθανότητες $P(w_i | w_{i-(n-1)}, \dots, w_{i-1})$ υπολογίζονται ως το πλήθος εμφανίσεων της ακολουθίας $w_{i-(n-1)}, \dots, w_{i-1}, w_i$ στο σώμα κειμένων προς το πλήθος εμφανίσεων της ακολουθίας$w_{i-(n-1)}, \dots, w_{i-1}$.

$$P(w_i | w_{i-(n-1)}, \dots, w_{i-1}) \cong \frac{C(w_{i-(n-1)}, \dots, w_{i-1}, w_i)}{C(w_{i-(n-1)}, \dots, w_{i-1})}$$

Πολλές ακολουθίες λέξεων, όμως, εμφανίζονται πολύ σπάνια ή καθόλου, ακόμα και σε πολύ μεγάλα σώματα κειμένων, με αποτέλεσμα πολλές από τις εκτιμήσεις πιθανοτήτων να είναι κακές και συχνά μηδενικές. Στην πράξη χρησιμοποιούνται τεχνικές εξομάλυνσης των εκτιμήσεων των πιθανοτήτων, καθώς και η παρακάτω λογαριθμική μορφή της $P(w_1,\ w_2,\ \dots,\ w_m)$, η οποία κανονικοποιεί επίσης το επιστρεφόμενο αποτέλεσμα ως προς το μήκος της

ακολουθίας.

$$\frac{1}{m}\log P(w_1, w_2, \dots, w_m) = \frac{1}{m}\sum_{i=1}^{m} \log P(w_i | w_{i-(n-1)}, \dots, w_{i-1})$$

Περισσότερες πληροφορίες για τα γλωσσικά μοντέλα *n*-γραμμάτων παρέχονται στη βιβλιογραφία (24). Στην παρούσα εργασία χρησιμοποιήθηκε το λογισμικό SRILM (23) για την εκπαίδευση ενός μοντέλου τριγραμμάτων (n = 3) πάνω σε 6,5 εκατομμύρια προτάσεις του σώματος κειμένων AQUAINT.[3]

Τα γνωρίσματα που προκύπτουν από το γλωσσικό μοντέλο είναι:

- Πιθανότητα γλωσσικού μοντέλου της αρχικής και παραχθείσας πρότασης:

  $\boldsymbol{lm}(αρχικής\ πρότασης)$
  $\boldsymbol{lm}(παραχθείσας)$

- Διαφορά πιθανότητας γλωσσικού μοντέλου των δύο προτάσεων:

  $\boldsymbol{lmDif} = lm(αρχικής\ πρότασης) - lm(παραχθείσας)$

  Υπολογίζουμε την διαφορά για να έχουμε μια αίσθηση πόσο επηρεάστηκε η γραμματική ποιότητα της αρχικής πρότασης από την εφαρμογή του κανόνα παράφρασης.

- Πιθανότητες γλωσσικού μοντέλου παραθύρων:
  Δημιουργούμε ένα «παράθυρο» το οποίο περιλαμβάνει το τμήμα της αρχικής πρότασης που ταιριάζει με το ένα μέρος του κανόνα, τις δύο προηγούμενες λέξεις και τις δύο επόμενες. Ομοίως, δημιουργούμε ένα παράθυρο που περιλαμβάνει το τμήμα της παραχθείσας πρότασης που ταιριάζει με το άλλο μέρος του κανόνα, τις δύο προηγούμενες και τις δύο επόμενες λέξεις. Υπολογίζουμε κατόπιν με το γλωσσικό μοντέλο τις πιθανότητες των δύο παραθύρων και τη διαφορά τους. Περιορίζουμε έτσι μέσω των παραθύρων τα συμφραζόμενα των κανόνων στις πολύ κοντινές τους λέξεις μόνο. Για

[3]Βλ. http://www.speech.sri.com/projects/srilm/και http://www.ldc.upenn.edu/Catalog/docs/LDC2002T31/.

παράδειγμα αν έχουμε το ακόλουθο ζεύγος προτάσεων:

"*He has spent much of his young* ***career*** *dominating opposing hitters while pitching for dominant teams.*"

↓

"*He has spent much of his young* ***employment*** *dominating opposing hitters while pitching for dominant teams.*"

που έχει προκύψει από τον κανόνα:

career *VBG_1*⇔***employment*** *VBG_1*

υπολογίζονται τα εξής γνωρίσματα:

$$\boldsymbol{lmWinS1} = \boldsymbol{lm}(his\ young\ career\ dominating\ opposing\ hitters)$$
$$\boldsymbol{lmWinS2} = \boldsymbol{lm}(his\ young\ employment\ dominating\ opposing\ hitters)$$
$$\boldsymbol{lmWinDif} = lmWinS1 - lmWinS2$$

4.3.2 Γνωρίσματα σημειακής αμοιβαίας πληροφορίας

Η σημειακή αμοιβαία πληροφορία (ΣΑΠ – Pointwise Mutual Information) είναι ένα μέτρο ομοιότητας που υπολογίζει κατά πόσον δύο λέξεις συνεμφανίζονται πάντα μαζί ή όχι σε τμήματα κειμένου (25). Το κάθε τμήμα μπορεί να είναι μια πρόταση, μια παράγραφος, μια ιστοσελίδα κ.τ.λ. Στην υλοποίηση της ΣΑΠ που χρησιμοποιήθηκε στην παρούσα εργασία, το κάθε τμήμα ήταν μια ιστοσελίδα.[4]

Η σημειακή αμοιβαία πληροφορία δύο λέξεων w_1 και w_2 υπολογίζεται ως εξής:

$$pmi(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

Όπου:

[4]Χρησιμοποιήσαμε την υλοποίηση ΣΑΠ της ιστοσελίδας http://cwl-projects.cogsci.rpi.edu/msr.

$\boldsymbol{P(w_1, w_2)}$**:** η πιθανότητα να βρεθούν στο ίδιο τμήμα οι λέξεις w_1, w_2.
$\boldsymbol{P(w_1)}$**:** η πιθανότητα να βρεθεί σε ένα τμήμα η λέξη w_1.
$\boldsymbol{P(w_2)}$**:** η πιθανότητα να βρεθεί σε ένα τμήμα η λέξη w_2.

Οι παραπάνω πιθανότητες μπορούν να εκτιμηθούν χρησιμοποιώντας ένα μεγάλο σώμα κειμένων (ή ιστοσελίδων)· στην υλοποίηση που χρησιμοποιήσαμε, οι εκτιμήσεις έχουν γίνει σε ένα πολύ μεγάλο σώμα κειμένου, τη συλλογή εγγράφων του Google. Συγκεκριμένα μετράται το πλήθος χτυπημάτων (hits) κατά την αναζήτηση κάθε όρου με τη μηχανή αναζήτησης της Google (28). Ασφαλώς η επιλογή του σώματος κειμένων μπορεί να επηρεάσει σημαντικά τις εκτιμήσεις.

Αν δύο λέξεις w_1, w_2 δεν συνεμφανίζονται ποτέ, τότε ο αριθμητής στο εσωτερικό του λογαρίθμου του $pmi(w_1, w_2)$ θα είναι ίσος με τον παρονομαστή και συνεπώς το αποτέλεσμα θα είναι μηδενικό. Αν οι δύο λέξεις εμφανίζονται πάντα μαζί, τότε $pmi(w_1, w_2) = -logP(w_1) = -logP(w_2)$. Στην υλοποίηση που χρησιμοποιήσαμε, οι τιμές του $pmi(w_1, w_2)$ είναι κανονικοποιημένες στο διάστημα [0,1].

Τα γνωρίσματα που βασίζονται στη ΣΑΠ είναι εννέα και χωρίζονται σε τρεις κατηγορίες:

Τρία γνωρίσματα ΣΑΠ που εξετάζουν κατά πόσον το αριστερό και το δεξιό μέρος του κανόνα ταιριάζουν με τα συμφραζόμενα

Για παράδειγμα, αν έχουμε την πρόταση:

This company ***produces*** *computer hardware.*

και τον κανόνα:

X ***produce*** *Y* ⇔ *X* ***lay*** *Y*

το δεξιό μέρος του κανόνα δεν ταιριάζει με τα συμφραζόμενα (*This company* ***lays*** *computer hardware*) και επομένως ο κανόνας αυτός δεν πρέπει να εφαρμοστεί στη συγκεκριμένη πρόταση.

Ο υπολογισμός του πρώτου από τα τρία αυτά γνωρίσματα γίνεται ως εξής:

$$PMI(r) = \frac{1}{n}\sum_{i=1}^{n}|PMI(w_i, LHS)| - \frac{1}{n'}\sum_{i=1}^{n'}|PMI(w_i', RHS)|$$

Όπου:

r: ο κανόνας παράφρασης τον οποίο εξετάζουμε,

$\boldsymbol{w_i}$**:** οι λέξεις της αρχικής πρότασης, αν εξαιρέσουμε τις λέξεις που ταιριάζουν με τον κανόνα (χωρίς, όμως, να εξαιρέσουμε τις λέξεις που ταιριάζουν με τις υποδοχές του κανόνα) και τις πολύ συχνές λέξεις της γλώσσας (stopwords),[5]

$\boldsymbol{w_i'}$**:** οι λέξεις της παραχθείσας πρότασης, αν εξαιρέσουμε τις λέξεις του κανόνα (αλλά όχι των υποδοχών του) και τις πολύ συχνές λέξεις της γλώσσας,

n: το πλήθος των λέξεων $\mathbf{w_i}$,

$\boldsymbol{n'}$**:** το πλήθος των λέξεων $\boldsymbol{w_i'}$,

LHS/RHS: οι λέξεις που αποτελούν το αριστερό/δεξί μέλος του κανόνα (Left Hand Side, Right Hand Side), χωρίς τις υποδοχές του κανόνα,

$\boldsymbol{PMI(w, ruleside)}$: είναι η ΣΑΠ μεταξύ μιας λέξης w και του αριστερού/δεξιού μέρος του κανόνα(LHS, RHS), όπου:

$$PMI(w, ruleside) = \frac{1}{m}\sum_{i=1}^{m} PMI(w, ruleside_i)$$

m**:** το πλήθος των λέξεων του *ruleside*,

w**:** η λέξη των συμφραζομένων που εξετάζουμε,

$\boldsymbol{ruleside_i}$**:** η i-στη λέξη του *ruleside*.

Όπως έχουμε ήδη αναφέρει, σε μια πρόταση ενδέχεται να μπορούν να εφαρμοστούν πολλοί κανόνες ταυτόχρονα. Το $PMI(r)$ υπολογίζεται ανεξάρτητα για κάθε κανόνα που εφαρμόζεται. Για παράδειγμα, έστω η πρόταση:

"*The two metro Atlanta business* ***owners*** *offer hard–to–find* ***items.***"

Μπορούν να εφαρμοστούν οι παρακάτω κανόνες:

[5]Βλ. http://www.textfixer.com/resources/common-english-words.txt.

JJ_1 **items** ⇔ JJ_1 **projects**

NN_1 **owners** ⇔ NN_1 **proprietors**

Για τον πρώτο κανόνα έχουμε:

LHS = items

RHS = projects

$$w_i \in \{\text{two}, \text{metro}, \text{Atlanta}, \text{business}, \text{owners}, \text{offer}, \text{hard–to–find}\}$$

Για το δεύτερο κανόνα έχουμε:

LHS = owners

RHS = proprietors

$w_i \in \{\text{two}, \text{metro}, \text{Atlanta}, \text{business}, \text{offer}, \text{hard–to–find}, \text{items}\}$

Για κάθε κανόνα παράφρασης r_i που εφαρμόζεται στην πρόταση που εξετάζουμε, υπολογίζουμε τα τρία παρακάτω γνωρίσματα, όπου k το πλήθος των κανόνων που εφαρμόζονται:

i. $\boldsymbol{PMI} = \frac{1}{k}\sum_{i=1}^{k} PMI(r_i)$

ii. $\boldsymbol{PMImax} = max_{i\in\{1,\dots,k\}}\, PMI(r_i)$

iii. $\boldsymbol{PMImin} = min_{i\in\{1,\dots,k\}}\, PMI(r_i)$

Τρία γνωρίσματα ΣΑΠ που εξετάζουν τη συνολική «συνοχή» της αρχικής και τελικής πρότασης

Τα γνωρίσματα αυτά είναι παρόμοια με τα προηγούμενα, αλλά εξετάζουν το μέσο $pmi(w_1, w_2)$ όλων των ζευγών λέξεων της αρχικής και της τελικής πρότασης. Αρχικά αφαιρούνται και πάλι από τις δύο προτάσεις οι συχνές λέξεις (stopwords). Στη συνέχεια υπολογίζεται το μέσο $pmi(w_1, w_2)$ της αρχικής και ομοίως της τελικής πρότασης:

$$GPMI(s) = \frac{1}{n} \sum_{i,j \in [1,n], i \neq j} pmi(w_i, w_j)$$

Όπου:

s: η πρόταση που εξετάζουμε,

n: το πλήθος των λέξεων της πρότασης, χωρίς τις πολύ συχνές λέξεις (stopwords).

Για παράδειγμα στην πρόταση:

s : "~~*The*~~ *two metro Atlanta business owners offer hard–to–find items.*"

Ο υπολογισμός του ΣΑΠ θα γινόταν ως:

$$GPMI(s) = \frac{1}{8}(pmi(two, metro) + pmi(two, Atlanta) + \cdots)$$

Προκύπτουν τα ακόλουθα γνωρίσματα:

iv. ***GPMI****(αρχικής πρότασης)*

v. ***GPMI****(τελικής πρότασης) και*

vi. ***GPMIdif*** *= GPMI(αρχικής πρότασης)-GPMI(τελικής πρότασης)*

Τρία γνωρίσματα που εξετάζουν τη συνολική «συνοχή» σημαντικών λέξεων της αρχικής και τελικής πρότασης

Τα γνωρίσματα αυτά είναι τα ίδια με τα τρία προηγούμενα, με τη διαφορά ότι λαμβάνονται υπόψη μόνο τα επίθετα, ουσιαστικά, κύρια ονόματα, επιρρήματα και ρήματα των προτάσεων.[6]

4.3.3 Γνωρίσματα ανάλυσης λανθάνουσας σημασίας

Η ανάλυση λανθάνουσας σημασίας (ΑΛΣ – Latent semantic analysis) (26) είναι μια

[6]Χρησιμοποιούμε τον επισημειωτή μερών του λόγου (POStagger) του Stanford που διατίθεται από τη διεύθυνση http://nlp.stanford.edu/software/tagger.shtml

στατιστική μέθοδος η οποία εντοπίζει συσχετίσεις ανάμεσα σε λέξεις ή μεγαλύτερα κομμάτια κειμένου με βάση τις κοινές εμφανίσεις τους σε κείμενα. Βασίζεται στην ανάλυση πινάκων και την ανασύνθεσή τους με συμπιεσμένη μορφή λιγότερων διαστάσεων.

Η επεξεργασία αυτή συνίσταται στην κατασκευή ενός πίνακα Α με στήλες τις λογικές ενότητες του σώματος κειμένων, π.χ. προτάσεις, παραγράφους ή έγγραφα. Κάθε γραμμή του πίνακα περιλαμβάνει μια μοναδική λέξη του σώματος κειμένων. Τα κελιά του πίνακα περιέχουν την συχνότητα (π.χ. tf-idf) με την οποία οι λέξεις εμφανίζονται στις αντίστοιχες ενότητες.

Στον πίνακα Α διαστάσεων m x n εφαρμόζουμε παραγοντοποίηση ιδιαζουσών τιμών (Singular Value Decomposition, SVD), μια μέθοδο η οποία διασπά τον αρχικό πίνακα σε τρεις καινούριους, ώστε αν τους πολλαπλασιάσουμε να σχηματίζουν τον αρχικό πίνακα.

$$\boldsymbol{A = T \cdot S \cdot D^T}$$

Πιο συγκεκριμένα, ο πίνακας S είναι ένας διαγώνιος πίνακας[7](rectangular diagonal matrix) που περιέχει τις ιδιάζουσες τιμές (singular values) του αρχικού πίνακα **A** σε φθίνουσα διάταξη**.** Οι δύο πίνακες ***T, D*** είναι ορθοκανονικοί και περιέχουν τα ιδιοδιανύσματα του $\boldsymbol{AA^T}$ και του $\boldsymbol{A^TA}$ αντίστοιχα. Θυμίζουμε ότι ένας πίνακας είναι ορθοκανονικός όταν ο ανάστροφός του (transpose) πίνακας είναι ίδιος με τον αντίστροφό του (inverse).

Αν κρατήσουμε στο διαγώνιο πίνακα $\boldsymbol{S}$ μόνο τις *k* μεγαλύτερες ιδιάζουσες τιμές, δημιουργώντας τον πίνακα $\boldsymbol{S_k}$, και κρατήσουμε στους πίνακες $\boldsymbol{T}$και $\boldsymbol{D}$ μόνο τα αντίστοιχα ιδιοδιανύσματα, δημιουργώντας τους $\boldsymbol{T_k}$ και $\boldsymbol{D_k}$, τότε συνθέτουμε ένα νέο πίνακα $\boldsymbol{A_k}$, ο οποίος προέρχεται από τον **A** εκφρασμένο σε διαφορετικό σύστημα συντεταγμένων. Οι στήλες του πίνακα $\boldsymbol{A_k}$ θα αντιστοιχούν πλέον σε «έννοιες» (concepts), αντί για τις αρχικές (και πολύ περισσότερες) ενότητες, ενώ κάθε γραμμή θα δείχνει το συσχετισμό της αντίστοιχης λέξης με κάθε έννοια. Μπορούμε πλέον να βρούμε τη «σημασιολογική» ομοιότητα μεταξύ δύο λέξεων $\mathbf{w_i}$, $\mathbf{w_j}$ υπολογίζοντας την ομοιότητα συνημιτόνου των αντιστοίχων διανυσμάτων-γραμμών του $\boldsymbol{A_k}$**.**

[7] Ονομάζουμε τον πίνακα διαγώνιο καταχρηστικά διότι δεν είναι τετραγωνικός, αλλά οι τιμές του $d_{i,j}$ είναι μη μηδενικές μόνο όταν i = j

$$LSA(\mathrm{w_i}, \mathrm{w_j}) = \frac{\overrightarrow{\mathrm{w_i}} \times \overrightarrow{\mathrm{w_j}}}{|\overrightarrow{\mathrm{w_i}}| \cdot |\overrightarrow{\mathrm{w_j}}|}$$

Όπου:

$\overrightarrow{\boldsymbol{w_i}}$**:** το διάνυσμα-γραμμή της λέξης $\mathrm{w_i}$ στον πίνακα A_k,

$\overrightarrow{\boldsymbol{w_2}}$**:** το διάνυσμα-γραμμή της λέξης $\mathrm{w_2}$ στον πίνακα A_k,

$|\overrightarrow{\boldsymbol{w_i}}|$**:** το μέτρο του $\overrightarrow{\mathrm{w_i}}$,

$|\overrightarrow{\boldsymbol{w_2}}|$**:** το μέτρο του $\overrightarrow{\mathrm{w_2}}$.

Επιστρέφοντας στη μέθοδο της παρούσας εργασίας, τα γνωρίσματα που βασίζονται στη ΑΛΣ είναι τέσσερα.[8] Για κάθε κανόνα *r* που εφαρμόζεται στο ζεύγος αρχικής-τελικής πρότασης που εξετάζουμε, υπολογίζουμε το:

$$LSA(r) = \frac{\sum_{\mathrm{w_i}, \mathrm{w_j} \in K} LSA(\mathrm{w_i}, \mathrm{w_j})}{|K|}$$

όπου *K* το σύνολο όλων των ζευγών λέξεων w_1, w_2 ώστε η w_1 να προέρχεται από το αριστερό μέρος (LHS) του κανόνα και η w_2 από το δεξιό (RHS). Για παράδειγμα έστω ότι έχουμε τον κανόνα:

JJ_1 ***teams*** ⟺ *JJ_1* ***task forces***

Τότε:

K ={(teams, task),(teams, forces)}

Θεωρώντας ότι στο ζεύγος αρχικής-τελικής πρότασης εφαρμόζονται *k* κανόνες παράφρασης, υπολογίζουμε τα ακόλουθα τρία γνωρίσματα:

i. $\boldsymbol{LSA} = \frac{1}{k}\sum_{i=1}^{k} LSA(r_i)$

ii. $\boldsymbol{LSAmax} = max_{i\in\{1,\dots,k\}} LSA(r_i)$

iii. $\boldsymbol{LSAmin} = min_{i\in\{1,\dots,k\}} LSA(r_i)$

Για το τέταρτο γνώρισμα που βασίζεται στην ΑΛΣ, μετράμε την ομοιότητα (με βάση την ΑΛΣ) ανάμεσα στα διανύσματα της αρχικής πρότασης και της παραχθείσας. Το διάνυσμα της κάθε πρότασης είναι ο βεβαρημένος μέσος όρος των διανυσμάτων των λέξεών της (27). Τα βάρη κάθε λέξης – όρου, δίνονται με έναν λογαριθμικό μετασχηματισμό της εντροπίας της

[8]Χρησιμοποιήθηκε η υλοποίηση LSAπου διατίθεται από τη διεύθυνση: http://lsa.colorado.edu/

κάθε λέξης. Περισσότερες πληροφορίες σχετικά με αυτόν τον μετασχηματισμό μπορούν να αναζητηθούν στο (27).

iv. $\boldsymbol{GLSA} = LSA(\alpha\rho\chi\iota\kappa\acute{\eta}\ \pi\rho\acute{o}\tau\alpha\sigma\eta, \tau\epsilon\lambda\iota\kappa\acute{\eta}\ \pi\rho\acute{o}\tau\alpha\sigma\eta)$

Κεφάλαιο 5: Πειραματικά δεδομένα και αποτελέσματα

5.1 Περισσότερες πληροφορίες για τα πειραματικά δεδομένα

Τα δεδομένα εκπαίδευσης και αξιολόγησης, όπως περιγράφηκαν και στην ενότητα 4.1, αποτελούνται από ζεύγη προτάσεων. Κάθε ζεύγος περιέχει μια αρχική πρόταση και μια τελική, που έχει δημιουργηθεί εφαρμόζοντας έναν ή περισσότερους κανόνες παράφρασης στην αρχική. Η ίδια αρχική πρόταση ενδέχεται να συμμετέχει σε πολλά ζεύγη, αφού εφαρμόζοντας διαφορετικούς κανόνες παράφρασης προκύπτουν διαφορετικές τελικές. Οι κανόνες παράφρασης προέρχονται από τη συλλογή κανόνων που κατασκεύασαν οι Zhao κ.ά. (10) με την μέθοδο της ενότητας 1.3. Στα δεδομένα συμπεριλαμβάνονται και πληροφορίες που δείχνουν, για κάθε ζεύγος, ποιοι κανόνες εφαρμόστηκαν, καθώς και τους βαθμούς των κανόνων από τα μοντέλα 1 και 3 των Zhao κ.ά. (βλ. ενότητα 2.3).

Τα ζεύγη προτάσεων δόθηκαν σε ανθρώπους-κριτές, από τους οποίους ζητήθηκε να βαθμολογήσουν κάθε ζεύγος με τρεις βαθμούς, ως προς τη γραμματική ορθότητα (grammaticality) της τελικής πρότασης, τη διατήρηση του νοήματος της αρχικής (meaning preservation) και τη συνολική ποιότητα (overall quality) της παράφρασης. Η αξιολόγηση από τους ανθρώπους-κριτές έγινε όπως περιγράφεται από τους Μαλακασιώτη κ.ά. (3) και περιγράφεται παρακάτω συνοπτικά.

Οι κριτές ήταν 15 και κάθε κριτής εξέτασε κατά μέσο όρο 148 ζεύγη. Οι διαθέσιμες τιμές των παραπάνω βαθμολογιών είναι 1, 2, 3 και 4. Όσον αφορά την γραμματική ορθότητα, η τιμή 1 σημαίνει ότι η παράφραση εμφανίζει πολλά γραμματικά λάθη σε σχέση με την αρχική, και η τιμή 4 ότι είναι απόλυτα γραμματικά ορθή. Αντίστοιχα, για την διατήρηση του νοήματος, 1 σημαίνει πως καμία από τις πληροφορίες που ανέφερε η αρχική πρόταση δεν αναφέρεται στην παραχθείσα ή ότι οι πληροφορίες αυτές έχουν διαστρεβλωθεί, ενώ η τιμή 4 σημαίνει ότι η παραχθείσα πρόταση περιλαμβάνει όλες τις πληροφορίες που είχε και η αρχική. Στην περίπτωση της συνολικής ποιότητας της παράφρασης, οι τιμές 1 - 2 δείχνουν ότι η πρόταση που προέκυψε δεν είναι καλή παράφραση και οι τιμές 3 - 4 δείχνουν πως είναι.

5.2 Πειράματα με γραμμικό διαχωριστή

Στα πρώτα πειράματα που εκτελέσαμε, χωρίσαμε τα ζεύγη προτάσεων σε δύο κατηγορίες: μια θετική κατηγορία που περιελάμβανε τα ζεύγη με βαθμό συνολικής ποιότητας 3 ή 4 και μια αρνητική με τα υπόλοιπα ζεύγη, δηλαδή με βαθμό συνολικής ποιότητας 1 ή 2. Ο στόχος ήταν να εκπαιδεύσουμε έναν ταξινομητή που να μπορεί να διαχωρίζει τα ζεύγη των δύο κατηγοριών. Η εκπαίδευση του ταξινομητή έγινε με 1435 ζεύγη και η αξιολόγηση με 1909 ζεύγη διαφορετικά από εκείνα της εκπαίδευσης. Στα πειράματα χρησιμοποιήθηκαν διάφοροι συνδυασμοί των γνωρισμάτων και μετρήθηκε το ποσοστό λάθους του ταξινομητή (error rate). Δοκιμάσαμε έναν ταξινομητή μεγίστης εντροπίας (ενότητα 4.2), δηλαδή μια περίπτωση γραμμικού διαχωριστή, καθώς και ένα βασικό (baseline) σύστημα σύγκρισης, το οποίο περιγράφεται στη συνέχεια.

5.2.1 Σύστημα σύγκρισης

Το βασικό σύστημα (baseline) αποφασίζει «αφελώς» αν ένα ζεύγος προτάσεων ανήκει στη θετική ή την αρνητική κατηγορία, χρησιμοποιώντας τους τέσσερις βαθμούς που συνοδεύουν κάθε κανόνα παράφρασης των Zhao κ.ά. (βλ. ενότητα 1.3). Πιο συγκεκριμένα, για κάθε κανόνα που έχει χρησιμοποιηθεί στο εξεταζόμενο ζεύγος, υπολογίζεται ο μέσος όρος των τεσσάρων βαθμών του και κατόπιν ο μέσος όρος των βαθμών όλων των εμπλεκόμενων κανόνων. Αν ο συνολικός μέσος όρος είναι μεγαλύτερος από ένα κατώφλι, το ζεύγος κατατάσσεται στη θετική κατηγορία, διαφορετικά στην αρνητική. Το κατώφλι επιλέγεται ώστε να επιτυγχάνει μεγάλο ποσοστό ορθότητας (accuracy) στα δεδομένα εκπαίδευσης. Το ποσοστό ορθότητας ορίζεται ως:

$$ορθότητα = \frac{TN + TP}{TN + TP + FN + FP}$$

όπου:

TN = #αρνητικών παραδειγμάτων που κατατάχτηκαν σωστά (true negatives),

TP = #θετικών παραδειγμάτων που κατατάχτηκαν σωστά (true positives),

FN = #θετικών παραδειγμάτων που κατατάχτηκαν λανθασμένα ως αρνητικά (false negatives),

FP = #αρνητικών παραδειγμάτων που κατατάχτηκαν λανθασμένα ως θετικά (false positives).

Αναζητήσαμε την καλύτερη τιμή του κατωφλίου στο διάστημα [0,1], ξεκινώντας από την τιμή 0 και αυξάνοντας την κατά 0,001 σε κάθε επανάληψη. Η τιμή που οδήγησε στο καλύτερο ποσοστό ορθότητας (0,5407) στα δεδομένα εκπαίδευσης και που χρησιμοποιήθηκε στη συνέχεια ήταν 0,137.

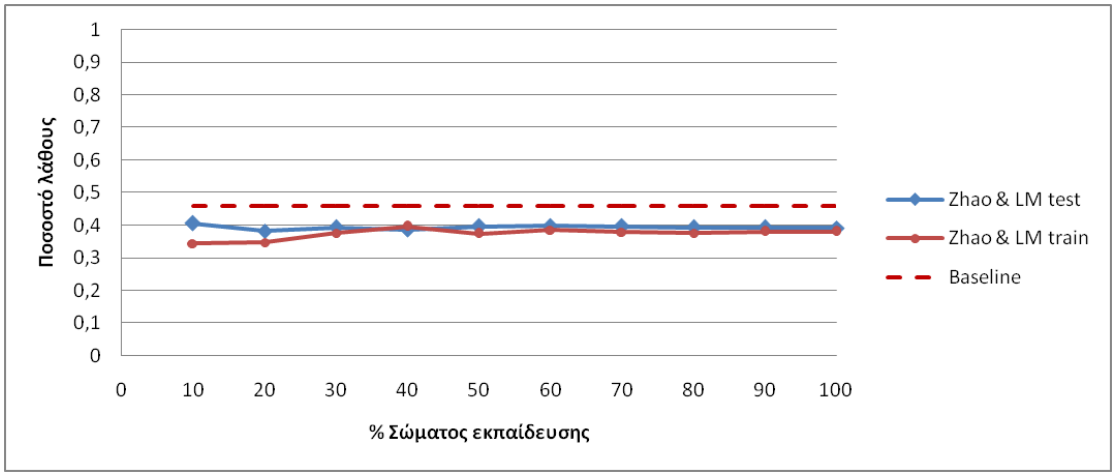
5.2.2 Αποτελέσματα ταξινομητή μεγίστης εντροπίας

Με τον ταξινομητή μεγίστης εντροπίας, δοκιμάσαμε διάφορους συνδυασμούς των γνωρισμάτων της ενότητας 4.3.

Αρχικά χρησιμοποιήσαμε ως γνωρίσματα τους βαθμούς που συνοδεύουν τους κανόνες των Zhao κ.α. (10), τους οποίους χρησιμοποιεί και το σύστημα σύγκρισης, καθώς και τα γνωρίσματα που βασίζονται στο γλωσσικό μοντέλο (βλ. ενότητα 4.3.1). Στο παρακάτω διάγραμμα απεικονίζονται οι καμπύλες μάθησης για αυτό το σύστημα (Zhao & LM), καθώς και τα αντίστοιχα αποτελέσματα του συστήματος σύγκρισης. Ο οριζόντιος άξονας δείχνει το ποσοστό του συνόλου των δεδομένων εκπαίδευσης που χρησιμοποιείται, ενώ ο κατακόρυφος άξονας δείχνει το ποσοστό λάθους.

Η καμπύλη «Zhao&LMtest» μετράει το ποσοστό λάθους του ταξινομητή όταν του ζητείται να κατατάξει τα δεδομένα αξιολόγησης, δηλαδή ζεύγη προτάσεων που δεν έχει συναντήσει κατά την εκπαίδευση. Καθώς προστίθενται δεδομένα εκπαίδευσης, αναμένεται αυτή η καμπύλη να κατεβαίνει, δηλαδή να μειώνεται το σφάλμα. Η καμπύλη «Zhao&LMtrain» μετράει το ποσοστό λάθους του ταξινομητή όταν του ζητείται να κατατάξει τα ίδια δεδομένα με τα οποία εκπαιδεύτηκε. Δεδομένου ότι συνήθως ένας ταξινομητής τα πηγαίνει καλύτερα στα δεδομένα στα οποία έχει εκπαιδευτεί και χειρότερα όταν του ζητείται να κατατάξει νέα δεδομένα, διαισθητικά η δεύτερη καμπύλη αποτελεί ένα είδος κάτω φράγματος για την πρώτη. Καθώς προστίθενται νέα δεδομένα εκπαίδευσης, αναμένεται η δεύτερη καμπύλη να ανεβαίνει (χειρότερα αποτελέσματα στα δεδομένα εκπαίδευσης), γιατί γίνεται δυσκολότερο το σύστημα να κάνει υπερ-εφαρμογή (overfitting) στα δεδομένα εκπαίδευσης.

Η εκπαίδευση του συστήματος σύγκρισης έγινε σε όλα τα δεδομένα εκπαίδευσης, αλλά δείχνουμε τα αποτελέσματά του ως ευθεία για ευκολία σύγκρισης.

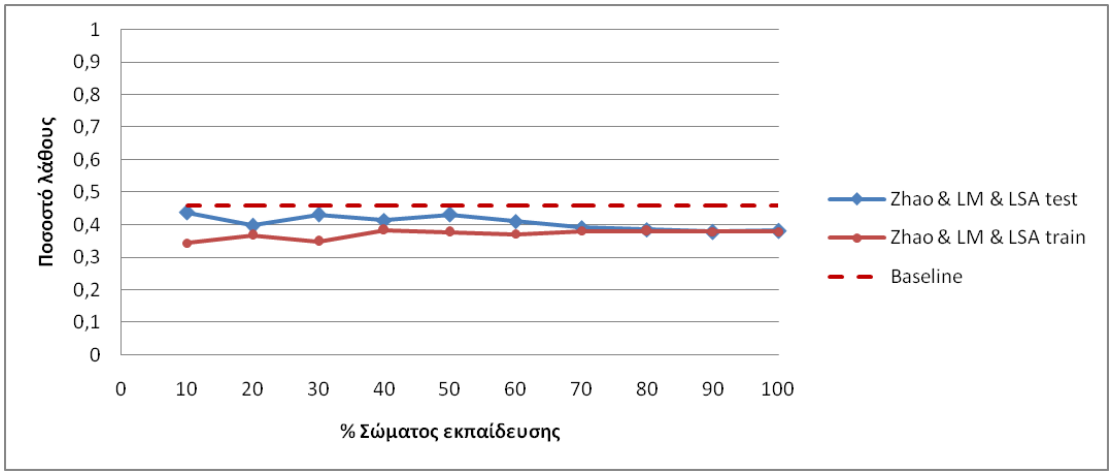


Διάγραμμα 1:Καμπύλες μάθησης συστήματος Zhao&LM

Πρέπει να τονίσουμε ότι, αφού μετράμε το ποσοστό λάθους του κάθε συστήματος, όσο χαμηλότερη είναι μια καμπύλη τόσο καλύτερο είναι το αντίστοιχο σύστημα. Από το διάγραμμα 1, συμπεραίνουμε ότι ο ταξινομητής μεγίστης εντροπίας με γνωρίσματα μόνο τους βαθμούς των κανόνων και τα γνωρίσματα γλωσσικού μοντέλου είναι ελαφρώς καλύτερος από το σύστημα σύγκρισης (baseline).

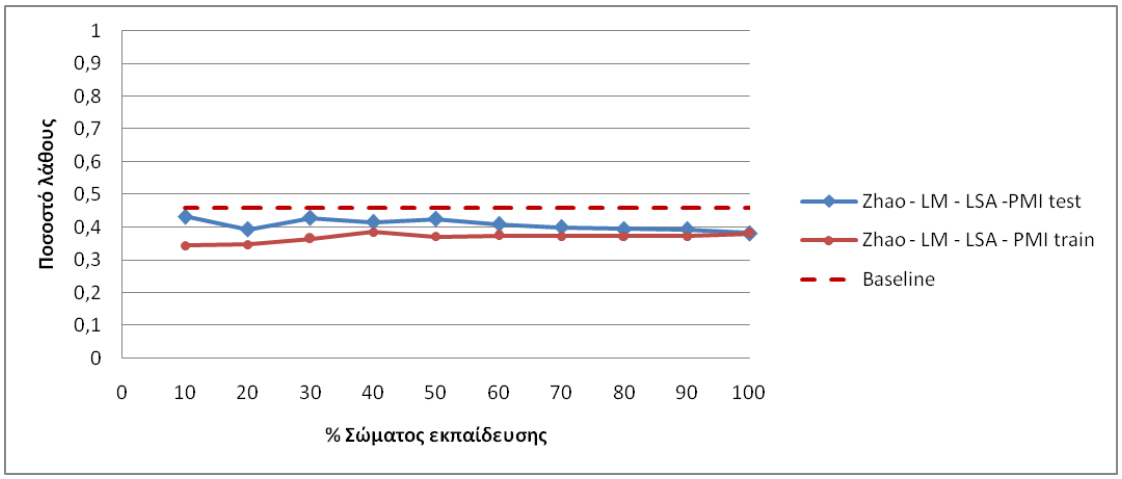
Η μικρή απόσταση μεταξύ των δύο καμπυλών μάθησης test και train μας δείχνει ότι δεν υπάρχουν περιθώρια να βελτιωθεί ο ταξινομητής προσθέτοντας περισσότερα δεδομένα εκπαίδευσης και ότι αντιμετωπίζουμε πρόβλημα περιορισμένου χώρου αναζήτησης (high bias problem). Στην περίπτωση αυτή, για αν βελτιωθεί η επίδοση του συστήματος συνιστάται συνήθως να προσθέσουμε περισσότερα γνωρίσματα στο ταξινομητή ή να δοκιμάσουμε μια πιο περίπλοκη υπόθεση (π.χ. μη γραμμικό διαχωριστή). Αποφασίσαμε να διερευνήσουμε πρώτα αν τα αποτελέσματα βελτιώνονται με περισσότερα γνωρίσματα.

Προσθέσαμε τα γνωρίσματα που βασίζονται στην ΑΛΣ, που περιγράφηκαν στην ενότητα 4.3.3. Στα πειράματα μας κρατήσαμε τις k = 300 ιδιάζουσες τιμές από τον πίνακα της ΑΛΣ. Ειδικά για τον υπολογισμό του γνωρίσματος GLSA (βλ. ενότητα 4.3.3), οι ιδιάζουσες τιμές που κρατήσαμε ήταν k = 249. Στο διάγραμμα παρακάτω φαίνονται οι αντίστοιχες καμπύλες μάθησης του νέου συστήματος (Zhao& LM & LSA).



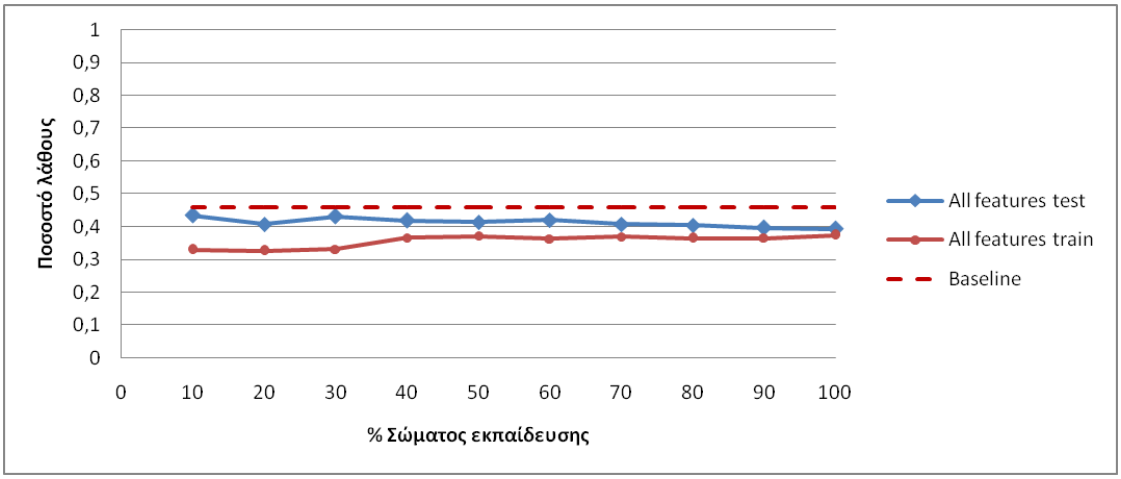
Διάγραμμα 2:Καμπύλες μάθησης συστήματος Zhao&LM&LSA

Συγκρίνοντας αυτό το διάγραμμα με το προηγούμενο διαπιστώνουμε ότι δεν υπάρχουν ουσιαστικές διαφορές. Το πρόβλημα περιορισμένου χώρου αναζήτησης (high bias) παραμένει και το ποσοστό λάθους μειώθηκε ελάχιστα. Η μικρή όμως βελτίωση μας ώθησε να προσθέσουμε και τα γνωρίσματα ΣΑΠ (ενότητα 2.5.2) στο ταξινομητή. Τα αποτελέσματα φαίνονται στο διάγραμμα 3.



Διάγραμμα 3:Καμπύλες μάθησης συστήματος Zhao&LM&LSA&PMI

Δυστυχώς και σε αυτή την περίπτωση δεν υπάρχει ουσιαστική βελτίωση και το πρόβλημα παραμένει. Ακόμα και όταν χρησιμοποιούμε όλα τα γνωρίσματα που αναλύσαμε στις ενότητες 2.5.1 έως 2.5.3, τα αποτελέσματα δεν βελτιώνονται ουσιαστικά, με το ποσοστό λάθους να παραμένει στα ίδια επίπεδα (βλ. διάγραμμα 4).



Διάγραμμα 4: Καμπύλες μάθησης συστήματος Allfeatures

Τα αποτελέσματα που παρουσιάστηκαν στα παραπάνω διαγράμματα συνοψίζονται στον Πίνακα 1, ως ποσοστά ορθότητας στα δεδομένα αξιολόγησης.

Συστήματα | Ποσοστό δεδομένων εκπαίδευσης	**10**	**20**	**30**	**40**	**50**
Zhao & LM test	0,5945521	0,618649	0,607124	0,61341	0,605029
Zhao & LM & LSA test	0,5625982	0,603457	0,568884	0,586695	0,569408
Zhao & LM & LSA & PMI test	0,5683604	0,608172	0,572551	0,585123	0,576218
All features test	0,5662651	0,592457	0,569408	0,582504	0,586171
Συστήματα | Ποσοστό δεδομένων εκπαίδευσης	**60**	**70**	**80**	**90**	**100**
Zhao & LM test	0,6034573	0,604505	0,607648	0,608172	0,609219
Zhao & LM & LSA test	0,5903614	0,608696	0,615506	0,622315	0,619172
Zhao & LM & LSA & PMI test	0,5914091	0,601886	0,6066	0,607124	0,619172
All features test	0,5809324	0,593504	0,596124	0,603981	0,607648

Πίνακας 1: Ποσοστά ορθότητας συστημάτων

5.3 Πειράματα με παλινδρόμηση διανυσμάτων υποστήριξης

Εφόσον η προσθήκη περισσοτέρων γνωρισμάτων στον ταξινομητή δεν κατάφεραν να λύσουν το πρόβλημα περιορισμένου χώρου αναζήτησης(high bias), αποφασίσαμε να δοκιμάσουμε πιο περίπλοκες υποθέσεις. Χρησιμοποιήσαμε παλινδρόμηση διανυσμάτων υποστήριξης (ΠΔΥ, Support Vector Regression–SVR, βλ. και ενότητα 4.2) με πυρήνα συνάρτησης ακτινικής βάσης (Radial Basis Function, RBF) (29) για την αξιολόγηση των παραφράσεων, που επιτρέπει τη μάθηση και μη γραμμικών υποθέσεων. Επίσης, πλέον δεν

κάνουμε δυαδική κατάταξη των ζευγών σε ορθές και μη παραφράσεις, αλλά χρησιμοποιούμε την ΠΔΥ για να βαθμολογήσουμε κάθε ζεύγος με έναν πραγματικό αριθμό, ο οποίος (ιδανικά) δείχνει πόσο καλή είναι η παράφραση.

Η ΠΔΥ εκπαιδεύτηκε στα ίδια 1435 ζεύγη προτάσεων, στα οποία εκπαιδεύτηκε και ο ταξινομητής μέγιστης εντροπίας. Η αξιολόγηση έγινε στα 1909 ζεύγη που αναφέραμε και στην αρχή της ενότητας 4.1. Πριν τη διαδικασία της εκπαίδευσης χρειάστηκε να ρυθμίσουμε τις παραμέτρους της ΠΔΥ (και του πυρήνα RBF), ώστε να επιτύχουμε καλύτερα αποτελέσματα. Η ρύθμιση έγινε με το κριτήριο της τετραγωνικής συσχέτισης (squared correlation coefficient - r^2) στις βαθμολογίες που δίνει η ΠΔΥ με αυτές των ανθρώπων στα δεδομένα εκπαίδευσης. Η συσχέτιση αυτή υπολογίζεται ως εξής:

$$r^2 = \frac{(\sum_{i=1}^{l} f(x_i) y_i - \sum_{i=1}^{l} f(x_i) \sum_{i=1}^{l} y_i)^2}{(l \sum_{i=1}^{l} f(x_i)^2 - (\sum_{i=1}^{l} f(x_i))^2)(l \sum_{i=1}^{l} y_i^2 - (\sum_{i=1}^{l} y_i)^2)}$$

Όπου:
$\boldsymbol{x_i}$**:** το διάνυσμα γνωρισμάτων του ζεύγους προτάσεων,
$\boldsymbol{y_i}$**:** η ανθρώπινη βαθμολογία για το διάνυσμα,
$\boldsymbol{f(x_i)}$**:** η πρόβλεψη βαθμολογίας που δίνει η ΠΔΥ,
$\boldsymbol{l}$**:** το μέγεθος του συνόλου δεδομένων.

Μια ακόμη αλλαγή, σε σχέση με τα προηγούμενα πειράματα, είναι ότι δεν χρησιμοποιούμε πλέον το βαθμό συνολικής ποιότητας (overall quality) που έχουν δώσει οι άνθρωποι-κριτές σε κάθε ζεύγος προτάσεων, επειδή οι Μαλακασιώτης κ.ά. (3) παρατήρησαν ότι οι άνθρωποι-κριτές συχνά δυσκολεύονταν να αποφασίσουν κατά πόσον ο βαθμός συνολικής ποιότητας θα έπρεπε να αντανακλά περισσότερο τη γραμματική ορθότητα, τη διατήρηση του νοήματος ή το πόσο μεγάλες είναι οι αλλαγές από την αρχική στην τελική πρόταση (π.χ. αν άλλαξε μόνο μία ή πολλές λέξεις). Ακολουθώντας τους Μαλακασιώτη κ.ά. θεωρούμε πλέον ότι ο σωστός βαθμός για κάθε ζεύγος προτάσεων, δηλαδή αυτός που ιδανικά θα έπρεπε να μάθει να επιστρέφει η ΠΔΥ είναι:

$$ideal\ score = \lambda_1 \cdot Grammaticality + \lambda_2 \cdot Meaning + \lambda_3 \cdot Diversity$$

Όπου:

***Grammaticality*:** ο βαθμός γραμματικής ορθότητας (1 – 4) που έχουν δώσει οι άνθρωποι-κριτές στο ζεύγος,
***Meaning*:** ο βαθμός διατήρησης νοήματος (1 – 4) που έχουν δώσει οι άνθρωποι-κριτές στο ζεύγος,
***Diversity*:** η απόσταση επεξεργασίας (1 – edit distance) μεταξύ της ακολουθίας λέξεων της αρχικής και τελικής πρότασης του ζεύγους, η οποία υπολογίζεται αυτόματα,
$\boldsymbol{\lambda_1, \lambda_2, \lambda_3}$**:** θετικά (ή μηδενικά) βάρη με $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Ο τρίτος παράγοντας (Diversity) μετράει κατά πόσο η τελική πρόταση του ζεύγους διαφέρει από την αρχική. Για παράδειγμα, αν έχουμε την αρχική πρόταση:

*He **said** he could not walk **at all**.*

θα θέλαμε να θεωρηθεί καλύτερο ένα ζεύγος με τελική πρόταση την:

*He **announced that** he could not walk **in the least**.*

από ένα ζεύγος με τελική πρόταση την:

*He **said that** he could not walk **at all**.*

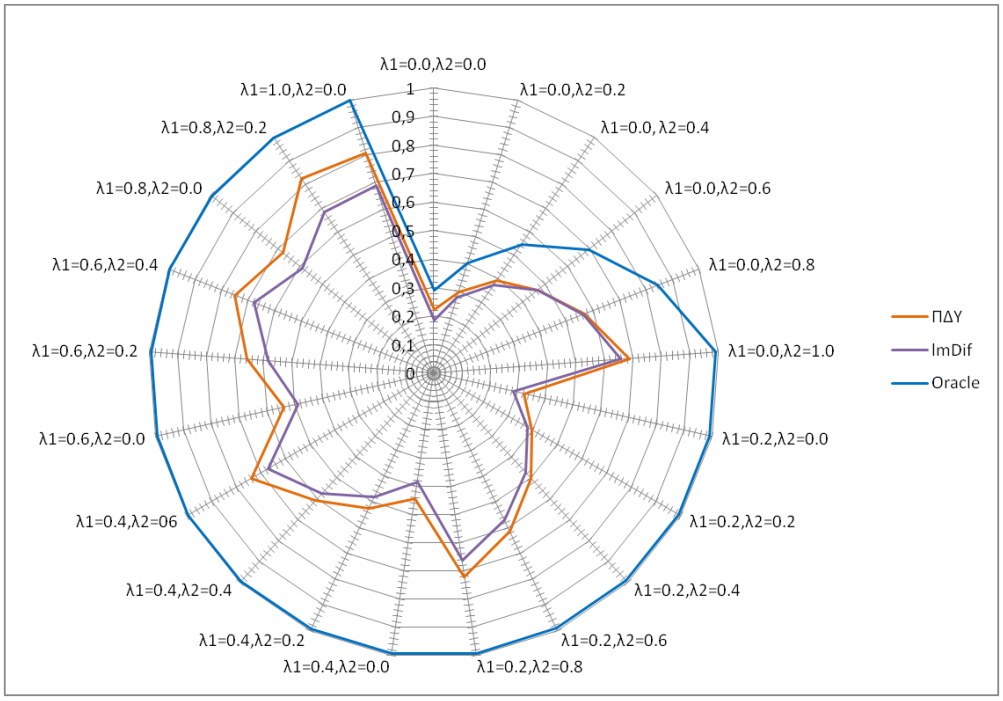
Η διαφοροποίηση (Diversity) δεν περιλαμβάνεται στους βαθμούς που έδωσαν οι άνθρωποι - κριτές αλλά υπολογίζεται αυτόματα ως το 1 μείον την απόσταση επεξεργασίας (1-edit distance) και είναι κανονικοποιημένη στο διάστημα [0, 1]. Αντίστοιχα και οι υπόλοιποι συντελεστές(Grammaticality, Meaning) κανονικοποιήθηκαν στο ίδιο διάστημα.

Τα βάρη $\lambda_1, \lambda_2, \lambda_3$ δεν παίρνουν προκαθορισμένες τιμές με κάποια διαδικασία εκπαίδευσης. Θεωρούμε ότι το βάρος κάθε συντελεστή πρέπει να είναι διαφορετικό ανάλογα με την εφαρμογή στην οποία θέλουμε να χρησιμοποιήσουμε τις παραφράσεις. Για παράδειγμα, σε ένα σύστημα ανάκτησης πληροφοριών ενδέχεται να μας ενδιαφέρει να παράγουμε όσο περισσότερες νέες λέξεις-κλειδιά μπορούμε κατά την παράφραση του ερωτήματος, διατηρώντας το νόημα, με αποτέλεσμα η διατήρηση νοήματος (Meaning) και η διαφοροποίηση (Diversity) να είναι ενδεχομένως πιο σημαντικές από τη γραμματική ορθότητα (Grammaticality). Αντιθέτως, σε ένα σύστημα παραγωγής κειμένων φυσικής

γλώσσας πρέπει να δοθεί περισσότερο βάρος στην γραμματική ποιότητα της παράφρασης. Θέλουμε, επομένως, να αξιολογούμε τις μεθόδους παράφρασης για πολλούς διαφορετικούς συνδυασμούς $\lambda_1, \lambda_2, \lambda_3$. Αυτή η διαδικασία αξιολόγησης εισήχθη από τους Μαλακασιώτη κ.ά. (3).

Πιο συγκεκριμένα, για κάθε (αρχική) πρόταση που μας δίνουν προς παράφραση, χρησιμοποιούμε την ΠΔΥ για να αξιολογήσουμε όλα τα ζεύγη αρχικής-τελικής πρότασης στα οποία συμμετέχει η δοθείσα αρχική πρόταση. Κρατάμε το ζεύγος που η ΠΔΥ αξιολόγησε ως καλύτερο και επιστρέφουμε την τελική πρόταση που αυτό περιέχει. Για να εξετάσουμε πόσο καλά τα πηγαίνει η ΠΔΥ, υπολογίζουμε (για κάθε αρχική πρόταση) τον ιδανικό βαθμό (ideal score) του ζεύγους που η ΠΔΥ θεώρησε καλύτερο και τον συγκρίνουμε με τον ιδανικό βαθμό του ζεύγους που θα επέλεγε ένα βασικό σύστημα σύγκρισης (baseline) ή ένα μαντείο (oracle). Το βασικό σύστημα επιλέγει το ζεύγος με τη μικρότερη διαφορά πιθανότητας γλωσσικού μοντέλου (lmDif, βλ. ενότητα 4.3.1), ενώ το μαντείο έχει στη διάθεσή του και τους ιδανικούς βαθμούς όλων των ζευγών και διαλέγει το ζεύγος με τον υψηλότερο ιδανικό βαθμό. Η διαδικασία αυτή επαναλαμβάνεται για πολλούς διαφορετικούς συνδυασμούς $\lambda_1, \lambda_2, \lambda_3$.

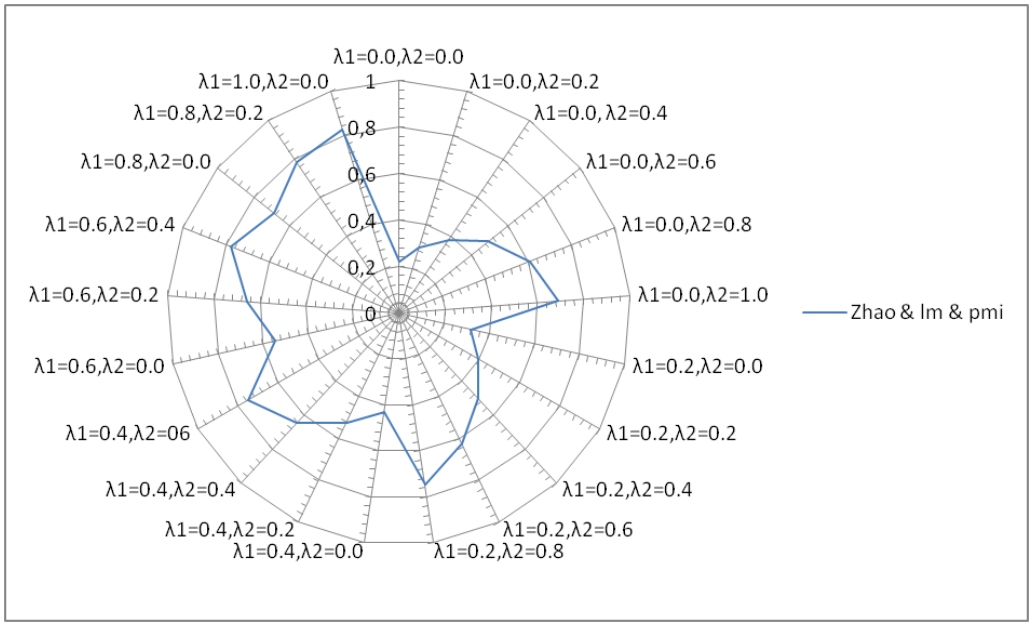
Δοκιμάσαμε 21 διαφορετικούς συνδυασμούς των λ_i, όπως φαίνεται στο Διάγραμμα 5. Κάθε ακτίνα του διαγράμματος αντιστοιχεί σε διαφορετικό συνδυασμό των τιμών των λ_i. Σε κάθε ακτίνα, η απόσταση από το κέντρο του διαγράμματος είναι ο ιδανικός βαθμός (ideal score) του ζεύγους που επέλεξε η αντίστοιχη μέθοδος. Το διάγραμμα δείχνει τις καμπύλες του βασικού συστήματος, του μαντείου και της ΠΔΥ. Αν η καμπύλη μιας μεθόδου περικλείει ολόκληρη την καμπύλη μιας άλλης, τότε η πρώτη μέθοδος είναι καλύτερη ανεξαρτήτως των τιμών των $\lambda_1, \lambda_2, \lambda_3$. Βλέπουμε από το διάγραμμα ότι η ΠΔΥ είναι σαφώς καλύτερη από τη βασική μέθοδο (κάτω φράγμα) και χειρότερη από τη μέθοδο του μαντείου (άνω φράγμα).



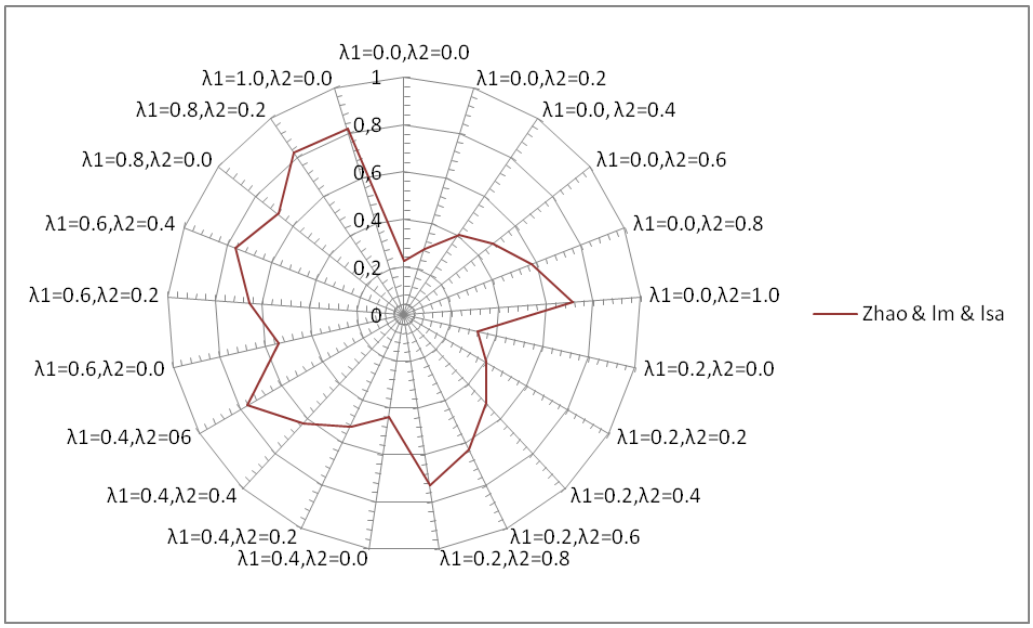
Διάγραμμα 5: Μέση Ανθρώπινη Βαθμολογία

Επιπλέον, παρατηρούμε ότι υπάρχουν κάποιες περιοχές στο διάγραμμα, κυρίως για τους συνδυασμούς όπου το λ_3 είναι πολύ μεγάλο σε σχέση με τα λ_1, λ_2, όπου ακόμα και η μέθοδος του μαντείου δεν έχει τέλεια αποτελέσματα. Αυτό είναι αναμενόμενο, αφού αν δώσουμε βάρος μόνο στην διαφοροποίηση της πρότασης (λ_3 να τείνει στο 1), χωρίς να ενδιαφερόμαστε για την διατήρηση της γραμματικής ποιότητας ή του νοήματος της αρχικής πρότασης (λ_1 και λ_2 να τείνουν στο 0), δύσκολα θα καταλήξουμε σε «καλές» παραφράσεις, δεδομένου ότι οι κανόνες παράφρασης επιφέρουν γενικά μικρές αλλαγές στην αρχική πρόταση.

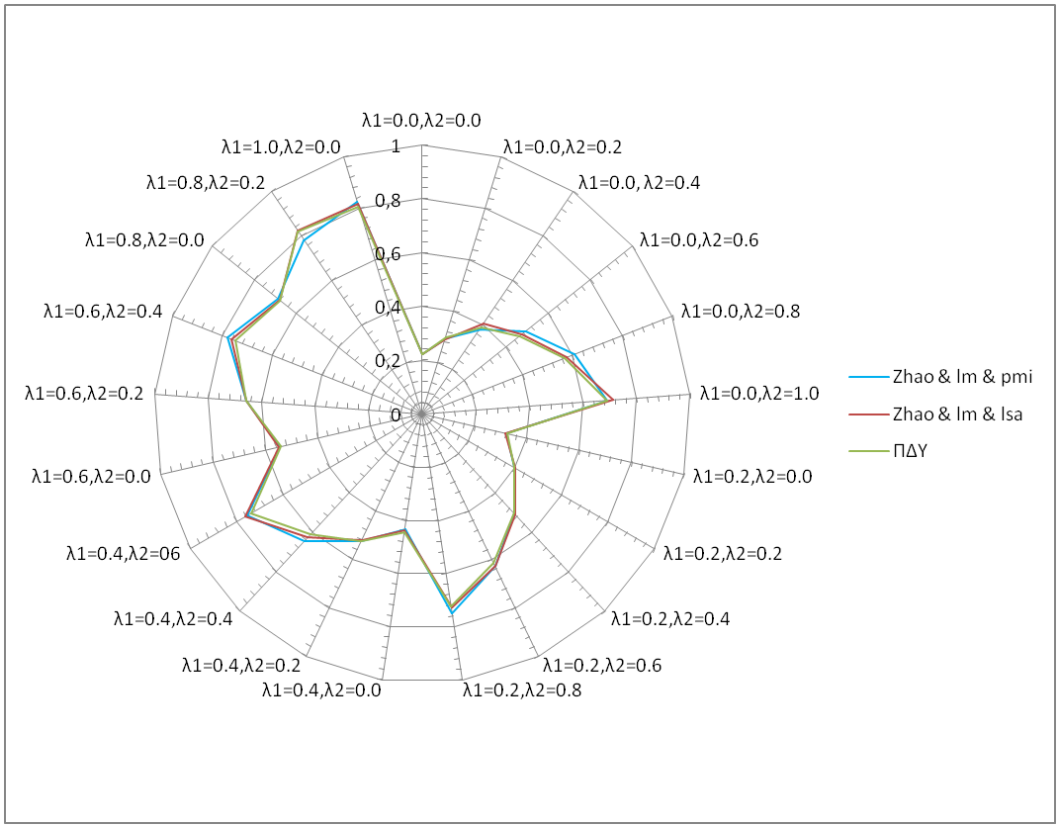
Στην περίπτωση της ΠΔΥ, τα γνωρίσματα είναι ίδια με αυτά που χρησιμοποιήθηκαν με τον ταξινομητή μέγιστης εντροπίας. Συνοπτικά χρησιμοποιούμε τα μέτρα που προκύπτουν από το γλωσσικό μοντέλο, τη σημειακή αμοιβαία πληροφορία και την ανάλυση λανθάνουσας σημασίας (LM, PMI, LSA). Συνολικά τα γνωρίσματα είναι 31. Επιπλέον έγιναν πειράματα με διαφορετικούς συνδυασμούς γνωρισμάτων, αλλά και με λιγότερα δεδομένα εκπαίδευσης, ωστόσο τα αποτελέσματα δεν έδειξαν έντονη διαφοροποίηση στην καμπύλη, όπως φαίνεται στα Διαγράμματα 5 – 7.



Διάγραμμα 6: Πειράματα με χρήση μόνο γλωσσικού μοντέλου και ΣΑΠ



Διάγραμμα 7: Πειράματα με χρήση γλωσσικού μοντέλου και ΑΛΣ



Διάγραμμα 8: Συγκεντρωτικές καμπύλες

Στον Πίνακα 2 φαίνονται αναλυτικά και οι ιδανικοί βαθμοί των επιλεγόμενων ζευγών, που παριστάνονται στο Διάγραμμα 5.

λ	ΠΔΥ	ImDif	Oracle
λ1=0.0,λ2=0.0	0,22178854	0,185117206	0,291715829
λ1=0.0,λ2=0.2	0,298498903	0,279649321	0,404496644
λ1=0.0, λ2=0.4	0,391555805	0,374181435	0,547245333
λ1=0.0,λ2=0.6	0,466493487	0,468713549	0,695155008
λ1=0.0,λ2=0.8	0,568533522	0,563245663	0,84313306
λ1=0.0,λ2=1.0	0,688888889	0,657777778	0,991111111
λ1=0.2,λ2=0.0	0,324994022	0,285871543	0,991111111
λ1=0.2,λ2=0.2	0,395495284	0,380403657	0,991111111
λ1=0.2,λ2=0.4	0,5010237	0,474935771	0,991111111
λ1=0.2,λ2=0.6	0,614172355	0,569467886	0,991111111
λ1=0.2,λ2=0.8	0,722666667	0,664	0,992888889
λ1=0.4,λ2=0.0	0,444767587	0,386625879	0,992888889
λ1=0.4,λ2=0.2	0,52461985	0,481157994	0,992888889
λ1=0.4,λ2=0.4	0,609666009	0,575690108	0,992888889
λ1=0.4,λ2=06	0,737777778	0,670222222	0,994666667
λ1=0.6,λ2=0.0	0,538823541	0,487380216	0,994666667
λ1=0.6,λ2=0.2	0,657716556	0,58191233	0,994666667
λ1=0.6,λ2=0.4	0,748444444	0,676444444	0,996444444
λ1=0.8,λ2=0.0	0,677218295	0,588134552	0,996444444
λ1=0.8,λ2=0.2	0,824	0,682666667	0,998222222
λ1=1.0,λ2=0.0	0,808888889	0,688888889	1

Πίνακας 2: Συγκεντρωτικά αποτελέσματα πειραμάτων ΠΔΥ (Όλα τα γνωρίσματα)

Κεφάλαιο 6: Συμπεράσματα και μελλοντική εργασία

Στη διάρκεια αυτής της εργασίας, μελετήθηκαν αρχικά μέθοδοι αυτόματης εξαγωγής και αξιολόγησης κανόνων παράφρασης που έχουν προταθεί στη βιβλιογραφία. Στη συνέχεια αναπτύχθηκε μια νέα μέθοδος αξιολόγησης κανόνων παράφρασης, η οποία χρησιμοποιεί επιβλεπόμενη μηχανική μάθηση και λαμβάνει υπόψη της και τα συμφραζόμενα στα οποία εφαρμόζεται ο κάθε κανόνας παράφρασης. Πειραματιστήκαμε με έναν ταξινομητή μεγίστης εντροπίας (maximum entropy) και με παλινδρόμηση διανυσμάτων υποστήριξης (support vector regression). Δοκιμάσαμε διάφορους συνδυασμούς γνωρισμάτων (features), που περιελάμβαναν γνωρίσματα βασισμένα σε ένα γλωσσικό μοντέλο *n*-γραμμάτων, γνωρίσματα βασισμένα στη σημειακή αμοιβαία πληροφορία (pointwise mutual information), γνωρίσματα βασισμένα στην ανάλυση λανθάνουσας σημασίας (latent semantic analysis), καθώς και γνωρίσματα που παράγονται από τη διαδικασία της εξαγωγής των κανόνων παράφρασης.

Στην περίπτωση του ταξινομητή μεγίστης εντροπίας, ο στόχος ήταν να διαχωρίζουμε τις παραγόμενες παραφράσεις (ουσιαστικά τις εφαρμογές των κανόνων) σε θετικές (καλές παραφράσεις) και αρνητικές (κακές). Τα πειραματικά αποτελέσματα έδειξαν ότι είχαμε πρόβλημα περιορισμένου χώρου αναζήτησης (high bias). Επιχειρήσαμε να αντιμετωπίσουμε το πρόβλημα προσθέτοντας σταδιακά περισσότερα γνωρίσματα, χωρίς όμως ιδιαίτερη επιτυχία· το ποσοστό λάθους παρέμεινε σε κάθε περίπτωση περίπου 40%. Στη συνέχεια, δοκιμάσαμε να αντιμετωπίσουμε το πρόβλημα με παλινδρόμηση διανυσμάτων υποστήριξης, χρησιμοποιώντας μη γραμμικό πυρήνα αλλά και αξιολογώντας το σύστημα με διαφορετικό τρόπο. Ο στόχος ήταν πλέον να μάθουμε να προβλέπουμε πόσο καλές ήταν οι παραγόμενες παραφράσεις, επιστρέφοντας αυτόματα ένα βαθμό όσο το δυνατόν πιο κοντά σε ένα γραμμικό συνδυασμό των βαθμών που είχαν δώσει άνθρωποι-κριτές για τη γραμματική ορθότητα και διατήρηση του νοήματος κάθε παράφρασης, καθώς και ενός αυτόματα υπολογιζόμενου βαθμού που εξέταζε πόσο διέφερε η παράφραση από την αρχική πρόταση. Εκτελέσαμε πειράματα με διαφορετικούς συνδυασμούς βαρών των τριών βαθμών, δείχνοντας ότι η μέθοδός μας, με όλα τα διαθέσιμα γνωρίσματα, ήταν καλύτερη από μια απλούστερη μέθοδο που χρησιμοποιούσε μόνο ένα γλωσσικό μοντέλο. Ωστόσο η βελτίωση που προσέφερε κάθε νέο γνώρισμα από μόνο του στα αποτελέσματα ήταν αμελητέα.

Θα ήταν ενδιαφέρον να εξεταστεί κατά πόσον μια παραλλαγή της μεθόδου της εργασίας θα μπορούσε να συνδυαστεί με μεθόδους παραγωγής παραφράσεων που δεν χρησιμοποιούν κανόνες παράφρασης. Για παράδειγμα, οι Zhao κ.ά. (Ενότητα 2.3) παράγουν υποψήφιες παραφράσεις δίνοντας την αρχική πρόταση σε πολλές μηχανές αυτόματης μετάφρασης που τη μεταφράζουν σε άλλες γλώσσες και κατόπιν πίσω στην αρχική γλώσσα· οι παραγόμενες υποψήφιες παραφράσεις θα μπορούσαν να αξιολογούνται χρησιμοποιώντας π.χ. παλινδρόμηση διανυσμάτων υποστήριξης με τα γνωρίσματα της παρούσας εργασίας.

Βιβλιογραφία

1. **Androutsopoulos, Ion and Malakasiotis, Prodromos.** A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research.* May 2010, pp. 135-187.
2. **Szpector, Idan, et al.** Contextual Preferences. *Proceedings of ACL-08: HLT.* 2008, pp. 683-691.
3. **Malakasiotis, P. and Androutsopoulos, I.** A Generate and Rank Approach to Sentence Paraphrasing. *Proceedings of the 2011 Conference on Empirical Methods on Natural Language Processing.* 2011.
4. **Bannard, Colin and Callison-Burch, Chris.** Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting of the ACL.* 2005, pp. 597-604.
5. **Koehn, P.** *Statistical Machine Translation.* 2010.
6. **Och Franz Josef, Ney Hermann.** A systematic comparison of various statistical alignment models. *Computational Linguistics.* March, 2003, Vol. 29, 1, pp. 19-51.
7. **Och, Franz Josef and Ney, Hermann.** Improved Statistical Alignment Models. *Proceedings of ACL.* 2000.
8. **Och Franz Josef, Ney Hermann.** The alignment template approach to statistical machine translation. *Computational Linguistics.* 2004, Vol. 4, 30, pp. 417-449.
9. **Callison-Burch, Chris.** Syntactic constraints on Paraphrases Extracted from Parallel Corpora. *Proceedings of EMNLP.* 2008, pp. 196-205.
10. **Zhao Shiqi, Wang Haifeng,Liu Ting, Li Sheng.** Extracting paraphrase patterns from bilingual parallel corpora. *Natural Language Engineering.* February 6, 2009, pp. 503-526.
11. **Press W.H, Teukolsky S.A, Vetterling W.T, Flannery B.P.** *Numerical recipes in C: the art of scientific computing.* Cambridge : Cambridge University Press, 1992. pp. 412-420.
12. **Kok, Stanley and Brockett, Chris.** Hitting the Right Paraphrases in Good Time. *Proceedings of HLT/NAACL.* 2010, pp. 145-153.
13. **Lovász, László.** Random walks on graphs: A survey.
14. **Sarkar Purnamrita, Moore Andrew.** A tractable approach to finding closest truncated-commute-time neighbours in large graphs. *Proceedings of the 23th Conference on Uncertainty in Artificial Intelligence.* 2007.
15. **Ravichandran, Deepak and Hovy, Eduard.** Learning surface text patterns for a question answering system. *Proceedings of ACL.* 2002.
16. **Sudo, Kiyoshi, Sekine, Satoshi and Grishman, Ralph.** An improved extraction pattern representation model for automatic IE pattern acquisition. *Proceedings of ACL.* 2003.
17. **Romano, Lorenza, et al.** Investigating a generic paraphrase-based approach for relation extraction. *Proceedings of EACL.* 2006.
18. **Szpector, Idan, Shnarch, Eyal and Dagan, Ido.** Instance-based Evaluation of Entailment Rule Acquisition. 2007, pp. 456-463.
19. **Lin, Dekang and Pantel, Patrick.** Discovery of inference rules for question answering. *Natural Language Engineering.* 2001, Vol. 7, 4, pp. 343-360.
20. **Barzilay, Regina and Lee, Lillian.** Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *Proceedings of HLT/NAACL.* 2003.
21. **Szpector, Idan, et al.** Scaling web-based acquisition of entailment relations. *Proceedings of EMNLP.* 2004.
22. **Brown Peter, Della Pietra Stephen, Della Pietra Vincent, Mercer Robert.** The mathematics of machine translation: Parameter estimation. *Computational Linguistics.* June, 1993, Vol. 19, 2, pp. 263-311.
23. **Andreas, Stolcke.** *SRILM —An extensible language modelling toolkit.*
24. **Jurafsky Daniel, H.Martin Steve.** N-grams. *Speech and Language Processing.* s.l. : Pearson Education Inc., 2009.
25. **Manning, Chris και Schutze, Hinrich.** Collocations. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press, 1999.
26. **Landauer, Thomas K, Foltz, Peter W and Laham, Darrell.** An Introduction to Latent Semantic Analysis. *Discourse Processes.* 1998, 25, pp. 259-284.
27. **Foltz, Peter W., Kintsch, Walter and Landauer, Thomas K.** The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes.* 1998, Vol. 25, 2&3, pp. 285-307.
28. **Lindsey, R.V, et al.** Be wary of what your computer reads: the effects of corpus selection on measuring semantic relatedness. *Proceedings of the 8th International Conference of Cognitive Modeling, ICCM.* 2007.
29. **Chang, Chih-Chung and Lin, Jen.** LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology.* 2011, Vol. 2, 3, pp. 1-27.
30. **Carletta, Jean.** Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics.* 1996, Vol. 22, 2.
31. **Sarkar Purnamrita, Moore Andrew, Prakash Amit.** Fast Incremental Proximity Search in Large Graphs. *Proceedings of the 25th International Conference on Machine Learning.* 2008.

32. **Pantel, Patrick Andre.** Clustering by committee. *Ph.D thesis.* 2003.
33. **Manning, Christopher and Klein, Dan.** Optimization, Maxent Models, and Conditional Estimation without Magic. *Tutorial at HLT-NAACL and ACL.* 2003.
34. **Jurafsky, Daniel και Martin, James H.** Alignment in Machine Translation. *Speech and Language Processing.* s.l. : Pearson Education Inc., 2009.
35. **Diab, Mona and Resnik, Philip.** An unsupervised method for word sense tagging using parallel corpora. *Proceedings of ACL.* 2002.
36. **Manning Chris, Schütze Hinrich.** Statistical Inference: n-gram Models over Sparse Data. *Foundations of Statistical Natural Language Processing.* Cambridge : MIT Press, 1999.