



Ονοματεπώνυμο : Σταυρούλα Ανατολάκη

Τμήμα Φοίτησης : Πληροφορική

Αριθμός Μητρώου : Π2019021

Ακαδημαϊκό email : p19anat@ionio.gr

Θέμα της Πτυχιακής Εργασίας :

« Χρήση εργαλείων Μηχανικής Μάθησης (Machine Learning) και Γλωσσικής Τεχνολογίας (Natural Language Processing) για την απλοποίηση κειμένου (text simplification).»

Επιβλέπουσα καθηγήτρια: Κάτια Κερμανίδου

Μέλη της Επιτροπής:

1) Αριστείδης Βραχάτης

2) Ανδρέας Καναβός

Περιεχόμενα

Κεφάλαιο 1. Εισαγωγή	3
Κεφάλαιο 2. Απλοποίηση κειμένου.....	4
2.1. Ορισμοί	4
2.2. Χρησιμότητα- Πεδία Εφαρμογών.....	6
2.3. Δυσκολίες- Προβλήματα στη χρήση	7
2.4. Τρόπος Υλοποίησης.....	9
Κεφάλαιο 3. Μηχανική Μάθηση	12
3.1 Ορισμός.....	12
3.2 Είδη μηχανικής μάθησης	15
3.3 Βασικές τεχνικές μηχανικής μάθησης	17
Κεφάλαιο 4. Βιβλιογραφία εργασιών μηχανικής μάθησης στην απλοποίηση κειμένου.....	19
4.1. Μελέτες που συλλέχθηκαν	19
4.2. Μεθοδολογία Έρευνας.....	63
4.2.1 Ερευνητική στρατηγική	64
4.2.2 Καθορισμός κριτηρίων εισόδου και αποκλεισμού	65
4.2.3 Διαδικασία Αναζήτηση άρθρων	66
4.2.4. Το εργαλείο του QuilBot	67
Κεφάλαιο 5. Αποτελέσματα Έρευνας.....	70
Κεφάλαιο 6. Συμπεράσματα- Προτάσεις για μελλοντική έρευνα	75
6.1. Συζήτηση των αποτελεσμάτων	77
6.2 Προτάσεις για Μελλοντική Έρευνα	79
Βιβλιογραφία	82

Κεφάλαιο 1. Εισαγωγή

Ο σκοπός της απλοποίησης κειμένου είναι να προσαρμόσει το περιεχόμενο των εγγράφων προκειμένου να γίνει πιο κατανοητό από έναν δεδομένο πληθυσμό ή να βελτιωθεί η απόδοση των εργασιών, όπως η σύνοψη ή η εξαγωγή πληροφοριών. Τα κύρια βήματα για τα αυτόματα συστήματα απλούστευσης κειμένου είναι αρκετά καλά καθορισμένα και διερευνημένα στην υπάρχουσα εργασία, αλλά η αξιολόγηση του αποτελέσματος της απλοποίησης παραμένει ανεπαρκώς μελετημένη. Πράγματι, σε αντίθεση με άλλες εργασίες, όπως η ανάκτηση και εξαγωγή πληροφοριών, η δόμηση ορολογίας ή η απάντηση ερωτήσεων, που αναμένουν πραγματικές και συναινετικές εκροές των συστημάτων, είναι δύσκολο να οριστεί μια τυπική έξοδος απλοποίησης. Υπάρχει σημαντική υποκειμενικότητα στη διαδικασία απλοποίησης και δεν είναι συναινετική επειδή βασίζεται σε μεγάλο βαθμό στη γνώση των ανθρώπων. Ως εκ τούτου, αρκετοί παράγοντες εμπλέκονται στη διαδικασία απλούστευσης και στην αξιολόγησή της.

Κεφάλαιο 2. Απλοποίηση κειμένου

2.1. Ορισμοί

Η απλοποίηση κειμένου είναι η διαδικασία μετατροπής ενός κειμένου σε ισοδύναμο που είναι πιο κατανοητό. Αυτή η απλοποίηση είναι επωφελής για πολλές ομάδες αναγνωστών, όπως μαθητές ξένων γλωσσών, ηλικιωμένους και άτομα με άλλες ειδικές ανάγκες ανάγνωσης και κατανόησης. Απλοποιημένα Κείμενα Χαρακτηρίζονται από απλό και άμεσο στυλ και μικρότερο λεξιλόγιο καθώς και απλούστερες προτάσεις. Αυτό σημαίνει ότι οι σύνθετες προτάσεις της διαδικασίας απλοποίησης συχνά χωρίζονται σε πολλές μικρότερες προτάσεις. Η απλοποίηση κειμένου έχει πολλές σημαντικές εφαρμογές όπως:

- Διευκόλυνση της πρόσβασης σε πληροφορίες σε άτομα με χαμηλό επίπεδο αλφαριθμητισμού.
- Δημιουργία ειδησεογραφικών άρθρων προσβάσιμων σε άτομα με διανοητική αναπηρία ή σε άτομα που χρειάζονται υποβοηθούμενη ανάγνωση.
- Διευκολύνοντας την πρόσβαση σε περιεχόμενο από μαθητές ξένων γλωσσών και βοήθεια στην προετοιμασία προσαρμοσμένου υλικού για την απόκτηση δεύτερης γλώσσας.
- Μεταμόρφωση κειμένου υψηλής τεχνικής πολυπλοκότητας (π.χ. διπλώματα ευρεσιτεχνίας, τεχνικά εγχειρίδια) για άτομα που δεν είναι εξοικειωμένα με τις περίπλοκες έννοιες αυτών των τύπων κειμένων.

Οι γενικεύσεις σχετικά με τα αποτελέσματα κατανόησης που προκύπτουν από αυθεντικά, απλοποιημένα και επεξεργασμένα κείμενα είναι δύσκολες λόγω των διαφορών στα σχέδια της έρευνας. Ωστόσο, η έρευνα δείχνει ότι η απλοποίηση γενικά βελτιώνει την κυριολεκτική κατανόηση. Συνολικά, κάποια τροποποίηση κειμένου τείνει να βελτιώνει την κατανόηση. Ενώ οι δομικές διαδικασίες απλοποίησης (δηλαδή η χρήση λιστών λέξεων ή τύπων αναγνωσιμότητας) μπορεί να φαίνονται σαφείς, τα αποτελέσματα της τροποποίησης του κειμένου δεν είναι καθόλου σαφή. Δηλαδή, τα κείμενα γενικά απλοποιούνται για να γίνουν πιο κατανοητά τροποποιώντας το λεξικό τους και τη συντακτική τους δομή. Ωστόσο, η διαδικασία απλοποίησης μπορεί να έχει ακούσιες συνέπειες και, σε ορισμένες περιπτώσεις, να παράγει κείμενα που είναι λιγότερο κατανοητά. Για παράδειγμα, η εξάρτηση από το λεξιλόγιο υψηλής συχνότητας μπορεί να οδηγήσει ακούσια στη χρήση πιο διαφορετικών λέξεων. Ομοίως, η απλοποίηση του λεξιλογίου μπορεί να

οδηγήσει σε μια συμπτυκνωμένη συντακτική δομή που είναι πιο περίπλοκη. Αυτό μπορεί να ισχύει ιδιαίτερα όταν οι συγγραφείς εξαρτώνται από τη διαίσθησή τους όταν απλοποιούν το κείμενο.

Ενώ οι θεωρίες που προωθούν τα οφέλη ή τα μειονεκτήματα των απλουστευμένων κειμένων είναι σημαντικές, οι πραγματικές αποδείξεις για τη δύναμη των απλοποιημένων κειμένων βρίσκονται σε πειραματικές μελέτες. Για παράδειγμα, οι Long και Ross (1993) συνέκριναν τις βαθμολογίες κατανόησης των αναγνωστών χρησιμοποιώντας τρεις εκδόσεις του ίδιου κειμένου: αυθεντικές, απλοποιημένες και επεξεργασμένες εκδόσεις. Ο Ross διαπίστωσε ότι οι μαθητές που διάβασαν το γλωσσικά απλοποιημένο κείμενο σημείωσαν σημαντικά υψηλότερη βαθμολογία σε στοιχεία πολλαπλών επιλογών που προορίζονταν να δοκιμάσουν την κατανόηση από εκείνους που διάβασαν την αυθεντική έκδοση. Διαπίστωσαν επίσης ότι οι μαθητές που διάβασαν την επεξεργασμένη έκδοση δεν σημείωσαν σημαντικά καλύτερη βαθμολογία από εκείνους που διάβασαν το αυθεντικό ή το απλοποιημένο κείμενο. Αυτά τα αποτελέσματα επαναλήφθηκαν σε μια μελέτη παρακολούθησης, η οποία έδειξε ότι δεν υπήρχαν σημαντικές διαφορές μεταξύ απλοποιημένων και επεξεργασμένων κειμένων, αλλά τα απλουστευμένα κείμενα έδειξαν σημαντικά οφέλη στην αναγνωστική κατανόηση έναντι των αυθεντικών κειμένων, ενώ τα επεξεργασμένα κείμενα έκαναν παρουσιάζουν κέρδη έναντι των αυθεντικών κειμένων. Μια μεταγενέστερη μελέτη που διεξήχθη από τον Tweissi (1998) διαπίστωσε επίσης ότι η απλοποίηση επηρέασε θετικά την αναγνωστική κατανόηση των μαθητών. Ωστόσο, ο Tweissi διαπίστωσε ότι ήταν το είδος της απλοποίησης και όχι το μέγεθος της απλοποίησης που έπαιξε τον πιο σημαντικό ρόλο για να βοηθήσουν τους μαθητές να κατανοήσουν το υλικό (π.χ. κείμενα με απλοποιημένο λεξικό οδήγησαν σε μεγαλύτερα κέρδη κατανόησης από ό,τι άλλοι τύποι τροποποιήσεων κειμένου). Συνολικά, αυτές οι μελέτες υποστηρίζουν τη βασική ιδέα πίσω από τα απλουστευμένα κείμενα: ότι η χρήση απλοποιημένης εισαγωγής οδηγεί σε πιο κατανοητή γλώσσα.

2.2. Χρησιμότητα- Πεδία Εφαρμογών

Ενώ η απλοποίηση κειμένου μπορεί να έχει τεχνικές εφαρμογές σε διάφορους αγωγούς, π.χ. ως πρόδρομος για τη μηχανική μετάφραση, ένας αυξανόμενος όγκος ερευνών έχει εξετάσει τη χρήση της για την παροχή εργαλείων βοήθειας ανάγνωσης σε συγκεκριμένες ομάδες ανθρώπων που μπορεί να επωφεληθούν από την υποστήριξη ανάγνωσης. Έρευνες για το ATS ως υποβοηθητική τεχνολογία επικεντρώθηκε στη χρήση του για μια σειρά από ομάδες χρηστών, συμπεριλαμβανομένων ατόμων με δυσλεξία, με αφασία, μαθητές δεύτερης γλώσσας, παιδιά, αναγνώστες χαμηλού γραμματισμού, και πιο στενά σε σχέση με την έρευνά μας, DHH ενήλικες και παιδιά. Οι τρεις προσεγγίσεις που προσδιορίζονται παραπάνω μπορεί να ωφελήσουν τις ομάδες χρηστών σε διαφορετικό βαθμό, ανάλογα με το προφίλ αλφαριθμητισμού τους. Προηγούμενη εργασία ρίχνει φως στους λόγους για αυτό, εντοπίζοντας ότι οι πηγές δυσκολίας ανάγνωσης για διαφορετικές ομάδες χρηστών μπορεί να διαφέρουν, και επομένως τα ευρήματα από τη μια ομάδα χρηστών μπορεί να μην γενικεύονται σε άλλη, καθώς διαφορετικές ομάδες χρηστών μπορεί να επωφεληθούν από τις διάφορες προσεγγίσεις στο ATS.

Έχουν γίνει πολλές μελέτες χρηστών σε αυτόν τον τομέα, εστιάζοντας σε διαφορετικές πτυχές των συστημάτων. Για παράδειγμα, προηγούμενες μελέτες ATS που επικεντρώθηκαν σε συγκεκριμένες ομάδες-στόχους χρηστών έχουν διερευνήσει εάν οι χρήστες επωφελούνται από τις διάφορες προσεγγίσεις στην απλοποίηση κειμένου ή συγκρίνοντας πώς η χρήση διαφορετικών συστημάτων επηρεάζει αυτά τα οφέλη. Λίγες μελέτες μέτρησαν οποιαδήποτε διαφορά σε αντικειμενικά μέτρα (π.χ. ερωτήσεις κατανόησης), αλλά ορισμένες μέτρησαν τη διαφορά στην υποκειμενική αντίληψη των χρηστών για το εάν επωφελήθηκαν από την τεχνολογία.

2.3. Δυσκολίες- Προβλήματα στη χρήση

Αν και έχουν γίνει πολλές προσπάθειες για τις τεχνικές TS τις τελευταίες δύο δεκαετίες, λίγες έχουν χρησιμοποιηθεί στην παραγωγή. Αυτά που χρησιμοποιούνται στην παραγωγή αναπτύσσονται γενικά με έμφαση ως βοήθημα προς τον χρήστη στη μετάφραση του κειμένου του σε απλοποιημένη γλώσσα. Ένας βασικός παράγοντας στη χαμηλή απορρόφηση των συστημάτων TS είναι η ανακρίβεια. Σε ορισμένες εφαρμογές επεξεργασίας φυσικής γλώσσας, μπορεί να είναι αποδεκτή χαμηλή ακρίβεια, καθώς η εφαρμογή εξακολουθεί να είναι χρησιμοποιήσιμη. Για παράδειγμα, στην ανάκτηση πληροφοριών, ακόμα κι αν ένα σύστημα έχει μέτρια ακρίβεια, θα επιτρέψει στον χρήστη να βρει μέρος των εγγράφων που αναζητούσε. Χωρίς το σύστημα ανάκτησης πληροφοριών, ο χρήστης δεν θα μπορούσε να βρει τα έγγραφα τόσο εύκολα. Ωστόσο, αυτό δεν μεταφέρεται με τον ίδιο τρόπο στο TS. Εάν ένα σύστημα δεν είναι ακριβές, τότε το κείμενο που προκύπτει δεν θα έχει νόημα. Αν το κείμενο δεν είναι κατανοητό, τότε σίγουρα δεν θα είναι πιο απλό από το πρωτότυπο. Εάν ένας χρήστης παρουσιάζεται τακτικά με ανακριβείς απλοποιήσεις, τότε δεν θα το βρει χρήσιμο. Η υποβοηθητική τεχνολογία πρέπει να είναι επακριβώς υποβοηθητική, διαφορετικά ο χρήστης θα είναι πιο μπερδεμένος και λιγότερο ικανός να αλληλεπιδράσει με το κείμενο από ό,τι στην αρχική του «πιο περίπλοκη» μορφή. Το TS είναι ένα σε μεγάλο βαθμό άλυτο έργο. Ενώ πολλοί τομείς της επεξεργασίας φυσικής γλώσσας και της επιστήμης των υπολογιστών έχουν ένα εμβληματικό σύστημα, μέθοδο ή τεχνική, το TS έχει πολλές ποικίλες προσεγγίσεις. Ενώ αυτές οι τεχνικές χρησιμοποιούν διαφορετικές μεθοδολογίες και μπορεί να έχουν διαφορετικά αποτελέσματα, σκοπός τους είναι πάντα η απλοποίηση του κειμένου. Το πεδίο κινείται γρήγορα και διεξάγεται τακτικά έρευνα σε νέους τομείς.

Επιπλέον, όλο και περισσότεροι άνθρωποι ενδιαφέρονται για το TS με αυξανόμενο αριθμό έργων και δημοσιεύσεων από έτος σε έτος. Ενώ έχουν εφαρμοστεί πολλές τεχνικές, υπάρχει ακόμη πολλή δουλειά που πρέπει να γίνει για τη σύγκριση, την αξιολόγηση και τη βελτίωση αυτά τα. Το κείμενο είναι ένα θεμελιώδες μέρος της καθημερινής μας αλληλεπίδρασης με τον κόσμο της πληροφορίας. Εάν το κείμενο είναι απλοποιημένο για έναν τελικό χρήστη, τότε αυτό μπορεί να βελτιώσει την εμπειρία και την ποιότητα ζωής του. Είτε διαβάζουμε την εφημερίδα, ελέγχουμε μηνύματα ηλεκτρονικού ταχυδρομείου ή ακολουθούμε οδηγίες, είναι πολύ σημαντικό να μπορούμε να κατανοήσουμε το κείμενο που χρησιμοποιείται για τη μετάδοση αυτών των

πληροφοριών. Το TS μπορεί να εφαρμοστεί για να μειώσει την πολυπλοκότητα των πληροφοριών και να αυξήσει την κατανόηση από τον χρήστη του κειμένου που συναντά στην καθημερινή του ζωή. Αυτό έχει μεγάλα πλεονεκτήματα τόσο για τους αναγνώστες όσο και για τους συγγραφείς. Ο αναγνώστης αποκτά καλύτερη κατανόηση του κόσμου γύρω του και οι συγγραφείς μπορούν να διασφαλίσουν ότι το γραπτό τους υλικό θα είναι κατανοητό από εκείνους τους παραλήπτες με χαμηλό επίπεδο ανάγνωσης.

Τα σχετικά πεδία της αυτόματης μετάφρασης και της σύνοψης κειμένων επιτρέπουν τη διασταύρωση και την κοινή χρήση τεχνικών. Για παράδειγμα, οι τεχνικές ευθυγράμμισης σώματος έχουν δανειστεί από τη σύνοψη και έχουν χρησιμοποιηθεί στατιστικές τεχνικές αυτόματης μετάφρασης, μαζί με τις μεθόδους αξιολόγησής τους. Αυτή η διασταύρωση σημαίνει ότι, καθώς προχωρούν αυτά τα πεδία, θα υπάρχουν διαθέσιμες νέες τεχνικές για το έργο της απλοποίησης. Καθώς οι τεχνικές αναπτύσσονται στο πλαίσιο του TS, θα είναι επίσης χρήσιμες στο πλαίσιο άλλων σχετικών τομέων. Η ανάγκη για απλοποιημένα Αγγλικά ειδικότερα αποδεικνύεται από τη δημοτικότητα του έργου Simple English Wikipedia (μια εναλλακτική στην αγγλική Wikipedia), το οποίο παρέχει απλουστευμένες εκδόσεις των άρθρων της Wikipedia. Υπάρχουν πάνω από 88.000 άρθρα που έχουν γραφτεί στο χέρι σε απλά αγγλικά για αυτό το έργο. Πολλές ομάδες με χαμηλά επίπεδα αγγλικών επωφελούνται. Το μέγεθος της Απλής Βικιπαίδειας υποδηλώνει την ανάγκη για απλά αγγλικά, ωστόσο η διαδικασία χειροτεχνίας αυτών των άρθρων είναι χρονοβόρα. Οι βελτιώσεις στην αυτοματοποίηση της απλούστευσης θα βοηθούσαν στην αντιμετώπιση αυτής της ανάγκης.

2.4. Τρόπος Υλοποίησης

Το TS εντάσσεται στον τομέα της επεξεργασίας φυσικής γλώσσας. Σε αυτό το πεδίο είναι πολύ παρόμοιο με άλλες τεχνικές όπως η αυτόματη μετάφραση, η μονόγλωσση δημιουργία κειμένου σε κείμενο, η σύνοψη κειμένου και η παραγωγή παράφρασης. Όλα αυτά τα πεδία βασίζονται το ένα στο άλλο για τεχνικές και πόρους και πολλές τεχνικές στο TS προέρχονται από αυτά τα άλλα πεδία. Το TS διαφέρει από τη σύνοψη κειμένου, καθώς το επίκεντρο της σύνοψης κειμένου είναι η μείωση του μήκους και του περιεχομένου της εισαγωγής. Ενώ τα απλουστευμένα κείμενα είναι συνήθως πιο σύντομα, αυτό δεν συμβαίνει απαραίτητα και η απλούστευση μπορεί να οδηγήσει σε μεγαλύτερη απόδοση — ειδικά κατά τη δημιουργία επεξηγήσεων. Η περίληψη στοχεύει επίσης στη μείωση του περιεχομένου — αφαιρώντας αυτό που μπορεί να είναι λιγότερο σημαντικό ή περιττό. Αυτό συνήθως δεν διερευνάται με απλοποίηση, όπου συνήθως διατηρείται όλο το περιεχόμενο. Ορισμένες προσπάθειες έχουν διερευνήσει τη χρήση της απλούστευσης παράλληλα με τα συστήματα σύνοψης. Εδώ, το TS χρησιμοποιείται για τη βελτίωση της αναγνωσιμότητας της τελικής περίληψης. Όταν μιλάμε για TS, οι λέξεις απλές και σύνθετες χρησιμοποιούνται συχνά μεταξύ τους. Για παράδειγμα, σε ένα παράλληλο σώμα απλοποιημένων αγγλικών και κανονικών αγγλικών, το πρώτο θα ονομάζεται απλό και το δεύτερο σύνθετο.

Σε ένα σύνολο τεχνικών αγγλικών και κανονικών αγγλικών, το πρώτο θα ονομάζεται σύνθετο και το δεύτερο απλό. Αυτό δείχνει ότι η απλότητα και η πολυπλοκότητα σχετίζονται μεταξύ τους και πρέπει να χρησιμοποιούνται με προσοχή. Όταν δημιουργούμε απλό κείμενο, στην πραγματικότητα σκοπεύουμε να δημιουργήσουμε κείμενο που είναι πιο απλό (και τόσο λιγότερο σύνθετο) από ό,τι ήταν αρχικά. Δύο άλλοι σημαντικοί όροι που πρέπει να οριστούν είναι η αναγνωσιμότητα και η κατανόηση. Στην αρχή, αυτά μπορεί να φαίνονται σαν τα ίδια πράγματα, ωστόσο μπορεί να μετρηθούν ανεξάρτητα ανάλογα με το πλαίσιο μιας εφαρμογής. Η αναγνωσιμότητα καθορίζει πόσο εύκολο μπορεί να είναι η ανάγνωση ενός κειμένου. Αυτό συνήθως διέπεται από παράγοντες όπως η πολυπλοκότητα της γραμματικής, το μήκος των προτάσεων και η εξοικείωση με το λεξιλόγιο. Κατανοησιμότητα είναι ο όγκος των πληροφοριών που μπορεί να αποκτήσει ένας χρήστης από ένα κομμάτι κειμένου. Αυτό μπορεί να επηρεαστεί από παράγοντες όπως η εξοικείωση του χρήστη με το λεξιλόγιο της πηγής, η κατανόηση βασικών εννοιών ή ο χρόνος και η προσοχή που δόθηκε για την ανάγνωση του κειμένου. Μπορεί ένα

κείμενο να έχει υψηλή αναγνωσιμότητα, αλλά χαμηλή κατανόηση. Για παράδειγμα: προσπάθεια ανάγνωσης ενός καλογραμμένου επιστημονικού άρθρου χωρίς επιστημονική κατάρτιση. Μπορεί επίσης να είναι πιθανό ένα κείμενο να έχει χαμηλή αναγνωσιμότητα, αλλά να είναι ακόμα κατανοητό. Για παράδειγμα: ένας συγγραφέας που επικοινωνεί ένα απλό σημείο χρησιμοποιεί παραπλανητικές γραμματικές δομές. Η αναγνωσιμότητα και η κατανοητότητα σχετίζονται και ένα κείμενο που είναι πιο ευανάγνωστο είναι πιθανό να είναι πιο κατανοητό, καθώς ο αναγνώστης θα βρει ευκολότερο χρόνο για να εξετάσει τις δύσκολες έννοιες. Ομοίως, ένα κείμενο που είναι εύκολα κατανοητό θα ενθαρρύνει τον αναγνώστη να συνεχίσει να διαβάζει, ακόμη και όταν είναι δύσκολη η αναγνωσιμότητα. Η απλότητα είναι διαισθητικά προφανής, αλλά δύσκολο να οριστεί.

Τα τυπικά μέτρα λαμβάνουν υπόψη παράγοντες όπως το μήκος της πρότασης, τον αριθμό των συλλαβών και άλλους παράγοντες επιφανειακού κειμένου. Αν και δίνουν μια καλή εκτίμηση, δεν είναι πάντα ακριβείς. Για παράδειγμα, πάρτε την περίπτωση του μήκους της πρότασης. Μια μεγάλη πρόταση μπορεί να χρησιμοποιεί πολλούς σύνθετους όρους και αναφορές (λέξεις που αναφέρονται σε προηγούμενη οντότητα: αυτός, αυτή, αυτό, κ.λπ.). Μια απλοποιημένη εκδοχή αυτού θα ήταν λεξιλογικά μεγαλύτερη, αλλά μπορεί να είναι πιο επεξηγητική. Στην περίπτωση δημιουργίας επεξηγήσεων, οι σύνθετες λέξεις επισυνάπτονται με σύντομους ορισμούς, αυξάνοντας το μήκος της πρότασης. Τα αυτόματα ευρετικά μέτρα θα κρίνουν αυτές τις προτάσεις που έχουν απλοποιηθεί ως πιο περίπλοκες.

Το τελικό κείμενο μπορεί να είναι μεγαλύτερο, ωστόσο είναι επίσης πιο κατανοητό και επομένως πιο απλό. Διαφορετικές μορφές απλούστευσης θα καλύψουν διαφορετικές ανάγκες. Δεν υπάρχουν δύο χρήστες που να είναι ακριβώς ίδιοι και ό,τι βρίσκει εύκολα κάποιος, μπορεί να το βρει δύσκολο για άλλον. Αυτό ισχύει τόσο σε επίπεδο διαφορετικών ομάδων χρηστών (ο χρήστης χαμηλού αλφαριθμητισμού απαιτεί διαφορετικές απλουστεύσεις για τον εκπαιδευόμενο της δεύτερης γλώσσας), αλλά και εντός ομάδων χρηστών. Παράγοντες όπως η διάλεκτος, η καθομιλούμενη και η εξοικείωση με το λεξιλόγιο και τη σύνταξη μπορούν όλοι να επηρεάσουν την κατανόηση ενός κειμένου από τον χρήστη. Αυτό σημαίνει ότι η απλοποίηση γίνεται καλύτερα σε γενικό επίπεδο. Το κείμενο που είναι πολύ απλό για έναν χρήστη μπορεί να είναι πιο περίπλοκο για έναν άλλο.

Ωστόσο, το κείμενο που γίνεται ελαφρώς απλούστερο για έναν χρήστη θα είναι γενικά ευκολότερο για τους περισσότερους άλλους χρήστες.

Κεφάλαιο 3. Μηχανική Μάθηση

3.1 Ορισμός

Η Μηχανική Μάθηση σύμφωνα με τους Michie et al (D. Michie, 1994) θεωρείται γενικά ότι περιλαμβάνει αυτόματες διαδικασίες υπολογισμού που βασίζονται σε λογικές ή δυαδικές πράξεις που μαθαίνουν μια εργασία από μια σειρά παραδειγμάτων. Εδώ ασχολούμαστε απλώς με την ταξινόμηση και είναι αμφισβητήσιμο τι πρέπει να μπει κάτω από την ομπρέλα της Μηχανικής Μάθησης. Η προσοχή έχει επικεντρωθεί στις προσεγγίσεις του δέντρου αποφάσεων, στις οποίες η ταξινόμηση προκύπτει από μια ακολουθία λογικών βημάτων. Αυτά είναι ικανά να αντιπροσωπεύουν το πιο σύνθετο πρόβλημα με επαρκή δεδομένα (αλλά αυτό μπορεί να σημαίνει τεράστιο ποσό!). Άλλες τεχνικές, όπως οι γενετικοί αλγόριθμοι και οι διαδικασίες επαγωγικής λογικής (ILP), βρίσκονται επί του παρόντος υπό ενεργό ανάπτυξη και κατ' αρχήν θα μας επιτρέψουν να ασχοληθούμε με γενικότερους τύπους δεδομένων, συμπεριλαμβανομένων περιπτώσεων όπου ο αριθμός και ο τύπος των χαρακτηριστικών μπορεί να ποικίλλει και όπου πρόσθετα επικαλύπτονται επίπεδα μάθησης, με ιεραρχική δομή ιδιοτήτων και κλάσεων κ.ο.κ. Η Μηχανική Μάθηση στοχεύει στη δημιουργία εκφράσεων ταξινόμησης αρκετά απλές ώστε να γίνονται εύκολα κατανοητές από τον άνθρωπο. Πρέπει να μιμούνται επαρκώς τον ανθρώπινο συλλογισμό για να παρέχουν εικόνα στη διαδικασία λήψης αποφάσεων. Όπως και οι στατιστικές προσεγγίσεις, η βασική γνώση μπορεί να αξιοποιηθεί κατά την ανάπτυξη, αλλά η λειτουργία υποτίθεται χωρίς ανθρώπινη παρέμβαση. Για να μάθετε είναι:

1. απόκτηση γνώσης, κατανόησης ή κυριαρχίας μέσω της εμπειρίας ή της μελέτης ή απόκτηση γνώσης (κάτι) ή απόκτηση δεξιοτήτων σε (κάποια τέχνη ή πρακτική)
2. να αποκτήσει εμπειρία ή ικανότητα ή δεξιότητα
3. να απομνημονεύσω (κάτι), να αποκτήσω με εμπειρία, παράδειγμα ή πρακτική.

Η Μηχανική Μάθηση μπορεί να οριστεί ως μια διαδικασία κατασκευής συστημάτων υπολογιστών που βελτιώνονται αυτόματα με την εμπειρία και εφαρμόζουν μια διαδικασία μάθησης. Η Μηχανική Μάθηση μπορεί ακόμα να οριστεί ως η αυτόματη εκμάθηση της θεωρίας από τα δεδομένα, μέσω μιας διαδικασίας συμπερασμάτων, προσαρμογής μοντέλων ή εκμάθησης από παραδείγματα: Αυτοματοποιημένη εξαγωγή χρήσιμων πληροφοριών από ένα σύνολο δεδομένων

με τη δημιουργία καλών πιθανολογικών μοντέλων. Ιδανικά κατάλληλο για περιοχές με πολλά δεδομένα ελλείπει γενικής θεωρίας.

Η κύρια εστίαση της έρευνας μηχανικής μάθησης είναι η αυτόματη παραγωγή μοντέλων και ένα μοντέλο είναι ένα μοτίβο, σχέδιο, αναπαράσταση ή περιγραφή που έχει σχεδιαστεί για να δείξει την κύρια λειτουργία ενός συστήματος ή έννοια, όπως κανόνες καθορίζουν τον κανόνα για την εκτέλεση μιας μαθηματικής πράξης και την απόκτηση ένα ορισμένο αποτέλεσμα, μια συνάρτηση από σύνολα τύπων σε τύπους, και μοτίβα (μοντέλο που μπορεί να χρησιμοποιηθεί για τη δημιουργία πραγμάτων ή τμημάτων ενός πράγματος από δεδομένα. Η μάθηση είναι ένα ΦΑΙΝΟΜΕΝΟ ΠΟΛΛΑΠΛΩΝ όπως περιγράφεται από τους Jaime et al (Jaime G. Carbonell , 1983) και δήλωσε επίσης ότι οι διαδικασίες μάθησης περιλαμβάνουν την απόκτηση νέας δηλωτικής γνώσης, την ανάπτυξη κινητικών και γνωστικών δεξιοτήτων μέσω διδασκαλίας ή πρακτικής, την οργάνωση της νέας γνώσης σε γενικές, αποτελεσματικές αναπαραστάσεις και την ανακάλυψη νέων γεγονότων και θεωριών μέσω της παρατήρησης. Η μελέτη και η υπολογιστική μοντελοποίηση των μαθησιακών διαδικασιών στις πολλαπλές εκφάνσεις τους αποτελεί το αντικείμενο της μηχανικής μάθησης.

Αν και η μηχανική μάθηση ήταν κεντρικό μέλημα της τεχνητής νοημοσύνης από τις πρώτες μέρες που ήταν δημοφιλής η ιδέα των «αυτο-οργανωμένων συστημάτων», οι περιορισμοί που ήταν εγγενείς στις πρώτες προσεγγίσεις νευρωνικών δικτύων οδήγησαν σε προσωρινή μείωση του όγκου της έρευνας. Πιο πρόσφατα, νέες συμβολικές μέθοδοι και τεχνικές έντασης γνώσης έχουν αποφέρει πολλά υποσχόμενα αποτελέσματα και αυτά με τη σειρά τους έχουν οδηγήσει στην τρέχουσα αναβίωση της έρευνας μηχανικής μάθησης. Η Μηχανική Μάθηση σύμφωνα με τους Michie et al (D. Michie, 1994) θεωρείται γενικά ότι περιλαμβάνει αυτόματες διαδικασίες υπολογισμού που βασίζονται σε λογικές ή δυαδικές πράξεις που μαθαίνουν μια εργασία από μια σειρά παραδειγμάτων. Εδώ ασχολούμαστε απλώς με την ταξινόμηση και είναι αμφισβητήσιμο τι πρέπει να μπει κάτω από την ομπρέλα της Μηχανικής Μάθησης. Η προσοχή έχει επικεντρωθεί στις προσεγγίσεις του δέντρου αποφάσεων, στις οποίες η ταξινόμηση προκύπτει από μια ακολουθία λογικών βημάτων. Αυτά είναι ικανά να αντιπροσωπεύουν το πιο σύνθετο πρόβλημα με επαρκή δεδομένα (αλλά αυτό μπορεί να σημαίνει τεράστιο ποσό!). Άλλες τεχνικές, όπως οι

γενετικοί αλγόριθμοι και οι διαδικασίες επαγωγικής λογικής (ILP), βρίσκονται επί του παρόντος υπό ενεργό ανάπτυξη και κατ' αρχήν θα μας επιτρέψουν να ασχοληθούμε με γενικότερους τύπους δεδομένων, συμπεριλαμβανομένων περιπτώσεων όπου ο αριθμός και ο τύπος των χαρακτηριστικών μπορεί να ποικίλλει και όπου πρόσθετα επικαλύπτονται επίπεδα μάθησης, με ιεραρχική δομή ιδιοτήτων και κλάσεων κ.ο.κ.

Η Μηχανική Μάθηση στοχεύει στη δημιουργία εκφράσεων ταξινόμησης αρκετά απλές ώστε να γίνονται εύκολα κατανοητές από τον άνθρωπο. Πρέπει να μιμούνται επαρκώς τον ανθρώπινο συλλογισμό για να παρέχουν εικόνα στη διαδικασία λήψης αποφάσεων. Όπως και οι στατιστικές προσεγγίσεις, η βασική γνώση μπορεί να αξιοποιηθεί κατά την ανάπτυξη, αλλά η λειτουργία υποτίθεται χωρίς ανθρώπινη παρέμβαση. Η μηχανική μάθηση ήταν ανέκαθεν αναπόσπαστο μέρος της τεχνητής νοημοσύνης σύμφωνα με τους Jaime et al (Jaime G. Carbonell, 1983) και η μεθοδολογία της εξελίχθηκε σε συνεννόηση, με τις κύριες ανησυχίες του χώρου. Ως απάντηση στις δυσκολίες κωδικοποίησης των συνεχώς αυξανόμενων όγκων γνώσης σε συστήματα μοντέλων τεχνητής νοημοσύνης, πολλοί ερευνητές έστρεψαν πρόσφατα την προσοχή τους στη μηχανική μάθηση ως μέσο για να ξεπεραστεί το εμπόδιο στην απόκτηση γνώσης. Αυτό το βιβλίο παρουσιάζει μια ταξινομική ανάλυση της μηχανικής μάθησης που οργανώνεται κυρίως από στρατηγικές μάθησης και δευτερευόντως από αναπαράσταση γνώσης και τομείς εφαρμογής. Μια ιστορική έρευνα που περιγράφει την ανάπτυξη διαφόρων προσεγγίσεων στη μηχανική μάθηση παρουσιάζεται από τα πρώιμα νευρωνικά δίκτυα έως τις παρούσες τεχνικές έντασης γνώσης.

3.2 Είδη μηχανικής μάθησης

Υπάρχουν αρκετοί τομείς μηχανικής εκμάθησης που θα μπορούσαν να αξιοποιηθούν για την επίλυση των προβλημάτων διαχείρισης email και η προσέγγισή μας εφάρμοσε τη μέθοδο μηχανικής εκμάθησης χωρίς επίβλεψη. Η μάθηση χωρίς επίβλεψη είναι μια μέθοδος μηχανικής μάθησης όπου ο αλγόριθμος παρουσιάζεται με παραδείγματα μόνο από τον χώρο εισόδου και ένα μοντέλο ταιριάζει σε αυτές τις παρατηρήσεις. Για παράδειγμα, ένας αλγόριθμος ομαδοποίησης θα ήταν μια μορφή μάθησης χωρίς επίβλεψη. «Η μάθηση χωρίς επίβλεψη είναι μια μέθοδος μηχανικής μάθησης όπου ένα μοντέλο είναι κατάλληλο για παρατηρήσεις. Διακρίνεται από την εποπτευόμενη μάθηση από το γεγονός ότι δεν υπάρχει *a priori*. Σε μάθηση χωρίς επίβλεψη, συλλέγεται ένα σύνολο δεδομένων από αντικείμενα εισόδου. Η μάθηση χωρίς επίβλεψη συνήθως αντιμετωπίζει τα αντικείμενα εισόδου ως ένα σύνολο τυχαίων μεταβλητών. Στη συνέχεια δημιουργείται ένα μοντέλο πυκνότητας άρθρωσης για το σύνολο δεδομένων. Το πρόβλημα της μάθησης χωρίς επίβλεψη περιλάμβανε μοτίβα μάθησης στην είσοδο όταν δεν παρέχονται συγκεκριμένες τιμές εξόδου» σύμφωνα με τον Russell (Russell, 2003).

Υπάρχουν διάφορες κατηγορίες στον τομέα της τεχνητής νοημοσύνης. Οι ταξινομήσεις των συστημάτων μηχανικής μάθησης είναι:

- Εποπτευόμενη μηχανική μάθηση: Η εποπτευόμενη μάθηση είναι μια τεχνική μηχανικής μάθησης για την εκμάθηση μιας συνάρτησης από δεδομένα εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούνται από ζεύγη αντικειμένων εισόδου (συνήθως διανύσματα) και επιθυμητές εξόδους. Η έξοδος της συνάρτησης μπορεί να είναι μια συνεχής τιμή (που ονομάζεται παλινδρόμηση), ή μπορεί να προβλέψει μια ετικέτα κλάσης του αντικειμένου εισόδου (που ονομάζεται ταξινόμηση). Το καθήκον του εποπτευόμενου εκπαιδευόμενου είναι να προβλέψει την τιμή της συνάρτησης για οποιοδήποτε έγκυρο αντικείμενο εισόδου αφού έχει δει έναν αριθμό παραδειγμάτων εκπαίδευσης (δηλαδή ζεύγη εισόδου και εξόδου στόχου). Για να το πετύχει αυτό, ο εκπαιδευόμενος πρέπει να γενικεύσει από τα παρουσιαζόμενα δεδομένα σε αόρατες καταστάσεις με «λογικό» τρόπο (βλ. επαγωγική προκατάληψη). (Σύγκριση με μάθηση χωρίς επίβλεψη.)
- Μη εποπτευόμενη μηχανική εκμάθηση: Η μη εποπτευόμενη εκμάθηση¹ είναι ένας τύπος μηχανικής εκμάθησης όπου δεν χρησιμοποιούνται χειροκίνητες ετικέτες εισόδων.

Διακρίνεται από τις εποπτευόμενες προσεγγίσεις μάθησης που μαθαίνουν πώς να εκτελούν μια εργασία, όπως η ταξινόμηση ή η παλινδρόμηση, χρησιμοποιώντας ένα σύνολο παραδειγμάτων προετοιμασμένων από τον άνθρωπο. Μάθηση χωρίς επίβλεψη σημαίνει ότι μας δίνονται μόνο τα X και κάποια (τελική) λειτουργία ανατροφοδότησης για την απόδοσή μας. Έχουμε απλώς ένα εκπαιδευτικό σύνολο διανυσμάτων χωρίς τιμές συνάρτησης αυτών. Το πρόβλημα σε αυτήν την περίπτωση, τυπικά, είναι να χωρίσουμε το σετ εκπαίδευσης σε υποσύνολα, $1 \dots R$, με κάποιο κατάλληλο τρόπο.

3.3 Βασικές τεχνικές μηχανικής μάθησης

Η ταξινόμηση του συστήματος μηχανικής μάθησης θα μπορούσε να εφαρμοστεί σε πολλές διαφορετικές διαστάσεις και έχουμε επιλέξει αυτές τις δύο διαστάσεις: Εκμάθηση συμπερασμάτων: Αυτή είναι μια μορφή ταξινόμησης με βάση την υποκείμενη στρατηγική που εμπλέκεται. Αυτές οι στρατηγικές θα εξαρτηθούν από το μέγεθος των συμπερασμάτων που εκτελεί το σύστημα εκμάθησης σχετικά με τις πληροφορίες που παρέχονται στο σύστημα. Τώρα οι στρατηγικές μάθησης διακρίνονται από το μέγεθος των συμπερασμάτων που εκτελεί ο εκπαιδευόμενος σχετικά με τις παρεχόμενες πληροφορίες. Έτσι, εάν ένα σύστημα υπολογιστή πραγματοποιεί ταξινόμηση email για παράδειγμα, οι γνώσεις του αυξάνονται, αλλά αυτό μπορεί να μην παράγει κανένα συμπέρασμα για τις νέες πληροφορίες, αυτό σημαίνει ότι όλες οι γνωστικές προσπάθειες είναι από την πλευρά του αναλυτή ή του προγραμματιστή.

Αλλά εάν ο ταξινομητής μηχανικής μάθησης ανακαλύψει ανεξάρτητα νέες θεωρίες ή υιοθετήσει νέες έννοιες, αυτό θα δώσει ένα πολύ σημαντικό συμπέρασμα. Αυτό ονομάζεται απόκτηση γνώσης από παράδειγμα ή πειράματα ή με παρατήρηση. Ένα παράδειγμα είναι: Όταν ένας μαθητής θέλει να λύσει στατιστικά προβλήματα σε ένα εγχειρίδιο – αυτή είναι μια διαδικασία που περιλαμβάνει συμπέρασμα, αλλά η λύση δεν είναι να ανακαλύψετε μια ολοκαίνουργια φόρμουλα χωρίς καθοδήγηση από έναν δάσκαλο ή ένα βιβλίο. Έτσι, καθώς αυξάνεται ο όγκος των συμπερασμάτων που μπορεί να κάνει ο εκπαιδευόμενος, μειώνονται τα βάρη που επιβαρύνουν τον δάσκαλο ή το εξωτερικό περιβάλλον. Σύμφωνα με τον Jaime (Jaime G. Carbonell, 1983), ο οποίος δήλωσε ότι είναι πολύ πιο δύσκολο να διδάξεις ένα άτομο εξηγώντας κάθε βήμα σε μια περίπλοκη εργασία παρά δείχνοντας σε αυτό τον τρόπο ότι παρόμοιες εργασίες είναι συνήθως Έγινε. Είναι ακόμη πιο δύσκολο να προγραμματίσετε έναν υπολογιστή για να εκτελέσετε μια σύνθετη εργασία από το να δώσουμε εντολή σε ένα άτομο να εκτελέσει την εργασία. καθώς ο προγραμματισμός απαιτεί σαφή προδιαγραφή όλων των προαπαιτούμενων λεπτομερειών, ενώ ένα άτομο που λαμβάνει οδηγίες μπορεί να χρησιμοποιήσει την προηγούμενη γνώση και την κοινή λογική για να συμπληρώσει τις περισσότερες κοσμικές λεπτομέρειες. Αναπαράσταση γνώσης: Αυτή είναι μια μορφή δεξιοτήτων που αποκτά ο εκπαιδευόμενος με βάση τον τύπο αναπαράστασης της γνώσης.

Υπάρχουν πολλά άλλα υπάρχοντα συστήματα μάθησης που χρησιμοποιούν πολλαπλές στρατηγικές και αναπαραστάσεις γνώσης και ορισμένα έχουν εφαρμοστεί σε περισσότερες από μία. Στη μέθοδο μηχανικής μάθησης που βασίζεται στη γνώση, δεν χρησιμοποιείται κανένα συμπέρασμα, αλλά ο εκπαιδευόμενος εμφανίζει τον μετασχηματισμό της γνώσης με διάφορους τρόπους:

- Εκμάθηση μέσω προγραμματισμού: Όταν ένας αλγόριθμος ή ένας κώδικας είναι γραμμένος για την εκτέλεση συγκεκριμένης εργασίας. Π.χ. ένας κωδικός γράφεται ως παιχνίδι εικασίας για τον τύπο του ζώου. Ένα τέτοιο πρόγραμμα θα μπορούσε να τροποποιηθεί από εξωτερική οντότητα.
- Μάθηση μέσω απομνημόνευσης: Αυτό γίνεται με την απομνημόνευση δεδομένων γεγονότων ή δεδομένων χωρίς να προκύψουν συμπεράσματα από τις εισερχόμενες πληροφορίες ή δεδομένα. Μάθηση από παραδείγματα: Πρόκειται για μια ειδική περίπτωση επαγωγικής μάθησης. Δεδομένου ενός συνόλου παραδειγμάτων και αντιπαραδειγμάτων μιας έννοιας, ο εκπαιδευόμενος προκαλεί μια γενική περιγραφή έννοιας που περιγράφει όλα τα θετικά παραδείγματα και κανένα από τα αντιπαραδείγματα.
- Η μάθηση από παραδείγματα είναι μια μέθοδος που έχει διερευνηθεί σε μεγάλο βαθμό στον τομέα της τεχνητής νοημοσύνης. Ο όγκος των συμπερασμάτων που εκτελείται από τον μαθητή είναι πολύ μεγαλύτερος από ό,τι όταν μαθαίνει από οδηγίες, (Jaime G. Carbonell, 1983).
- Μάθηση από την παρατήρηση: Αυτή είναι μια προσέγγιση μάθησης χωρίς επίβλεψη και είναι μια πολύ γενική μορφή επαγωγικής μάθησης που περιλαμβάνει συστήματα ανακάλυψης, εργασίες σχηματισμού θεωρίας, δημιουργία κριτηρίων ταξινόμησης για τη διαμόρφωση ταξινομικών ιεραρχιών και παρόμοια εργασία που πρέπει να εκτελείται χωρίς το όφελος ενός εξωτερικού δασκάλου. Η μάθηση χωρίς επίβλεψη απαιτεί από τον μαθητή να κάνει περισσότερα συμπεράσματα από οποιαδήποτε προσέγγιση όπως εξηγήθηκε προηγουμένως. Δεν παρέχεται στον εκπαιδευόμενο ένα σύνολο δεδομένων ή παράδειγμα μιας συγκεκριμένης έννοιας. Η παραπάνω ταξινόμηση των στρατηγικών μάθησης θα πρέπει να βοηθήσει κάποιον να συγκρίνει διάφορα συστήματα μάθησης ως προς τους υποκείμενους μηχανισμούς τους, ως προς τη διαθέσιμη εξωτερική πηγή πληροφοριών και ως προς τον βαθμό στον οποίο ανταποκρίνονται στην προοργανωμένη γνώση.

Κεφάλαιο 4. Βιβλιογραφία εργασιών μηχανικής μάθησης στην απλοποίηση κειμένου

4.1. Μελέτες που συλλέχθηκαν

Η κατηγοριοποίηση των ερευνών με βάση το περιεχόμενο είναι ουσιώδης για την οργάνωση της επιστημονικής γνώσης και την παροχή σαφών κατευθυντήριων γραμμών για την περαιτέρω έρευνα. Οι επισκοπήσεις παρέχουν μια συνολική εικόνα της υπάρχουσας βιβλιογραφίας, τα πρωτογενή ερευνητικά άρθρα συμβάλλουν με νέα δεδομένα, οι μελέτες περίπτωσης προσφέρουν πρακτικές αναλύσεις συγκεκριμένων φαινομένων, τα θεωρητικά άρθρα προτείνουν νέες θεωρίες, και τα μεθοδολογικά άρθρα αναπτύσσουν νέες τεχνικές για τη συλλογή και ανάλυση δεδομένων. Κάθε μία από αυτές τις κατηγορίες έχει τη δική της μοναδική αξία και συνεισφέρει στην εξέλιξη της επιστημονικής έρευνας και γνώσης. Γι αυτό και η κατηγοριοποίηση των ερευνών με βάση το περιεχόμενο είναι ένας βασικός τρόπος οργάνωσης και παρουσίασης της επιστημονικής γνώσης. Διαφορετικοί τύποι ερευνητικών άρθρων έχουν διαφορετικούς σκοπούς και προσφέρουν μοναδικές συνεισφορές στην ακαδημαϊκή κοινότητα. Στην παρούσα εργασία θα χρησιμοποιηθούν πέντε κύριες κατηγορίες ερευνών: Επισκοπήσεις, Πρωτογενής Έρευνα, Μελέτες Περίπτωσης, Θεωρητικά Άρθρα και Μεθοδολογικά Άρθρα.

Επισκόπηση (Review Articles)

Τα επισκοπικά άρθρα παρουσιάζουν μια σύνοψη της υπάρχουσας βιβλιογραφίας σε ένα συγκεκριμένο πεδίο, αναλύοντας και σχολιάζοντας τα ευρήματα από διάφορες μελέτες. Αυτά τα άρθρα είναι ιδιαίτερα χρήσιμα για την κατανόηση της γενικής εικόνας ενός θέματος και για την αναγνώριση κενών στη βιβλιογραφία που χρειάζονται περαιτέρω διερεύνηση. Με αυτόν τον τρόπο, προσφέρει μια ολοκληρωμένη εικόνα της προόδου στον τομέα και των μελλοντικών προκλήσεων.

Πρωτογενής Έρευνα (Original Research Articles)

Τα πρωτογενή ερευνητικά άρθρα παρουσιάζουν νέα δεδομένα και ευρήματα από πρωτογενή έρευνα που διεξήχθη από τους συγγραφείς. Αυτά τα άρθρα συμβάλλουν άμεσα στην πρόοδο της γνώσης σε ένα πεδίο, παρέχοντας νέες πληροφορίες και ανακαλύψεις. Μέσω της παρουσίασης πρωτογενών δεδομένων, αυτό το άρθρο προωθεί την κατανόηση και την ανάπτυξη του πεδίου.

Μελέτες Περίπτωσης (Case Studies)

Οι μελέτες περίπτωσης παρουσιάζουν λεπτομερείς αναλύσεις συγκεκριμένων περιπτώσεων ή γεγονότων. Αυτού του είδους τα άρθρα είναι χρήσιμα για την κατανόηση πολύπλοκων φαινομένων σε πραγματικό περιβάλλον και συχνά προσφέρουν πρακτικές γνώσεις που μπορούν να εφαρμοστούν σε παρόμοιες καταστάσεις.

Θεωρητικά Άρθρα (Theoretical Articles)

Τα θεωρητικά άρθρα προτείνουν νέες θεωρίες ή μοντέλα και ενδέχεται να αναλύουν υπάρχουσες θεωρίες με νέα δεδομένα. Αυτά τα άρθρα είναι σημαντικά για την ανάπτυξη νέων τρόπων σκέψης και για την προώθηση της επιστημονικής γνώσης μέσω της θεμελίωσης νέων θεωρητικών πλαισίων.

Μεθοδολογικά Άρθρα (Methodological Articles)

Τα μεθοδολογικά άρθρα περιγράφουν και αξιολογούν νέες μεθόδους ή τεχνικές για τη συλλογή και την ανάλυση δεδομένων. Αυτά τα άρθρα είναι κρίσιμα για την πρόοδο της έρευνας, καθώς παρέχουν εργαλεία και μεθόδους που μπορούν να χρησιμοποιηθούν από άλλους ερευνητές για τη διεξαγωγή αξιόπιστων και έγκυρων μελετών.

Ο παρακάτω πίνακας, αποτυπώνει σχηματικά την κατηγορία και τις σημειώσεις των βασικότερων βιβλιογραφικών αναφορών που έχουν συλλεχθεί:

Αρθρογράφος	Τίτλος Άρθρου	Κατηγορία	Σημειώσεις
Aina Garí et al. (2018)	Συγκριτική μελέτη ενσωματώσεων λέξεων και άλλων χαρακτηριστικών για ανίχνευση λεξιλογικής πολυπλοκότητας στα γαλλικά	Original Research Article	Ανάλυση της πολυπλοκότητας γαλλικών λέξεων και συνδυασμών χαρακτηριστικών για την πρόβλεψη της πολυπλοκότητας. Χρησιμοποιήθηκαν πλατφόρμες Manulex και FLElex.
Yaoyuan Zhang et al.	A Constrained Sequence-to-Sequence Neural Model for Sentence Simplification	Original Research Article	Νέο μοντέλο περιορισμένης δημιουργίας νευρώνων για την απλοποίηση προτάσεων. Τα αποτελέσματα δείχνουν καλύτερη απόδοση από διάφορες γραμμές βάσης.
Alessia Battisti & Sarah Ebling	A Corpus for Automatic Readability Assessment and Text Simplification of German	Methodological Article	Δημιουργία ενός corpus για την αυτόματη αξιολόγηση αναγνωσιμότητας και απλοποίηση κειμένου στα γερμανικά. Περιέχει πληροφορίες για τη δομή του κειμένου.
Kristina Toutanova et al.	A Dataset and Evaluation Metrics for Abstractive Compression	Methodological Article	Εισαγωγή ενός συνόλου δεδομένων πολλαπλών αναφορών για τη συμπίεση αφηρημένων προτάσεων και

	of Sentences and Short Paragraphs		σύντομων παραγράφων. Χρησιμοποιήθηκαν δεδομένα από MASC και OANC.
Sanja Stajner & Hannah Bechara et al.	A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation	Original Research Article	Μελέτη της ποιότητας έναντι της ποσότητας των δεδομένων εκπαίδευσης στην απλοποίηση κειμένου. Τα αποτελέσματα δείχνουν ότι η ποιότητα των δεδομένων είναι σημαντική.

Αναλυτικά όλες οι βιβλιογραφικές αναφορές που έχουν συλλεχθεί:

1. Aina Garí, Μαριάννα Απιδιανάκη, Alexandre Allauzen. Συγκριτική μελέτη ενσωματώσεων λέξεων και άλλων χαρακτηριστικών για ανίχνευση λεξιλογικής πολυπλοκότητας στα γαλλικά. Conférence sur le Traitement Automatique des Langues Naturelles, Μάιος 2018, Rennes, Γαλλία. fahal-01838512ff

Σε αυτή τη μελέτη, οι συγγραφείς διενήργησαν πείραμα αναφορικά με την γραμματική αφομοίωση λέξεων αποσκοπώντας στη μέτρηση της πολυπλοκότητας των γαλλικών λέξεων. Συνδύασαν επίσης τις μετρήσεις αυτές με άλλα χαρακτηριστικά που έχουν αποδειχθεί ότι είναι κατάλληλα για την πρόβλεψη πολυπλοκότητας (text complexity identification). Τα αποτελέσματά της έρευνας κατέδειξαν ότι η αφομοίωση αποδίδει καλύτερα από άλλες λειτουργίες μεμονωμένα, αλλά δεν υπερέρχει των συστημάτων συχνότητας αυτής της γλώσσας.

Οι ερευνητές έκαναν χρήση δύο μεγάλων λεξιλογικών πόρων για τα γαλλικά, τις πλατφόρμες Manulex και FLElex.

Τα μοτίβα που παρατηρήθηκαν και οι βαθμολογίες που λήφθηκαν από τα συστήματα που εκπαιδεύτηκαν στο Manulex και στο FLElex είναι, γενικά, παρόμοια. Θα περίμενε κανείς ότι το

FLElex θα είχε ως αποτέλεσμα ένα σύστημα που θα παράγει καλύτερες βαθμολογίες λόγω της λεπτότερης ευαισθησίας του σε σχέση με το Manulex, αλλά το πολύ μεγαλύτερο μέγεθος του τελευταίου πιθανώς περιορίζει αυτό το αποτέλεσμα. Μεταξύ όλων των συνδυασμών χαρακτηριστικών, η συχνότητα σε συνδυασμό με τον αριθμό των χαρακτήρων ή με φωνήματα φαίνεται να είναι οι καλύτεροι προγνωστικοί παράγοντες πολυπλοκότητας τόσο για το Manulex όσο και για το FLElex. Ο αριθμός των φωνημάτων είναι το χαρακτηριστικό με τη χειρότερη απόδοση όταν χρησιμοποιείται μόνο του και λαμβάνει τη χαμηλότερη βαθμολογία σε όλα τα μέτρα.

Για τη διενέργεια του πειράματος, υλοποιήθηκε ένα απλό νευρωνικό δίκτυο τροφοδοσίας που έχει ως μαθησιακό στόχο έναν αριθμό που αντιστοιχεί στο πρώτο επίπεδο εμφάνισης μιας λέξης στον αντίστοιχο πόρο.

Οι ερευνητές κατέληξαν στο συμπέρασμα ότι το σύστημά που διαμόρφωσαν “μαθαίνει” μια βαθμολογία πολυπλοκότητας από τα επίπεδα πολυπλοκότητας που εισάγονται με βάση την ηλικία απόκτησης (Manulex) και την επάρκεια δεύτερης γλώσσας (FLElex). Η αξιολόγηση σε μια εργασία παραγωγής συνωνύμων φαίνεται να δείχνει ότι ενώ οι αφομοιώσεις λέξεων επιτυγχάνουν καλύτερα αποτελέσματα από άλλα χαρακτηριστικά μεμονωμένα, τα συστήματα που βασίζονται στη συχνότητα –ακόμη και μια απλή βασική συχνότητα– είναι καλύτερα και πιο κατάλληλα για αυτήν την εργασία. Τα συστήματα με την καλύτερη απόδοση χρησιμοποιούν συχνότητες ή τις συνδυάζουν με χαρακτήρες ή φωνήματα, δύο χαρακτηριστικά που έχουν αποδειχθεί χρήσιμα σε προηγούμενες εργασίες.

2. A Constrained Sequence-to-Sequence Neural Model for Sentence Simplification

Yaoyuan Zhang, Zhenxu Ye, Yansong Feng, Dongyan Zhao, Rui Yan

Οι ερευνητές εφάρμοσαν ένα νέο μοντέλο περιορισμένης δημιουργίας νευρώνων για να απλοποιήσουν προτάσεις, αποδίδοντας το νόημα τους με απλουστευμένες λέξεις. Τα τελικά αποτελέσματα στα ευθυγραμμισμένα σύνολα δεδομένων Wikipedia και Simple Wikipedia κατέδειξαν ότι η μέθοδός τους έχει αλύτερη απόδοση από διάφορες γραμμές βάσης.

Για το πείραμα, δημιουργήθηκαν απλοποιημένες προτάσεις χρησιμοποιώντας ένα μοντέλο αλληλουχίας (sequence to sequence model), εκπαιδευμένο σε μια παράλληλη βάση δεδομένων (corpus), δηλαδή την Αγγλική Βικιπαίδεια και την Απλή Αγγλική Βικιπαίδεια. Επίσης απλοποιήθηκαν οι λέξεις που αποτέλεσαν το δείγμα διαλογής χωρίς ωστόσο το τυπικό μοντέλο αλληλουχίας να μπορεί να εγγυηθεί την ύπαρξη τέτοιων λέξεων. Επομένως, προτάθηκε ένα μοντέλο περιορισμένης ακολουθίας σε ακολουθία (Constrained Seq2Seq) με τη δεδομένη απλοποιημένη λέξη(εις) ως περιορισμούς κατά τη δημιουργία προτάσεων.

Στη συνέχεια άλλαξαν λέξεις σύμφωνα με αυτή τη βάση δεδομένων και προχώρησαν στο μοντέλο δημιουργίας νευρωνικών προτάσεων, υποθέτοντας ότι οι αντικαταστάσεις λέξεων είναι σωστές με βάση προηγούμενες μελέτες για στοίχιση συνωνύμων. Πιο συγκεκριμένα, με δεδομένη μια πρόταση εισαγωγής, η διαδικασία απλοποίησης διεξήχθη σε δύο βήματα:

- Βήμα 1. Σύμφωνα με προηγούμενες μελέτες σχετικά με τη λεξιλογική υποκατάσταση, προσδιορίστηκαν πρώτα σύνθετες λέξεις στην εισαγωγική πρόταση και στη συνέχεια τις αντικαταστάθηκαν με τα απλούστερα συνώνυμά τους.
- Βήμα 2. Δεδομένων των απλουστευμένων λέξεων από το πρώτο βήμα ως περιορισμών, προτάθηκε έναν περιορισμένο μοντέλο seq2seq που κωδικοποιεί την πρόταση εισόδου ως διάνυσμα (κωδικοποιητής) και αποκωδικοποιεί το διάνυσμα σε απλοποιημένη πρόταση (αποκωδικοποιητής). Η διαδικασία δημιουργίας εξαρτάται από τις απλοποιημένες λέξεις και αποτελείται από την οπίσθια και τη μπροστινή παραγωγή.

Σημειωτέον ότι το πείραμα διεξήχθη με βάση τα σύνολα δεδομένων Αγγλικής Wikipedia και Simple English Wikipedia για να συγκριθούν οι προτεινόμενη μέθοδος με αρκετούς αντιπροσωπευτικούς αλγόριθμους, όπως:

Μωυσής. Είναι ένα τυπικό μοντέλο μηχανικής μετάφρασης που βασίζεται σε φράσεις (Koehn et al., 2007).

SBMT. Το SBMT είναι ένα συντακτικό μοντέλο μηχανικής μετάφρασης (Xu et al., 2016), το οποίο υλοποιείται στην εργαλειοθήκη ανοιχτού κώδικα Joshua (Post et al., 2013). Το μοντέλο απλοποίησης έχει βελτιστοποιηθεί στη μέτρηση SARI και αξιοποιεί το σύνολο δεδομένων PPDB (Pavlick and Callison-Burch, 2016) ως μια πλούσια πηγή πράξεων απλοποίησης.

Λεξική αντικατάσταση. Αυτή η μέθοδος αντικαθιστά μόνο τις σύνθετες λέξεις με τις απλοποιημένες λέξεις που χρησιμοποιήθηκαν ως περιοριστικές λέξεις στο μοντέλο.

Seq2Seq. Το μοντέλο αλληλουχίας σε ακολουθία είναι το υπερσύγχρονο μοντέλο νευρωνικής μηχανικής μετάφρασης (Cho et al., 2014) με τον μηχανισμό προσοχής που εφαρμόζεται (Bahdanau et al., 2014).

Περιορισμένη Seq2Seq. Εδώ προτάθηκε ένα νέο μοντέλο δημιουργίας νευρικών προτάσεων που βασίζεται στο παράδειγμα αλληλουχίας σε ακολουθία με μία λέξη περιορισμού. Χρησιμοποιήθηκε το Multi-Constrained Seq2Seq για να δηλωθεί το σενάριο όταν υπάρχουν περισσότερες από μία λέξεις περιορισμού.

Συμπερασματικά, διατυπώνεται η πρόταση χρήσης μιας νέας μεθόδου δύο βημάτων για την απλοποίηση προτάσεων συνδυάζοντας απλοποίηση σε επίπεδο λέξης και απλοποίηση σε επίπεδο πρότασης. Εκτελέστηκαν πειράματα στα παράλληλα σύνολα δεδομένων της Wikipedia και της απλής Wikipedia και τα αποτελέσματα δείχνουν ότι οι μέθοδοι που χρησιμοποιήθηκαν ξεπερνούν τις διάφορες βασικές γραμμές με καλύτερη αναγνωσιμότητα, ευελιξία και απλότητα.

3. A Corpus for Automatic Readability Assessment and Text Simplification of German, Alessia Battisti, Sarah Ebling

Στην εν λόγω έρευνα, παρουσιάστηκε ένα corpus για χρήση στην αυτόματη αξιολόγηση αναγνωσιμότητας και αυτόματη απλοποίηση κειμένου στα γερμανικά. Το σώμα συντάσσεται από πηγές Ιστού και αποτελείται από περίπου 211.000 προτάσεις. Ως νέα συνεισφορά, περιέχει πληροφορίες για τη δομή του κειμένου, την τυπογραφία και τις εικόνες, οι οποίες μπορούν να

αξιοποιηθούν ως μέρος των προσεγγίσεων μηχανικής μάθησης για την αξιολόγηση της αναγνωσιμότητας και την απλοποίηση του κειμένου.

Το σώμα που προκύπτει περιέχει 6.217 έγγραφα (5.461 μονόγλωσσα έγγραφα συν 378 έγγραφα για κάθε πλευρά των παράλληλων δεδομένων). Τα έγγραφα που είναι μόνο μονόγλωσσα περιέχουν κατά μέσο όρο λιγότερες προτάσεις από την απλοποιημένη γερμανική των παράλληλων δεδομένων (μέσο μήκος εγγράφου σε προτάσεις 31,64 έναντι 55,75). Η μέση διάρκεια της πρότασης είναι σχεδόν ίση και στα δύο σύνολα. Ως εκ τούτου, τα μονόγλωσσα κείμενα είναι πιο σύντομα από τα απλοποιημένα γερμανικά κείμενα στα παράλληλα δεδομένα. Σε σύγκριση με τα γερμανικά αντίστοιχα, τα απλουστευμένα γερμανικά κείμενα στα παράλληλα δεδομένα έχουν σαφώς υποστεί μια διαδικασία λεξιλογικής απλοποίησης: Το λεξιλόγιο είναι μικρότερο κατά 51% (33.384 έναντι 16.352 τύπων), το οποίο είναι συγκρίσιμο με το ποσοστό μείωσης που αναφέρεται για το Newsela Corpus (50,8%).

Εν κατακλείδει, οι ερευνητές κατήρτισαν ένα corpus δεδομένων που έχει καταρτιστεί για χρήση στην αυτόματη αξιολόγηση αναγνωσιμότητας και στην αυτόματη απλοποίηση κειμένου στα γερμανικά. Ενώ τέτοιες εργασίες έχουν αντιμετωπιστεί για άλλες γλώσσες, η έρευνα για τα γερμανικά είναι σπάνια. Τα χαρακτηριστικά που αξιοποιούνται ως μέρος των προσεγγίσεων μηχανικής μάθησης για την αξιολόγηση αναγνωσιμότητας μέχρι στιγμής περιλαμβάνουν τυπικά επιφανειακά ή/και (βαθύτερα) γλωσσικά χαρακτηριστικά. Το σώμα που παρουσιάζεται σε αυτό το άρθρο περιέχει επιπλέον πληροφορίες για τη δομή του κειμένου, την τυπογραφία και τις εικόνες. Αυτά τα χαρακτηριστικά έχουν αποδειχθεί ότι είναι ενδεικτικά απλών έναντι πολύπλοκων κειμένων τόσο θεωρητικά όσο και, χρησιμοποιώντας το σώμα που περιγράφεται σε αυτή την εργασία, εμπειρικά. Οι πληροφορίες σχετικά με τη δομή του κειμένου, την τυπογραφία και τις εικόνες μπορούν επίσης να αξιοποιηθούν ως μέρος μιας προσέγγισης νευρωνικής μηχανικής μετάφρασης για την απλοποίηση του κειμένου. Ένα σύνολο παράλληλων εγγράφων που χρησιμοποιούνται στην αυτόματη μετάφραση απαιτεί επιπλέον στοίχιση προτάσεων, οι οποίες έλειπαν από το σώμα.

4.A Dataset and Evaluation Metrics for Abstractive Compression of Sentences and Short Paragraphs, Kristina Toutanova, Chris Brockett, Ke M. Tran, Saleema Amershi

Οι ερευνητές στην εν λόγω έρευνα, εισήχθη ένα σύνολο δεδομένων πολλαπλών αναφορών που δημιουργήθηκε με μη αυτόματο τρόπο προς το σκοπό συμπίεσης μιας αφηρημένης πρότασης ή και σύντομης παραγράφου. Χρησιμοποιήθηκαν μεμονωμένες προτάσεις ή και παράγραφοι δύο προτάσεων από διάφορα είδη στην ενότητα γραπτού κειμένου του Manually Annotated Sub-Corpus (MASC) του Open American National Corpus (OANC), συμπληρωμένα με πρόσθετα δεδομένα από τη γραπτή ενότητα του OANC. Οι παράγραφοι με δύο προτάσεις αντιπροσωπεύουν περίπου το 23% των παραγράφων με περισσότερες από μία προτάσεις στο OANC. Οι παράγραφοι δύο προτάσεων που δειγματίστηκαν περιείχαν τουλάχιστον 25 λέξεις. Οι συμπίεσεις δημιουργήθηκαν χρησιμοποιώντας το UHRS, ένα εσωτερικό σύστημα crowd-sourcing παρόμοιο με το Mechanical Turk της Amazon, σε δύο κύκλους σχολιασμού, έναν για σμίκρυνση και έναν δεύτερο για την αξιολόγηση της ποιότητας συμπίεσης.

Η πρόοδος στην αυτοματοποιημένη συμπίεση κειμένου μετριέται τυπικά με σύγκριση των εξόδων του μοντέλου σε επίπεδο σώματος. Ωστόσο, για την εκπαίδευση μοντέλων με διακριτικό τρόπο και για την εκτέλεση λεπτομερών συγκρίσεων συστημάτων, είναι επίσης απαραίτητο να υπάρχει αξιολόγηση των εξόδων του συστήματος σε επίπεδο μεμονωμένων εισροών.

Τέθηκαν σε εφαρμογή και αξιολογήθηκαν τέσσερα συστήματα συμπίεσης. Αυτά περιλαμβάνουν τόσο βασισμένα σε διαγραφή όσο και αφηρημένα μοντέλα: (1) ILP, μια προσέγγιση γραμμικού προγραμματισμού ακέραιου αριθμού για συμπίεση με βάση τη διαγραφή (Clarke and Lapata, 2008), (2) T3, ένα μοντέλο βασισμένο σε δενδρικούς μετατροπείς για την αφηρημένη συμπίεση (Cohn και Lapata, 2008), (3) Seq2seq, ένα μοντέλο νευρωνικών δικτύων για συμπίεση βάσει διαγραφής (Filippova et al., 2015) και (4) NAMAS, ένα νευρωνικό μοντέλο για αφηρημένη συμπίεση και σύνοψη (Rush et al., 2015).

Οι ερευνητές εισήγαγαν ένα μεγάλο αφηρημένο σύνολο δεδομένων πολλαπλών αναφορών που συλλέγεται με μη αυτόματο τρόπο και ποσοτικοποίησαν τον αντίκτυπο των λειτουργιών επεξεργασίας και του περιβάλλοντος στην ποιότητα της ανθρώπινης συμπίεσης, δείχνοντας ότι οι λειτουργίες αντικατάστασης και αναδιατύπωσης διατηρούν περισσότερο το νόημα από άλλες λειτουργίες και ότι η συμπίεση βελτιώνει την ποιότητα. Επιπλέον, στην πρώτη συστηματική

μελέτη των μετρήσεων αυτόματης αξιολόγησης για τη συμπίεση κειμένου, καταδεικνύεται η σημασία της χρήσης πολλαπλών αναφορών και κατάλληλων γλωσσικών ενοτήτων και την ενσωμάτωση της ανάκλησης.

5. A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation, Sanja Stajner and Hannah Bechara and Horacio Saggion

Στην εν λόγω έρευνα μελετήθηκε η επίδραση της ποιότητας έναντι της ποσότητας των δεδομένων εκπαίδευσης στην αποτελεσματικότητα μιας προσέγγισης MT στην απλοποίηση κειμένου. Διεξήχθησαν 40 πειράματα με πρώτη βάση προτάσεις από την Αγγλική Wikipedia και την Απλή Αγγλική Wikipedia, οι οποίες ελέγχθηκαν ως προς : (1) την ομοιότητα μεταξύ των αρχικών και απλοποιημένων προτάσεων στα σύνολα δεδομένων εκπαίδευσης και ανάπτυξης και (2) τα μεγέθη αυτών των συνόλων δεδομένων. Τα αποτελέσματα υποδηλώνουν ότι στην τυπική προσέγγιση PB-SMT για την απλοποίηση κειμένου, η ποιότητα των συνόλων δεδομένων έχει μεγαλύτερο αντίκτυπο στην απόδοση του συστήματος.

Επιπλέον, επισημανθηκαν αρκετές σημαντικές διαφορές μεταξύ της διαγλωσσικής MT και της μονόγλωσσας MT που χρησιμοποιείται στην απλοποίηση κειμένου ενώ η έρευνα απέδειξε ότι το BLEU δεν είναι καλό μέτρο της απόδοσης του συστήματος στην εργασία απλοποίησης κειμένου.

Προκειμένου να διερευνηθούν λοιπόν τα δύο πρώτα ζητήματα, διεξήχθησαν 40 πειράματα μετάφρασης χρησιμοποιώντας τα στοιχισμένα ζεύγη προτάσεων από τη μεγαλύτερη υπάρχουσα βάση δεδομένων (corpus) TS (Wikipedia TS corpus), ελέγχοντας τα σύνολα δεδομένων εκπαίδευσης και ανάπτυξης για: (1) ομοιότητα προτάσεων (από την άποψη του S -Βαθμολογία BLEU), και (2) μέγεθος. Τα αποτελέσματά δείχνουν ότι μόνο το πρώτο μπορεί να επηρεάσει σημαντικά την έξοδο MT. Προκειμένου να διερευνηθεί το τελευταίο ζήτημα, εξετάστηκαν τα μοντέλα σε δύο διαφορετικά σύνολα δοκιμών ενώ πραγματοποιήθηκε και αξιολόγηση της παραγωγής πολλών συστημάτων από εξειδικευμένους φιλόλογους.

Τα πειράματά που διεξήχθησαν σε 40 διαφορετικά, προσεκτικά σχεδιασμένα σύνολα δεδομένων από το μεγαλύτερο διαθέσιμο σώμα TS με ευθυγράμμιση προτάσεων (Wikipedia TS corpus),

παρέχουν πολύτιμες πληροφορίες σχετικά με το πόσο μεγάλη επίδραση έχει το μέγεθος και η ποιότητα των δεδομένων εκπαίδευσης στην απόδοση του Συστήματος PB-SMT που προσπαθεί να μάθει να μεταφράζει από «πρωτότυπες- σύνθετες» σε «απλές» προτάσεις. Τα αποτελέσματα υποδεικνύουν ότι η χρήση των ζευγών προτάσεων με χαμηλές βαθμολογίες S-BLEU για εκπαίδευση και συντονισμό μοντέλων PB-SMT για TS τείνουν να προκαλούν την επιδείνωση της ευχέρειας ορθής αποτύπωσης και ακόμη και να αλλάζουν το νόημα της εξόδου. Επιπλέον, φαίνεται ότι τα μεγέθη των συνόλων δεδομένων εκπαίδευσης και ανάπτυξης δεν παίζουν σημαντικό ρόλο στο πόσο επιτυχημένο είναι το μοντέλο. Φαίνεται ότι τα προσεκτικά επιλεγμένα ζεύγη προτάσεων στα σύνολα δεδομένων εκπαίδευσης και ανάπτυξης (δηλαδή ζεύγη προτάσεων με μέτρια ομοιότητα) οδηγούν στις καλύτερες επιδόσεις των συστημάτων PB-SMT ανεξάρτητα από το μέγεθος των συνόλων δεδομένων.

6. Can Text Simplification Help Machine Translation?

Sanja STAJNER and Maja POPOVIC

Αυτό το άρθρο διερευνά τη χρήση της απλοποίησης κειμένου ως βήμα προεπεξεργασίας για στατιστική αυτόματη μετάφραση γραμματικά πολύπλοκων γλωσσών που δεν διαθέτουν βασείς δεδομένων. Τα πειράματά που διεξήχθησαν σχετικά με τη μετάφραση από τα Αγγλικά στα Σερβικά δείχνουν ότι αυτή η προσέγγιση μπορεί να βελτιώσει την ορθή γραμματική αποτύπωση της παραγωγής μετάφρασης και να μειώσει την τεχνική προσπάθεια μετά την επεξεργασία (αριθμός λειτουργιών μετά την επεξεργασία).

Επιπλέον, η χρήση πιο επιθετικών μεθόδων απλοποίησης κειμένου (οι οποίες όχι μόνο απλοποιούν τη δεδομένη πρόταση αλλά και απορρίπτουν άσχετες πληροφορίες παράγοντας συντακτικά πολύ απλές προτάσεις) βελτιώνει επίσης τη διατήρηση του νοήματος (επάρκεια) της μεταφραστικής παραγωγής.

Στην έρευνα αυτή διευρύνθηκε εάν είναι δυνατό να βελτιωθεί η απόδοση της αυτόματης μετάφρασης για γλώσσες που δεν διαθέτουν πόρους, εισάγοντας ένα βήμα προεπεξεργασίας στο οποίο οι προτάσεις-πηγής απλοποιούνται πρώτα από ένα σύστημα αυτόματης απλοποίησης

κειμένου (ATS). Οι ερευνητές εστίασαν σε Αγγλικά-Σερβικά MT και εφάρμοσαν δύο υπερσύγχρονα συστήματα ATS ως βήμα προεπεξεργασίας για την απλοποίηση της αρχικής αγγλικής πρότασης πριν την τροφοδοτήσουν σε ένα σύστημα SMT που βασίζεται σε φράσεις. Εκμεταλλεύτηκαν επίσης δύο διαφορετικούς τύπους συστημάτων ATS, ένα πιο συντηρητικό (το οποίο, ενώ απλοποιεί την πρόταση εισόδου λεξιλογικά και συντακτικά, διατηρεί όλες τις πληροφορίες που περιέχονται στην αρχική πρόταση) και το πιο επιθετικό (που, ενώ απλοποιεί την πρόταση εισόδου λεξικά και συντακτικά, προσπαθεί επίσης να μειώσει τον όγκο των πληροφοριών απορρίπτοντας άσχετες πληροφορίες και λεπτομέρειες υψηλού επιπέδου). Με αυτόν τον τρόπο, καταγραφηκαν δύο διαφορετικά σενάρια χρήσης στο MT: (1) όταν είναι σημαντικό να διατηρηθούν όλες οι πληροφορίες που περιέχονται στο κείμενο προέλευσης (π.χ. μεταφράσεις ολόκληρων κειμένων ή εγγράφων) και (2) όταν αρκεί να εξαχθεί μια συγκεκριμένη μόνο ουσιαστική πληροφορία του κειμένου πηγής (π.χ. όπως εαν περιηγηθεί κάποιος σε άρθρα ειδήσεων και αναζητήσει τις πιο σημαντικές ειδήσεις).

Τα αποτελέσματα της ανθρώπινης αξιολόγησης των άρθρων που μεταφράστηκαν χρησιμοποιώντας τις δύο προαναφερθείσες προσεγγίσεις, όσον αφορά τη γραμματικότητα (ευρύτητα) και τη διατήρηση του νοήματος (ακρίβεια) του προϊόντος, και την ανάλυση της προσπάθειας μετά την επεξεργασία (αριθμός μετά- λειτουργίες επεξεργασίας) δείχνει ότι και οι δύο προσεγγίσεις βελτιώνουν την έξοδο MT.

Στην εν λόγω έρευνα χρησιμοποιήθηκαν οι κάτωθι αλγόριθμοι:

TS-A: Ένας συνδυασμός λεξιλογικού συστήματος TS με το EventSimplify που εκτελεί συντακτική απλοποίηση με σημαντική μείωση περιεχομένου. Αυτό είναι το πιο «επιθετικό» σύστημα από όλα τα προαναφερθέντα συστήματα που εκτελούν μείωση περιεχομένου, δηλαδή είναι το σύστημα που εκτελεί το υψηλότερο επίπεδο μείωσης περιεχομένου και επιτυγχάνει την πιο ευανάγνωστη (απλή) έξοδο (λόγω του μεγάλου αριθμού λειτουργιών διαχωρισμού προτάσεων).

TS-C: Το λεξικοσυντακτικό σύστημα TS ανήκει στα «συντηρητικά» συστήματα ATS που δεν εκτελούν καμία μείωση περιεχομένου και επομένως, διατηρούν πλήρως το αρχικό νόημα της πρότασης.

Χρησιμοποιήθηκαν 100 άρθρα ειδήσεων από το EMM NewsBrief3 για τα οποία η έξοδος του συστήματος ATS EventSimplify ήταν δωρεάν. Οι ερευνητές εστίασαν περαιτέρω στην έξοδο του σχήματος απλοποίησης βάσει συμβάντων (το οποίο πέτυχε την υψηλότερη αναγνωσιμότητα και των τεσσάρων παρεχόμενων σχημάτων) και εφάρμοσαν το σύστημα λεξικής απλοποίησης πάνω από αυτό, προκειμένου να αποκτήσουν ένα πλήρες σύστημα απλοποίησης που περιλαμβάνει τη λεξιλογική απλοποίηση και τη συντακτική απλοποίηση και μείωση περιεχομένου (TS-A). Στη συνέχεια, εφαρμόσαμε το σύστημα TS-C σε όλα αυτά τα 100 πρωτότυπα άρθρα.

Περαιτέρω διερευνηθεί εάν τα υπερσύγχρονα συστήματα αυτόματης απλοποίησης κειμένου (ATS) μπορούν να βελτιώσουν την αυτόματη μετάφραση από τα Αγγλικά στα Σερβικά (MT) εάν χρησιμοποιηθούν ως βήμα προεπεξεργασίας για την απλοποίηση των προτάσεων πηγής πριν από τη μετάφραση τους με το σύστημα SMT. Τέθηκε σε εφαρμογή και αυτή η μέθοδος για αυτήν την υπόθεση χρησιμοποιώντας δύο συστήματα ATS, ένα πιο «συντηρητικό» (TS-C) που εκτελεί μόνο λεξιλογικές και συντακτικές απλοποιήσεις και ένα πιο «επιθετικό» (TS-A) που εκτελεί περισσότερες λεξιλογικές και συντακτικές αλλαγές αλλά και εκτελεί σημαντική μείωση περιεχομένου, οδηγώντας έτσι σε απώλεια ορισμένων λεπτομερειών πληροφοριών. Όλα τα παρουσιαζόμενα αποτελέσματα υποδεικνύουν ότι η χρήση του TS-C μπορεί να βελτιώσει την ευχέρεια της εξόδου MT και να μειώσει την τεχνική και γνωστική προσπάθεια μετά την επεξεργασία μέσω της μείωσης των σφαλμάτων αναδιάταξης. Η χρήση του TS-A εισάγει ακόμη περισσότερες βελτιώσεις για την επάρκεια, την ευχέρεια και όλους τους τύπους λειτουργιών επεξεργασίας, αλλά με κόστος απώλειας ορισμένων λεπτομερειών στις πληροφορίες. Αυτή η προσέγγιση, ωστόσο, θα μπορούσε να είναι πολύ χρήσιμη για εργασίες όπου το κύριο νόημα του κειμένου είναι κρίσιμο και η απώλεια ορισμένων λεπτομερειών είναι οικονομικά προσιτή.

Επιπλέον, τα αποτελέσματα έδειξαν ότι η χρήση ενός συστήματος TS ως βήμα προεπεξεργασίας σε έναν αγωγό MT είναι χρήσιμη μόνο για ένα υποσύνολο προτάσεων, ενώ οι υπόλοιπες προτάσεις είτε φθίνουν ποιοτικά είτε παραμένουν αμετάβλητες. Επομένως, μια μέθοδος για το φιλτράρισμα των προτάσεων σε δύο ή τρεις κατηγορίες (το TS βελτιώνει/το TS χειροτερεύει ή το TS βελτιώνει/το TS δεν επηρεάζει/το TS χειροτερεύει) θα ήταν πολύ χρήσιμο και θα πρέπει να διερευνηθεί στη μελλοντική εργασία.

7. CombiNMT: An Exploration into Neural Text Simplification Models

Michael Cooper, Matthew Shardlow

Αυτή η εργασία παρουσιάζει μια μελέτη αναπαραγωγής των μοντέλων απλοποίησης νευρωνικού κειμένου. Παράλληλα με τα αποτελέσματα αναπαραγωγής, παρουσίασαν τις βελτιώσεις που ονομάζονται CombiNMT. Αυτά επιτεύχθηκαν χρησιμοποιώντας μια ενημερωμένη υλοποίηση του OpenNMT και ενσωματώνοντας το σώμα Newsela μαζί με το αρχικό σύνολο δεδομένων της Wikipedia, καθώς και βελτιώνοντας και τα δύο σύνολα δεδομένων για την επιλογή παραδειγμάτων εκπαίδευσης υψηλής ποιότητας. Η εργασία αυτή παρουσιάζει δύο νέα συστήματα, το CombiNMT995, το οποίο είναι αποτέλεσμα αντιστοιχισμένων προτάσεων με ομοιότητα συνημιτόνου 0,995 ή μικρότερη, και το CombiNMT98, το οποίο, ομοίως, εκτελείται με ομοιότητα συνημιτόνου 0,98 ή μικρότερη. Χρησιμοποιώντας και την ανθρώπινη αξιολόγηση από εξειδικευμένο προσωπικό, αυξάνοντας τόσο τον αριθμό των σχολιαστών όσο και τον αριθμό των προτάσεων που σχολιάζονται, με σκοπό την αύξηση της ποιότητας των αποτελεσμάτων, το CombiNMT998 παρουσιάζει σημαντική βελτίωση σε σχέση με οποιοδήποτε από τα συστήματα Απλοποίησης Νευρωνικού Κειμένου (NTS).

Σε αυτήν τη μελέτη χρησιμοποιήθηκε το σύνολο δεδομένων από την αρχική μελέτη, μια επεξεργασμένη έκδοση του δημοσίως διαθέσιμου συνόλου δεδομένων Hwang που περιλαμβάνει ευθυγραμμίσεις μεταξύ της τυπικής Αγγλικής Wikipedia και της Απλής Αγγλικής Wikipedia, καθώς και του Newsela Corpus, που αποτελείται από 1,9 χιλιάδες αγγλικά και απλουστευμένα αγγλικά άρθρα ειδήσεων, που έχει αποδειχθεί χρήσιμο για απλοποίηση. Χρησιμοποιήθηκαν τυπικά αγγλικά άρθρα και αυτά τα άρθρα βαθμολογήθηκαν με την απλότητα 3 σε μια κλίμακα 1-5. Αυτό ήταν το τελευταίο επίπεδο βαθμολόγησης που περιελάμβανε και τα 1,9 χιλιάδες άρθρα στο τυπικό αγγλικό σύνολο.

Τα αποτελέσματα από το προεπιλεγμένο σύστημα NTS που παρουσιάστηκε από σχολιάστηκαν επίσης, για να καταδειχθεί πώς συγκρίνεται η ομάδα των σχολιαστών με αυτά που

χρησιμοποιούνται από την αρχική ομάδα. Σε αντίθεση με το αρχικό σύστημα, χρησιμοποιήθηκαν μόνο αγγλόφωνοι εθελοντές σχολιαστές, οι οποίοι κατέχουν τουλάχιστον πτυχίο Bachelor, για να διασφαλιστεί ένα καλό επίπεδο γραπτής κατανόησης της αγγλικής γλώσσας. Παρουσιάστηκαν στους σχολιαστές οι πρώτες 120 προτάσεις από το δοκιμαστικό σύνολο, μαζί με την έξοδο από κάθε σύστημα που εκτελεί τις ίδιες προτάσεις.

Το CombiNMT995 εκπαιδεύτηκε χρησιμοποιώντας τη διαμόρφωση που περιγράφεται στο (Nisioi et al., 2017) και νωρίτερα σε αυτό το έγγραφο, με το συνδυασμένο σύνολο δεδομένων μετά την αφαίρεση οποιουδήποτε ζεύγους προτάσεων με ομοιότητα συνημιτόνου μικρότερη από 0,995, με το σύνολο ανάπτυξης που χρησιμοποιείται στο (Nisioi et al., 2017). CombiNMT98 το οποίο εκπαιδεύτηκε με παρόμοιο τρόπο, εκτός από τιμή ομοιότητας συνημιτόνου μικρότερη από 0,98.

Χρησιμοποιήθηκε η περιγραφόμενη διαμόρφωση για την Απλοποίηση Νευρωνικού Κειμένου (NTS) και η Απλοποίηση Νευρωνικού Κειμένου με συστήματα αναπαραγωγής ενσωματώσεων word2vec (NTSw2v). Αυτά τα συστήματα εκπαιδεύτηκαν χρησιμοποιώντας το ίδιο σύνολο δεδομένων.

Το NTSw2v απέδωσε συγκριτικά με το σύστημα CombiNMT995, το οποίο, λαμβάνοντας υπόψη ότι το αρχικό NTSw2v είχε τα πιο εντυπωσιακά αποτελέσματα στην αρχική εργασία, δείχνει ότι ένα σύνολο δεδομένων αυξημένης ποιότητας μπορεί να μειώσει το απαραίτητο μέγεθος του απαιτούμενου συνόλου δεδομένων. Η χρήση του συνόλου δεδομένων σε ομοιότητα συνημιτόνου κάτω του 0,995 έδωσε στο σύνολο δεδομένων το μέγεθος που χρειαζόταν για να μειωθεί η πιθανότητα υπερπροσαρμογής. Είναι δύσκολο να προταθεί η συγκεκριμένη επίδραση του Newsela στο σύστημα, και μάλιστα να ποσοτικοποιηθεί κάθε πρόθεση που παρουσιάζεται, ωστόσο τα φιλτραρισμένα δεδομένα της Wikipedia από μόνα τους δεν θα είχαν το μέγεθος για να δημιουργήσουν ένα σύστημα εφαρμόσιμο σε γενικευμένο κείμενο. Η αρχική μελέτη συντόνισε το μοντέλο NTS για να βρει το καλύτερο μέγεθος δέσμης και τον καλύτερο αριθμό υπόθεσης. Αυτό θα μπορούσε να μειώσει την πιθανότητα υπερβολικής τοποθέτησης στα μοντέλα τους, λόγω του γεγονότος ότι η αναζήτηση δέσμης εξετάζει πολλαπλές εναλλακτικές παράλληλα και μειώνει τον αριθμό των κακών εναλλακτικών και μειώνει το μέγεθος της δέσμης της, ενώ αυξάνει την ποιότητα της πρόβλεψης.

Τα αποτελέσματα δείχνουν μια βελτίωση στις βαθμολογίες ανθρώπινης κρίσης, καθώς και στη βαθμολογία SARI σε σχέση με την αρχική βασική γραμμή NTS.

8. Creating Data in Icelandic for Text Normalization

Helga Svala Sigurðardóttir, Anna Björk Nikulásdóttir, Jón Guðnason

Οι ερευνητές εδώ παρουσίασαν το Regína, ένα σύστημα που βασίζεται σε κανόνες που μπορεί να ταξινομεί αυτόματα τα δεδομένα για ένα σύστημα μετατροπής κειμένου σε ομιλία (TTS). Κανονικά δεδομένα δεν υπάρχουν γενικά, επομένως δημιουργήθηκαν αρκετά καλά δεδομένα για πιο προηγμένες μεθόδους τυποποίησης κειμένου (TN). Σημειώθηκε επίσης με μη αυτόματο τρόπο το πρώτο κανονικοποιημένο σώμα στα ισλανδικά, 40.000 προτάσεις, και αναπτύχθηκε το Regína, ένα σύστημα TN που βασίζεται σε κανονικές εκφράσεις. Το νέο σύστημα έχει ακρίβεια 89,82% σε σύγκριση με το χειροκίνητο σχολιασμένο σώμα σε μη τυπικές λέξεις και έδειξε σημαντική βελτίωση στην ακρίβεια σε σύγκριση με ένα παλαιότερο σύστημα κανονικοποίησης για τα Ισλανδικά. Το κανονικοποιημένο σώμα και το Regína θα κυκλοφορήσουν ως συστήματα ανοιχτού κώδικα.

Τα δεδομένα που χρησιμοποιήθηκαν είναι 40.000 προτάσεις (741.909 λέξεις) από την έκδοση του 2017 του Ισλανδικού Gigaword Corpus (IGC). Χρησιμοποιήθηκαν προτάσεις που περιλαμβάνουν πολλές NSW, όπως αριθμούς, συντμήσεις και σύμβολα. Οι προτάσεις σχολιάστηκαν χειροκίνητα και αποτελούν το πρώτο σώμα κανονικοποίησης για τα Ισλανδικά. Για ένα μικρό πείραμα σχετικά με τη συμφωνία μεταξύ των σχολιαστών, τρία άτομα από το Πανεπιστήμιο του Ρέικιαβικ κανονικοποίησαν 30 προτάσεις με 205 NSW, χρησιμοποιώντας τις οδηγίες των ερευνητών. Οι σχολιαστές επέκτειναν τις λέξεις χωρίς να λαμβάνουν υπόψη μια σημειωτική τάξη. Η συμφωνία μεταξύ των σχολιαστών για τις NSW ήταν $\kappa = 0,85$.

Για να συγκριθεί το Regína με τον παλιό πρόγραμμα τυποποίησης Thrax, το Texta Haukur τροποποιήθηκαν 400 προτάσεις από ολόκληρο το σύνολο δεδομένων και με τα δύο συστήματα. 147 από αυτά περιείχαν NSW και παρατηρήθηκαν για πιο ουσιαστικά αποτελέσματα. Η Regína

είχε βαθμολογία ακρίβειας 83,67%, με 20 προτάσεις που περιέχουν 22 λέξεις που δεν ταίριαζαν με τον χειροκίνητο σχολιασμό. Το Textahaukur είχε βαθμολογία ακρίβειας 61,22%, με 55 προτάσεις να περιέχουν 106 εσφαλμένα τυποποιημένες λέξεις.

Εξετάστηκαν τα λάθη που έκαναν και το Regína και το Textahaukur. Η Regína είχε κάποιες συντομογραφίες που δεν επεκτάθηκαν λόγω πιθανής ασάφειας. Διαφορετικά, η πλειονότητα των σφαλμάτων αφορούσε λάθος αποτύπωση διευρυμένου αριθμού. Αυτά ήταν επίσης τα πιο συνηθισμένα λάθη για το Textahaukur. Τα πιο σοβαρά σφάλματα ήταν μια ισχυρή τάση αλλαγής υποθέσεων στη μέση ενός διακριτικού. Τα χειροκίνητα σχολιασμένα δεδομένα έγιναν αναπόφευκτα ένα σύνολο δεδομένων ανάπτυξης, καθώς ήταν πάντα ορατά για τον προγραμματιστή του Regína. Ωστόσο, αυτό είναι αποκλειστικά ένα πρόβλημα για τη σύγκριση του συστήματος με το σώμα. Το Regína θα χρησιμοποιηθεί για την τυποποίηση κειμένου για σύνθεση TTS. Αν και η ακριβής επέκταση μπορεί να διαφέρει από άτομο σε άτομο, αυτό δεν υποδηλώνει λανθασμένη τυποποίηση.

9. Identification of complex words and passages in medical documents in French

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, Horacio Saggion

Στην εν λόγω έρευνα χρησιμοποιήθηκαν 100 γαλλικές λέξεις κλινικής ορολογίας που επιλέχθηκαν τυχαία από το σώμα CAS (Grabar et al., 2018). Αυτό το σώμα περιέχει συνολικά 41.384 λέξεις. Οι κλινικές περιπτώσεις είναι ιατρικά έγγραφα παρόμοια με κλινικές αναφορές. Περιγράφουν το ιατρικό υπόβαθρο των ασθενών, τον λόγο της συμβουλευτικής τους, τη διαδικασία υγειονομικής περίθαλψης και τις θεραπείες που προτείνονται και πραγματοποιούνται, καθώς και το αποτέλεσμα. Τέτοια κλινικά έγγραφα μπορούν να συναντήσουν οι ασθενείς στην καθημερινή τους ζωή, γεγονός που παρακινεί τη μελέτη τους και τη χρήση τους στην τρέχουσα εργασία.

Τα έγγραφα αναλύονται συντακτικά από τον αναλυτή Cordial (Laurent et al., 2009) για να τα χωρίσει σε συντακτικές ομάδες (κομμάτια). Το σώμα περιέχει συνολικά 15.053 κομμάτια. Στη συνέχεια, τα έγγραφα σχολιάζονται χειροκίνητα από εννέα σχολιαστές, προκειμένου να επισημανθούν τα κομμάτια με δυσκολία κατανόησης. Οι σχολιαστές είναι όλοι φυσικοί ομιλητές

της Γαλλικής και δεν έχουν ιατρικές γνώσεις ή εκπαίδευση. Λίγοι από αυτούς είναι χρόνια πάσχοντες ώστε να έχουν μια εξοικείωση με την ορολογία. Κατά τη διαδικασία σχολιασμού, οι σχολιαστές συμβουλευτήκαν να μην χρησιμοποιούν λεξικά ή πληροφορίες που είναι διαθέσιμες στο Διαδίκτυο. Έπρεπε να κάνουν τους σχολιασμούς με βάση τις δικές τους γνώσεις. Οι σχολιαστές ανάγνωσαν ολόκληρα έγγραφα, όπου τα κομμάτια βρίσκονται ανάμεσα σε αγκύλες. Για κάθε τμήμα, οι σχολιαστές πρέπει να υποδείξουν εάν δεν μπορούν να το καταλάβουν (με κόκκινο) ή αν δεν είναι βέβαιοι ότι το καταλαβαίνουν (στο μπλε). Στην περίπτωση που καταλαβαίνουν ένα δεδομένο κομμάτι, δεν χρειάζεται να το σχολιάσουν.

Στο τέλος, κάθε έγγραφο σχολιάζεται από τουλάχιστον τέσσερις σχολιαστές, ενώ ορισμένα έγγραφα σχολιάζονται από έως και έξι σχολιαστές. Υπολογίστηκε η μεταβλητή κάπα του Fleiss (Fleiss, 1971) για τέσσερις σχολιαστές που σχολίασαν όλα τα έγγραφα, το οποίο δίνει ένα χαμηλό 0,175 κάπα. Για ορισμένα ζεύγη σχολιαστών, η κάπα εμφανίζει ελαφρώς υψηλότερες τιμές (0,292 και 0,316). Αυτό σημαίνει ότι αυτή η εργασία σχολιασμού είναι πολύ υποκειμενική και εξαρτάται σε μεγάλο βαθμό από τις γνώσεις και την εμπειρία του κάθε ατόμου.

Το σώμα στη συνέχεια τμηματοποιείται σε προτάσεις που περιέχουν ένα κομμάτι στόχο ανά πρόταση. Συνολικά, αντλήθηκαν 9.709 προτάσεις με 3.482 σύνθετα και 6.227 μη σύνθετα κομμάτια. Στη συνέχεια, για κάθε εκπαίδευση και αξιολόγηση, αναδιανεμήθηκαν τα δεδομένα με διαφορετικό αριθμό σποράς και τα χωρίστηκαν σε τρία μέρη: 70% για εκπαίδευση, 15% για επικύρωση και 15% για δοκιμή.

Στην εν λόγω έρευνα, χρησιμοποιήθηκαν οι εξής αλγόριθμοι:

FastText Embedding (Bojanowski et al., 2016), ένα γλωσσικό μοντέλο προεκπαιδευμένο σε μεγάλα μη εργαστηριακά corpora, χρησιμοποιείται για την αναπαράσταση λέξεων.

CamemBERT Embedding (Martin et al., 2020) μια γαλλική έκδοση του BERT (Devlin et al., 2018), είναι προεκπαιδευμένο σε μεγάλο αριθμό κειμένων χρησιμοποιώντας μια παραλαγμένη προσέγγιση κάλυψης διαστήματος.

Το CNN ή το Convolutional Neural Network (LeCun et al., 1995) χρησιμοποιείται συχνά για ταξινόμηση εικόνων και άλλες εργασίες, αλλά είναι επίσης αποτελεσματικό για την αναγνώριση σύνθετων λέξεων (Aroyehun et al., 2018; Sheang, 2019). Το μοντέλο αυτής της εργασίας ακολουθεί την προσέγγιση του (Sheang, 2019) προσθέτοντας περισσότερες λίστες συχνότητων, αφαιρώντας ορισμένα γλωσσικά χαρακτηριστικά και αντικαθιστώντας το GloVe με το FastText και το CamemBERT.

Για την εκπαίδευση του μοντέλου, πρώτα, οι διανυσματικές αναπαραστάσεις όλων των χαρακτηριστικών για μια πρόταση προσαρτήθηκαν ως δισδιάστατο διάνυσμα ανάλογα με τον αριθμό των χαρακτηριστικών. Το CatBoost (Dorogush et al., 2018) είναι μια βαθμίδα ενίσχυσης στη βιβλιοθήκη Decision Trees που χρησιμοποιείται συχνά για εργασίες κατάταξης, παλινδρόμησης και ταξινόμησης. Το CatBoost είναι μια βιβλιοθήκη εκμάθησης συνόλου που συνδυάζει πολλαπλούς αλγόριθμους μηχανικής μάθησης (Δέντρα απόφασης) για να αποκτήσει ένα καλύτερο μοντέλο. Για να εκπαιδευτεί το μοντέλο, οι ερευνητές ακολούθησαν το ίδιο βήμα προετοιμασίας δεδομένων όπως στο μοντέλο CNN και στη συνέχεια το ανέπτυξαν σε ένα μακρύ διάνυσμα. Το μοντέλο εκπαιδεύτηκε για 1200 επαναλήψεις με εκμάθηση 0,03.

Το XGBoost (Chen & Guestrin, 2016) σημαίνει Extreme Gradient Boosting, είναι μια επεκτάσιμη, κατανεμημένη βιβλιοθήκη μηχανικής μάθησης δέντρων αποφάσεων με ενισχυμένη κλίση παρόμοια με την CatBoost. Είναι μια από τις κορυφαίες βιβλιοθήκες μηχανικής μάθησης για προβλήματα παλινδρόμησης, ταξινόμησης και κατάταξης. Για την εκπαίδευση του μοντέλου, τα δεδομένα προετοιμάζονται όπως στο μοντέλο CatBoost και στη συνέχεια εκπαιδεύονται με το μέγιστο βάθος 10, τον ρυθμό εκμάθησης 0,03 και τον αριθμό των εκτιμητών 500.

Συμπερασματικά, οι μελετητές έχουν προτείνει τρεις ταξινομητές με βάση το CNN, το CatBoost και το XGBoost με διαφορετικά σύνολα χαρακτηριστικών όπως FastText, BERT, μήκος λέξης, συλλαβή λέξεων, πλήθος φωνηέντων, κατάταξη λέξεων, TFIDF, συχνότητα LangGen και καθαρή συχνότητα για ανίχνευση σύνθετων λέξεων σε γαλλικά βιοϊατρικά έγγραφα. Δημιούργησαν επίσης το σύνολο δεδομένων και παρέχουν το σύνολο των πειραμάτων για την αξιολόγηση. Τα αποτελέσματα έδειξαν ότι τα μοντέλα αποδίδουν αρκετά καλά και το μοντέλο του CNN ξεπερνά τα άλλα.

10. Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks

Yau-Shian Wang, Hung-Yi Lee

Σε αυτό το άρθρο, προτάθηκε η εκπαίδευση ενός αυτόματου κωδικοποιητή που κωδικοποιεί κείμενο εισόδου σε αναγνώσιμες από τον άνθρωπο προτάσεις και έτσι επιτυγχάνεται μη ζευγοειδής αφηρημένη σύνοψη. Ο αυτόματος κωδικοποιητής αποτελείται από μια γεννήτρια και έναν ανακατασκευαστή. Η γεννήτρια κωδικοποιεί το κείμενο εισόδου σε μια συντομότερη ακολουθία λέξεων και η ανακατασκευή ανακτά την πληροφορία από την έξοδο της γεννήτριας. Για να γίνει η έξοδος της γεννήτριας αναγνώσιμη από τον άνθρωπο, ένας παράγοντας διάκρισης περιορίζει την έξοδο της γεννήτριας ώστε να μοιάζει με προτάσεις που έχουν γραφτεί από άνθρωπο. Λαμβάνοντας την έξοδο της γεννήτριας ως σύνοψη του κειμένου εισόδου, επιτυγχάνεται η αφηρημένη σύνοψη χωρίς ζεύγη εγγράφου-σύνοψης ως δεδομένα εκπαίδευσης. Τα υποσχόμενα αποτελέσματα εμφανίζονται τόσο στα αγγλικά όσο και στα κινεζικά σώματα.

Εν προκειμένω οι ερευνητές δανείστηκαν την ιδέα του GAN για να κάνουν την έξοδο της γεννήτριας αναγνώσιμη από τον άνθρωπο. Η κύρια πρόκληση στην εφαρμογή του GAN στη δημιουργία προτάσεων είναι η διακριτή φύση της φυσικής γλώσσας. Για τη δημιουργία μιας ακολουθίας λέξεων, η γεννήτρια έχει συνήθως μη διαφορικά μέρη, όπως argmax ή άλλες συναρτήσεις δείγματος που προκαλούν την αποτυχία του αρχικού GAN. Χρησιμοποιώντας το GAN, προτάθηκε ένα μοντέλο που κωδικοποιεί κείμενο ως αναγνώσιμη από τον άνθρωπο περίληψη, που μαθαίνεται χωρίς ζεύγη εγγράφου-σύνοψης.

11. Medical Text Simplification Using Reinforcement Learning

(TESLEA): Deep Learning–Based Text Simplification Approach

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, Vijay Mago

Αυτή η μελέτη στοχεύει να αναπτύξει μια προσέγγιση απλοποίησης κειμένου (TS) που βασίζεται σε βαθιά μάθηση που μετατρέπει το σύνθετο ιατρικό κείμενο σε μια απλούστερη έκδοση, διατηρώντας παράλληλα την ποιότητα του παραγόμενου κειμένου.

Αναπτύχθηκε μια προσέγγιση TS που χρησιμοποιεί ενισχυτική μάθηση και γλωσσικά μοντέλα που βασίζονται σε μετασχηματιστές. Η απόδοση συνάφειας, η απόδοση Flesch-Kincaid και η απόδοση λεξιλογικής απλότητας βελτιστοποιήθηκαν για να βοηθήσουν στην απλούστευση σύνθετων ιατρικών παραγράφων με πυκνότητα ορολογίας στις απλούστερες εκδόσεις τους, διατηρώντας παράλληλα την ποιότητα του κειμένου. Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας 3568 σύνθετες-απλές ιατρικές παραγράφους και αξιολογήθηκε σε 480 παραγράφους μέσω της βοήθειας αυτοματοποιημένων μετρήσεων και ανθρώπινου σχολιασμού.

Η προτεινόμενη μέθοδος ξεπέρασε τις προηγούμενες γραμμές βάσης στις βαθμολογίες Flesch-Kincaid (11,84) και πέτυχε συγκρίσιμη απόδοση με άλλες βασικές γραμμές όταν μετρήθηκε με τη χρήση βαθμολογιών ROUGE-1 (0,39), ROUGE-2 (0,11) και SARI (0,40). Η μη αυτόματη αξιολόγηση έδειξε ότι η ποσοστιαία συμφωνία μεταξύ ανθρώπινων σχολιαστών ήταν περισσότερο από 70% όταν ελήφθησαν υπόψη παράγοντες όπως η ευχέρεια, η συνοχή και η επάρκεια.

Αναπτύχθηκε με επιτυχία μια μοναδική προσέγγιση ιατρικής TS που αξιοποιεί την ενισχυτική μάθηση και απλοποιεί με ακρίβεια πολύπλοκες ιατρικές παραγράφους, αυξάνοντας έτσι την αναγνωσιμότητά τους. Η προτεινόμενη προσέγγιση TS μπορεί να εφαρμοστεί για την αυτόματη δημιουργία απλοποιημένου κειμένου για σύνθετα δεδομένα ιατρικού κειμένου, τα οποία θα ενίσχυαν την προσβασιμότητα της βιοϊατρικής έρευνας σε ένα ευρύτερο κοινό.

Δεδομένης μιας περίπλοκης ιατρικής παραγράφου, ο στόχος αυτής της εργασίας ήταν να δημιουργήσει μια απλοποιημένη παράγραφο που να είναι συνοπτική και να αποτυπώνει τις σημαντικές πληροφορίες που εκφράζονται στο σύνθετο κείμενο. Για να επιτευχθεί αυτό, προτάθηκε ένα μοντέλο απλοποίησης που βασίζεται σε RL, το οποίο βελτιστοποιεί πολλαπλές ανταμοιβές κατά τη διάρκεια της εκπαίδευσης και συντονίζεται χρησιμοποιώντας ένα σύνολο δεδομένων ιατρικού TS σε επίπεδο παραγράφου.

Η Cochrane Database of Scientific Reviews είναι μια βάση δεδομένων υγειονομικής περίθαλψης με πληροφορίες για ένα ευρύ φάσμα κλινικών θεμάτων. Κάθε ανασκόπηση περιλαμβάνει μια περίληψη σε απλή γλώσσα (PLS) γραμμένη από τους συγγραφείς που ακολουθούν οδηγίες για τη δομή των περιλήψεων. Τα PLS υποτίθεται ότι είναι σαφή, κατανοητά και προσβάσιμα, ειδικά για ένα ευρύ κοινό που δεν είναι εξοικειωμένο με τον τομέα της ιατρικής. Τα PLS είναι εξαιρετικά ετερογενή στη φύση τους και δεν είναι ζευγαρωτά (δηλαδή, για κάθε σύνθετη πρόταση μπορεί να μην υπάρχει αντίστοιχη απλούστερη έκδοση). Το πρωτοποριακό σύνολο δεδομένων που αναπτύχθηκε από τους Devaraj et al χρησιμοποιείται σε αυτή τη μελέτη για την εκπαίδευση των μοντέλων και είναι διαθέσιμο στο κοινό. Η λύση TS που προτείνεται για την απλοποίηση σύνθετου ιατρικού κειμένου χρησιμοποιεί ένα μοντέλο απλοποίησης που βασίζεται σε RL, το οποίο βελτιστοποιεί πολλαπλές ανταμοιβές (ανταμοιβή συνάφειας, ανταμοιβές βαθμού Flesch-Kincaid και ανταμοιβές λεξιλογικής απλότητας) για να επιτύχει μια πιο ολοκληρωμένη και συνοπτική απλοποίηση.

Το βασικό μοντέλο γλώσσας που χρησιμοποιήθηκε σε αυτή τη μελέτη για την εκτέλεση της απλοποίησης ήταν το BART, το οποίο είναι ένα μοντέλο κωδικοποιητή-αποκωδικοποιητή που βασίζεται σε μετασχηματιστή και προεκπαιδεύτηκε χρησιμοποιώντας μια αντικειμενική συνάρτηση διαλογής υλικού. Το τμήμα του αποκωδικοποιητή του μοντέλου έχει αυτοπαλινδρομικό χαρακτήρα, καθιστώντας το πιο κατάλληλο για εργασίες δημιουργίας προτάσεων. Επιπλέον, το μοντέλο BART επιτυγχάνει ισχυρές επιδόσεις σε εργασίες δημιουργίας φυσικής γλώσσας, όπως η σύνοψη, η οποία αποδείχθηκε σε σύνολα δεδομένων XSum και CNN/Daily Mail. Σε αυτήν την περίπτωση, χρησιμοποιείται μια έκδοση του BART που έχει ρυθμιστεί με ακρίβεια σε σύνολο δεδομένων XSUM.

Από τα αποτελέσματα που προέκυψαν, μπορεί να συναχθεί το συμπέρασμα ότι το TESLEA είναι αποτελεσματικό στη δημιουργία απλούστερου κειμένου σε σύγκριση με τεχνικές περιλήψεις, τις χρυσές παραπομπές (δηλαδή απλές ιατρικές παραγράφους που αντιστοιχούν σε πολύπλοκες ιατρικές παραγράφους) και τα βασικά μοντέλα. Χειροκίνητες αξιολογήσεις του TESLEA- Το κείμενο που δημιουργήθηκε διεξήχθη με τη βοήθεια ειδικών του τομέα χρησιμοποιώντας τους προαναφερθέντες παράγοντες και διεξήχθη περαιτέρω έρευνα.

12. Neural Versus Non-Neural Text Simplification: A Case Study

Islam Nassar, Michelle Ananda-Rajah, Gholamreza Haffari

Στην έρευνα αυτή προτάθηκε ένα αρθρωτό σύστημα βασισμένο σε κανόνες για την απλούστευση κειμένου και αποδείχθηκε η δυνατότητά του να ξεπερνά τις επιδόσεις του υπερσύγχρονου συστήματος απλούστευσης που βασίζεται σε νευρωνικά δίκτυα από την άποψη της απλότητας. Έγινε σύγκρισή της εξόδου και των δύο συστημάτων για να επισημάνθούν οι διαφορές μεταξύ των δύο προσεγγίσεων. Επιπλέον, παρουσιάστηκε μια προσαρμογή του συστήματος για να χειρίζεται συγκεκριμένες εργασίες, όπου χρησιμοποιήθηκε μια υβριδική προσέγγιση του συστήματος που βασίζεται σε κανόνες και της αυτόματης μετάφρασης που βασίζεται σε φράσεις για την απλοποίηση των περιλήψεων ιατρικών εξιτηρίων σε μια κατάσταση χαμηλών πόρων.

Το προτεινόμενο σύστημα απλοποίησης βάσει κανόνων περιλαμβάνει δύο ενότητες που αντιστοιχούν στις δύο κύριες λειτουργίες απλοποίησης κειμένου: λεξιλογική και συντακτική απλοποίηση. Οι συγγραφείς χρησιμοποίησαν μια νευρωνική προσέγγιση για την απλοποίηση κειμένου όπως προτείνεται από (Zhang & Lapata, 2017) και μεταγλωττίστηκε ως DRESS (Deep REinforcement Sentence Simplification). Αντιμετωπίζουν την απλοποίηση κειμένου ως μια εργασία μοντελοποίησης αλληλουχίας σε ακολουθία. Αντλούν έμπνευση από τη Νευρωνική Μηχανική Μετάφραση, όπου εκπαιδεύουν ένα μοντέλο κωδικοποιητή-αποκωδικοποιητή σε ένα μονόγλωσσο παράλληλο σώμα σύνθετων-απλών αγγλικών. Για να ενθαρρύνουν περαιτέρω ένα απλούστερο αποτέλεσμα, εκπαιδεύουν το μοντέλο τους σε ένα πλαίσιο ενισχυτικής μάθησης όπου η συνάρτηση ανταμοιβής είναι ένας σταθμισμένος συνδυασμός της συνάφειας, της απλότητας και της ευχέρειας της πρότασης εξόδου. Ως μέσο αξιολόγησης για τη συνάφεια, χρησιμοποιούν έναν αυτόματο κωδικοποιητή ακολουθίας που βασίζεται σε LSTM για να αποκτήσουν μια διανυσματική αναπαράσταση και για τις δύο προτάσεις πηγής και εξόδου, ενώ στη συνέχεια όρισαν την απόδοση συνάφειας ως την ομοιότητα συνημιτόνου μεταξύ αυτών των δύο διανυσμάτων.

Όσον αφορά την απόδοση της ευχέρειας, χρησιμοποιούν ένα μοντέλο γλώσσας LSTM εκπαιδευμένο σε απλές προτάσεις για να λάβουν μια κανονικοποιημένη βαθμολογία αμηχανίας για την πρόταση εξόδου. Για την απόδοση απλότητας, χρησιμοποιούν τη μέτρηση SARI (Xu et al. 2016) που μετρά τις επικαλύψεις n-gram μεταξύ προτάσεων πηγής, εξόδου και αναφοράς. Τέλος, για να ενθαρρύνουν τη λεξιλογική απλοποίηση, χρησιμοποιούν ένα ξεχωριστό προ-εκπαιδευμένο μοντέλο κωδικοποιητή-αποκωδικοποιητή, εκπαιδευμένο σε ένα μη ενισχυμένο περιβάλλον σε ένα παράλληλο σώμα σύνθετων-απλές προτάσεων, για να λάβουν πιθανότητες λεξικής υποκατάστασης με βάση μια δεδομένη πρόταση πηγής. Η χρήση του τελευταίου μοντέλου ευνοεί τις λειτουργίες λεξικής απλοποίησης, αλλά δεν λαμβάνει υπόψη την ευχέρεια της συνολικής παραγωγής. Επομένως, η έξοδος του συστήματός τους προσδιορίζεται με γραμμικό συνδυασμό των δύο μοντέλων κωδικοποιητή-αποκωδικοποιητή.

Όσον αφορά τα αποτελέσματα της ιατρικής απλοποίησης, κοιτάζοντας τον μέσο όρο των λέξεων ανά πρόταση, είναι προφανές ότι αυτό το σύστημα τείνει να παράγει μεγαλύτερες απλουστευμένες προτάσεις. Αυτό οφείλεται στη φύση της ιατρικής απλοποίησης που απαιτεί επεξεργασία.

13. Optimizing Statistical Machine Translation for Text Simplification

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen and Chris Callison-Burch

Σε αυτό το άρθρο, οι συγγραφείς διεξήγαγαν μια εις βάθος προσαρμογή της στατιστικής αυτόματης μετάφρασης για την απλοποίηση κειμένου, εκμεταλλευόμενοι παραφράσεις μεγάλης κλίμακας από δίγλωσσα κείμενα και μικρό αριθμό χειροκίνητων απλοποιήσεων με πολλαπλές αναφορές. Η εργασία αυτή είναι η πρώτη που σχεδιάζει αυτόματες μετρήσεις που είναι αποτελεσματικές για τον συντονισμό και την αξιολόγηση συστημάτων απλοποίησης, οι οποίες θα διευκολύνουν την επαναληπτική ανάπτυξη για αυτήν την εργασία. Οι συγγραφείς παρουσιάζουν μια πλήρη προσαρμογή ενός πλαισίου μηχανικής μετάφρασης που βασίζεται στη σύνταξη για την απλοποίηση.

Για τη διαμόρφωση ενός αγωγού SMT χρειάστηκαν:

1) δύο νέες συντονίσιμες μετρήσεις για συγκεκριμένες απλοποιήσεις. 2) κανόνες παράφρασης μεγάλης κλίμακας που προέρχονται αυτόματα από δίγλωσσα παράλληλα σώματα, τα οποία είναι πιο φυσικά και άφθονα διαθέσιμα από τα χειροκίνητα απλοποιημένα κείμενα. 3) πλούσια χαρακτηριστικά απλοποίησης σε επίπεδο κανόνα. και 4) πολλαπλές απλουστεύσεις αναφοράς που συλλέγονται μέσω crowdsourcing για συντονισμό και αξιολόγηση. Το πείραμα σημειωτέον επικεντρώνεται κυρίως στη λεξική απλοποίηση (αναγραφή λέξεων ή φράσεων με απλούστερες εκδόσεις) και σε μικρότερο βαθμό σε κανόνες συντακτικής επανεγγραφής που απλοποιούν την εισαγωγή. Η εστίαση στη λεξιλογική απλοποίηση δεν επηρεάζει τη γενικότητα της εργασίας που παρουσιάζεται, καθώς η διαγραφή ή ο διαχωρισμός προτάσεων θα μπορούσε να εφαρμοστεί ως βήματα πριν ή μετά την επεξεργασία.

Οι συγγραφείς εκπαιδύσαν αλγόριθμους, όπως MERT (Och, 2003) ή PRO (Hopkins and May, 2011), στη συνέχεια βελτιστοποιούν απευθείας τις παραμέτρους του μοντέλου έτσι ώστε η ποιότητα απλοποίησης από άκρο σε άκρο να είναι βέλτιστη. Οι συγγραφείς προτείνουν τη χρήση δύο νέων μετρήσεων ελαφρού βάρους: FKBLEU που μετρά ρητά την αναγνωσιμότητα και SARI που τη μετρά σιωπηρά συγκρίνοντας με την είσοδο και τις αναφορές. Οι συγγραφείς παρουσίασαν μια αποτελεσματική προσαρμογή των τεχνικών στατιστικής αυτόματης μετάφρασης. Προτείνουν δύο νέες κατευθύνσεις: σχεδιασμό συντονισμένων μετρήσεων που συσχετίζονται με ανθρώπινες κρίσεις και χρήση εμπλουτισμένων με απλότητα κανόνων παράφρασης που προέρχονται από μεγαλύτερα δεδομένα από το σύνολο δεδομένων της Wikipedia Normal-Simple. Κρίνουν δε ότι μπορεί να είναι δυνατό να σχεδιαστεί μια καθολική μέτρηση που να λειτουργεί για πολλαπλές εργασίες δημιουργίας κειμένου σε κείμενο (συμπεριλαμβανομένης της απλοποίησης προτάσεων, της συμπίεσης και της διόρθωσης σφαλμάτων), ταυτόχρονα χρησιμοποιώντας την ίδια ιδέα σύγκρισης της παραγωγής του συστήματος με πολλαπλές αναφορές. Η μέτρηση θα μπορούσε ενδεχομένως να περιλαμβάνει συντονισμένες παραμέτρους ή σταθμισμένες ανθρώπινες κρίσεις σε αναφορές για την αντιμετώπιση διαφορετικών εργασιών.

14. Paragraph-level Simplification of Medical Texts

Ashwin Devaraj, Byron C. Wallace, Iain J. Marshall, Junyi Jessy Li

Οι συγγραφείς προτείνουν μια νέα μέτρηση που βασίζεται σε βαθμολογίες πιθανότητας από ένα μοντέλο καλυμμένης γλώσσας προεκπαιδευμένο σε επιστημονικά κείμενα. Δείχνουν ότι αυτό το αυτοματοποιημένο μέτρο διαφοροποιεί καλύτερα τις τεχνικές και τις απλές περιλήψεις από τις υπάρχουσες ευρετικές. Πιο συγκεκριμένα, διερευνούν νέα δεδομένα και σύγχρονα μοντέλα παραγωγής κειμένου υπό όρους (Lewis et al., 2020) για την απλοποίηση των ιατρικών εγγράφων. Εισάγουν ένα σύνολο δεδομένων ζευγαρωτών (τεχνικών, απλοποιημένων) κειμένων που προέρχονται από τη βάση δεδομένων Cochrane των Συστηματικών Ανασκοπήσεων, η οποία αποτελείται από συνθέσεις στοιχείων για ένα ευρύ φάσμα κλινικών θεμάτων. Ουσιαστικά, κάθε κριτική περιλαμβάνει μια περίληψη σε απλή γλώσσα (PLS) γραμμένη από τους συγγραφείς. Τα PLS γράφονται απευθείας από τις πλήρεις κριτικές με τη δική τους δομή και οδηγίες. Δεν πρόκειται για απλοποιημένες εκδόσεις των αντίστοιχων τεχνικών περιλήψεων των κριτικών, ούτε είναι περιλήψεις των περιλήψεων.

Συνέταξαν ένα σύνολο δεδομένων τεχνικών περιλήψεων βιοϊατρικών συστηματικών ανασκοπήσεων και αντίστοιχων PLS από τη βάση δεδομένων Cochrane Database of Systematic Reviews, η οποία περιλαμβάνει χιλιάδες συνόψεις στοιχείων (όπου οι συγγραφείς παρέχουν μια επισκόπηση όλων των δημοσιευμένων στοιχείων σχετικά με μια συγκεκριμένη κλινική ερώτηση ή θέμα). Τα πρότυπα PLS του Cochrane (Cochrane, 2013) συνιστούν ότι «το PLS πρέπει να είναι γραμμένο σε απλά αγγλικά, τα οποία μπορούν να γίνουν κατανοητά από τους περισσότερους αναγνώστες χωρίς πανεπιστημιακή εκπαίδευση». Τα PLS δεν είναι παράλληλα με κάθε πρόταση αφηρημένη. Αντίθετα, είναι δομημένα ετερογενώς (Kadic et al., 2016). Για να εξάγουν το σύνολο δεδομένων συνέγκριναν την ηλεκτρονική διεπαφή στη βάση δεδομένων για άρθρα που περιέχουν PLS, εξάγοντας το ακατέργαστο κείμενο των τεχνικών περιλήψεων και PLS για εκείνα που προσδιόρισαν. Με αυτόν τον τρόπο απέκτησαν 7820 ζεύγη μετά την αφαίρεση προβληματικών συνδέσμων (π.χ. σφάλματα HTTP 404). Επίσης, απέκλεισαν κριτικές με άτυπη μορφοποίηση που θα απαιτούσαν εκτενή χειροκίνητη επιθεώρηση.

Τα αποτελέσματα που παράγονται μέσω μοντέλων που έχουν εκπαιδευτεί με τον προτεινόμενο στόχο UL επιτυγχάνουν σημαντικά χαμηλότερες βαθμολογίες Flesch-Kincaid από αυτές που επιτυγχάνονται τόσο από τις τεχνικές περιλήψεις όσο και από το PLS αναφοράς, ενώ το μοντέλο που εκπαιδεύτηκε χωρίς UL παρήγαγε κείμενα με υψηλότερο επίπεδο ανάγνωσης από το PLS.

Περιέργως, οι ρυθμίσεις UL-Newsela και UL-και οι δύο, που χρησιμοποιούν το σύνολο δεδομένων Newsela για να παράγουν σταθμίσεις απίθανης πιθανότητας, δεν απέδωσαν μείωση στα εκτιμώμενα επίπεδα βαθμών. Αυτά τα αποτελέσματα υποδηλώνουν ότι: (1) Το BART είναι ικανό να πραγματοποιεί απλοποίηση ιατρικών κειμένων έτσι ώστε τα αποτελέσματα να απολαμβάνουν μειωμένα επίπεδα ανάγνωσης σε σύγκριση με εκείνα των τεχνικών περιλήψεων. (2) Η προτεινόμενη χρήση του UL για τη ρητή τιμωρία του μοντέλου για την παραγωγή ορολογίας επιτρέπει τη δημιουργία κειμένου με ακόμη μεγαλύτερη αναγνωσιμότητα από το PLS αναφοράς. Τα επίπεδα ανάγνωσης ακόμη και των απλουστευμένων αποτελεσμάτων, ωστόσο, είναι στα επίπεδα του τελευταίου γυμνασίου/πρώιμου κολεγίου. Αυτό θα μπορούσε να αντανakλά τις σχετικά μικρές διαφορές στις βαθμολογίες αναγνωσιμότητας μεταξύ των περιλήψεων και του PLS γενικά. Έδειξαν επίσης ότι οι συγκαλυμμένες βαθμολογίες πιθανοτήτων SciBERT είναι πιο χρήσιμες ως διαχωρισμός μεταξύ τεχνικών περιλήψεων και PLS από τις τυπικές μετρήσεις αναγνωσιμότητας, οι οποίες χρησιμοποιούν ενδείξεις σε επίπεδο επιφάνειας, όπως μετρήσεις λέξεων και προτάσεων.

Εξέτασαν το σημαντικό καθήκον της απλοποίησης του ιατρικού κειμένου. Βρήκαν έναν νέο πόρο για αυτήν την εργασία που αποτελείται από τεχνικές περιλήψεις που συνοψίζουν ιατρικά στοιχεία σε συνδυασμό με εκδόσεις του ίδιου σε απλή γλώσσα. Πρότειναν ένα νέο μοντέλο γλώσσας (MLM) ως αξιολογικό μέτρο της τεχνικής ικανότητας του κειμένου, το οποίο ποσοτικοποιεί την τεχνικότητα υπολογίζοντας την πιθανότητα ύπαρξης διακριτικών στο κείμενο εισόδου σε σχέση με ένα MLM που βασίζεται σε μετασχηματιστή και έχει εκπαιδευτεί σε ένα τεχνικό σώμα. Αποδείχθηκε επίσης ότι αυτή η μέτρηση διέκρινε καλύτερα τις τεχνικές περιλήψεις από το PLS από τις πιο παραδοσιακές έννοιες της αναγνωσιμότητας. Πρότειναν τέλος μοντέλα για αυτοματοποιημένη απλούστευση με βάση το BART (Lewis et al., 2020), επεκτείνοντας τον εκπαιδευτικό στόχο ενσωματώνοντας μια ρητή ποινή για την παραγωγή όρων «jargon». Διαπίστωσαν ότι αυτή η μέθοδος μπορεί να βελτιώσει τα αποτελέσματα του μοντέλου (δηλαδή, μπορεί να αυξήσει την απλότητα και την αφαιρετικότητα των περιλήψεων) σύμφωνα με τις μετρήσεις που εξετάζονται.

15.Semantic Structural Evaluation for Text Simplification

Elior Sulem, Omri Abend, Ari Rappoport

Σε αυτό το άρθρο οι συγγραφείς προτείνουν το πρώτο μέτρο για την αντιμετώπιση των δομικών πτυχών της απλοποίησης κειμένου, που ονομάζεται SAMSA (Απλοποίηση Αυτόματη αξιολόγηση Μέτρο μέσω Σημασιολογικού Σχολιασμού). Αξιοποιούν τις πρόσφατες εξελίξεις στη σημασιολογική ανάλυση για να αξιολογήσουν την ποιότητα της απλοποίησης αποσυνθέτοντας τα δεδομένα εισόδου με βάση τη σημασιολογική του δομή και συγκρίνοντάς τα με τα αποτελέσματα. Η SAMSA παρέχει μια διαδικασία αυτόματης αξιολόγησης χωρίς αναφορά, αποφεύγοντας τα προβλήματα που αντιμετωπίζουν οι μέθοδοι που βασίζονται σε αναφορές λόγω του τεράστιου χώρου έγκυρων απλοποιήσεων για μια δεδομένη πρόταση. Τα πειράματά τους με ανθρώπινη αξιολόγηση δείχνουν τόσο την ουσιαστική συσχέτιση της SAMSA με τις ανθρώπινες κρίσεις, όσο και την ανεπάρκεια των υφιστάμενων μέτρων που βασίζονται σε αναφορές στην αξιολόγηση της δομικής απλοποίησης.

Η κύρια υπόθεση της SAMSA είναι ότι μια δομικά σωστή απλοποίηση είναι εκείνη όπου: (1) κάθε πρόταση περιέχει ένα μεμονωμένο γεγονός από την είσοδο, (2) η κύρια σχέση καθενός από τα γεγονότα και οι συμμετέχοντες τους διατηρούνται στην έξοδο.

Το άρθρο παρουσίασε την πρώτη μέτρηση με επίγνωση δομής για απλοποίηση κειμένου, το SAMSA, και τα πρώτα πειράματα αξιολόγησης που στοχεύουν άμεσα το στοιχείο δομικής απλοποίησης, ξεχωριστά από το λεξικό στοιχείο. Οι δομικές και λεξιλογικές διαστάσεις της απλοποίησης συνδέονται χαλαρά και ότι τα πρωτόκολλα αξιολόγησης TS θα πρέπει να αξιολογούν και τα δύο. Οι συγγραφείς αποδεικνύουν εμπειρικά ότι τα ισχυρά μέτρα που αξιολογούν την ποιότητα της λεξιλογικής απλοποίησης (ιδίως το SARI), αποτυγχάνουν να συσχετιστούν με τις ανθρώπινες κρίσεις όταν εκτελείται δομική απλοποίηση από τα αξιολογούμενα συστήματα. Τα πειράματά μας δείχνουν ότι το SAMSA συσχετίζεται καλά με τις ανθρώπινες κρίσεις σε τέτοιες ρυθμίσεις, γεγονός που καταδεικνύει τη χρησιμότητά του για την αξιολόγηση και τον συντονισμό στατιστικών συστημάτων απλούστευσης και δείχνει ότι η δομική αξιολόγηση παρέχει μια συμπληρωματική προοπτική για την ποιότητα της απλούστευσης.

16. Sentence Simplification with Deep Reinforcement Learning

Xingxing Zhang and Mirella Lapata

Αυτό το άρθρο περιγράφει ένα μοντέλο, το οποίο ονομάζεται DRESS (ως συντομογραφία του Deep REinforcement Sentence Simplification), εξερευνά το χώρο των πιθανών απλοποιήσεων ενώ μαθαίνει να βελτιστοποιεί μια συνάρτηση απόδοσης που ενθαρρύνει τα αποτελέσματα που είναι απλά, άπταιστα και διατηρούν το νόημα της εισαγωγής. Πειράματα σε τρία σύνολα δεδομένων δείχνουν ότι το μοντέλο μας ξεπερνά τα ανταγωνιστικά συστήματα απλοποίησης.

Σε αυτό το άρθρο οι συγγραφείς προτείνουν ένα μοντέλο απλοποίησης που βασίζεται σε γνώσεις από τη νευρωνική μηχανική μετάφραση. Κεντρικό στοιχείο αυτής της προσέγγισης είναι μια αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή που υλοποιείται από επαναλαμβανόμενα νευρωνικά δίκτυα. Ο κωδικοποιητής διαβάζει την ακολουθία πηγής σε μια λίστα αναπαραστάσεων συνεχούς χώρου από την οποία ο αποκωδικοποιητής δημιουργεί την ακολουθία στόχο. Αν και το μοντέλο μας χρησιμοποιεί την αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή ως ραχοκοκαλιά, πρέπει επίσης να πληροί τους περιορισμούς που επιβάλλονται από την ίδια την εργασία απλοποίησης, δηλαδή, η προβλεπόμενη έξοδος πρέπει να είναι απλούστερη, να διατηρεί το νόημα της εισόδου και να είναι γραμματική. Για να ενσωματώσει αυτή τη γνώση, το μοντέλο εκπαιδεύεται σε ένα πλαίσιο ενισχυτικής μάθησης (Williams, 1992): διερευνά το χώρο των πιθανών απλοποιήσεων ενώ μαθαίνει να μεγιστοποιεί μια αναμενόμενη συνάρτηση ανταμοιβής που ενθαρρύνει αποτελέσματα που πληρούν περιορισμούς που αφορούν την απλοποίηση. Τα αποτελέσματα δείχνουν ότι η ενισχυτική μάθηση προσφέρει ένα εξαιρετικό μέσο για την εισαγωγή προηγούμενων γνώσεων στην εργασία απλοποίησης, επιτυγχάνοντας καλά αποτελέσματα σε τρία σύνολα δεδομένων.

17. Simple and Effective Text Simplification Using Semantic and Neural Methods

Elior Sulem, Omri Abend, Ari Rappoport

Οι συγγραφείς παρουσιάζουν έναν απλό και αποτελεσματικό αλγόριθμο διαχωρισμού που βασίζεται σε έναν αυτόματο σημασιολογικό αναλυτή. Μετά τη διαίρεση, το κείμενο είναι κατάλληλο για περαιτέρω ακριβείς λειτουργίες απλοποίησης. Συγκεκριμένα, δείχνουν ότι η νευρωνική μηχανική μετάφραση μπορεί να χρησιμοποιηθεί αποτελεσματικά σε αυτήν την περίπτωση. Ο διαχωρισμός με βάση τη σημασιολογική ανάλυση, όπως προτείνεται εδώ, μειώνει την αποτυχία τροποποίησης του πηγής με οποιονδήποτε τρόπο, άλλων συστημάτων. Η εκτενής αυτόματη και ανθρώπινη αξιολόγηση δείχνει ότι η προτεινόμενη μέθοδος συγκρίνεται ευνοϊκά με την τελευταία λέξη της τεχνολογίας στη συνδυασμένη λεξιλογική και δομική απλοποίηση.

Αποδεικνύεται επίσης ότι η ρητή ενσωμάτωση του διαχωρισμού των προτάσεων στο σύστημα απλοποίησης θα μπορούσε επίσης να μειώσει τον συντηρητισμό, ο οποίος αποτελεί σοβαρό περιορισμό των συστημάτων TS που βασίζονται σε NMT. Πράγματι, πειραματιζόμενοι οι ερευνητές με ένα υπερσύγχρονο νευρωνικό σύστημα (Nisioi et al., 2017), διαπιστώνεται ότι το 66% των προτάσεων εισόδου παραμένουν αμετάβλητες, ενώ καμία από τις αντίστοιχες αναφορές δεν είναι πανομοιότυπη με την πηγή. Η ανθρώπινη και αυτόματη αξιολόγηση των αναφορών (σε σχέση με άλλες αναφορές), επιβεβαιώνει ότι οι αναφορές είναι όντως απλούστερες από την πηγή, υποδηλώνοντας ότι ο παρατηρούμενος συντηρητισμός είναι υπερβολικός. Οι μέθοδοι των ερευνητών για την εκτέλεση διαχωρισμού προτάσεων ως προεπεξεργασία επιτρέπουν στο σύστημα TS να εκτελεί άλλες δομικές (π.χ. διαγραφές) και λεξιλογικές (π.χ. αντικαταστάσεις λέξεων) λειτουργίες, αυξάνοντας έτσι τόσο τη δομική όσο και τη λεξιλογική απλότητα.

Οι διαχωρισμένες προτάσεις εκτελούνται μέσω του υπερσύγχρονου νευρικού συστήματος TS NTS (Nisioi et al., 2017), που έχει δημιουργηθεί χρησιμοποιώντας το πλαίσιο νευρωνικής μηχανικής μετάφρασης OpenNMT (Klein et al., 2017). Η αρχιτεκτονική περιλαμβάνει δύο επίπεδα LSTM, με κρυφές καταστάσεις 500 μονάδων το καθένα, καθώς και παγκόσμια προσοχή σε συνδυασμό με τροφοδοσία εισόδου (Luong et al., 2015). Η εκπαίδευση γίνεται με πιθανότητα εγκατάλειψης 0,3 (Srivastava et al., 2014). Αυτό το μοντέλο χρησιμοποιεί πιθανότητες

ευθυγράμμισης μεταξύ των προβλέψεων και των αρχικών προτάσεων, αντί για μοντέλα που βασίζονται σε χαρακτήρες, για να ανακτήσει τις αρχικές λέξεις.

Οι συγγραφείς παρουσίασαν το πρώτο σύστημα απλοποίησης που συνδυάζει σημασιολογικές δομές και νευρωνική μηχανική μετάφραση, δείχνοντας ότι ξεπερνά τα υπάρχοντα λεξιλογικά και δομικά συστήματα. Η προτεινόμενη προσέγγιση αντιμετωπίζει τον υπερβολικό συντηρητισμό των συστημάτων που βασίζονται σε MT για TS, τα οποία συχνά αποτυγχάνουν να τροποποιήσουν την πηγή με οποιονδήποτε τρόπο. Το σημασιολογικό συστατικό εκτελεί διαχωρισμό προτάσεων χωρίς να βασίζεται σε ένα εξειδικευμένο σώμα, αλλά μόνο σε έναν σημασιολογικό αναλυτή. Η θεώρηση της διάσπασης προτάσεων ως αποσύνθεση μιας πρότασης στις σκηνές της υποστηρίζεται περαιτέρω από την πρόσφατη εργασία σχετικά με τη δομική αξιολόγηση TS (Sulem et al., 2018), η οποία προτείνει τη μέτρηση SAMSA.

18. Split and Rephrase

Shashi Narayan, Claire Gardent, Shay B. Cohen, Anastasia Shimorina

Οι συγγραφείς προτείνουν μια νέα εργασία απλοποίησης προτάσεων (Split-and-Rephrase) με στόχο να χωρίσουν μια σύνθετη πρόταση σε ένα νόημα που διατηρεί την ακολουθία μικρότερων προτάσεων. Οι συγγραφείς υποστηρίζουν ότι εφόσον οι μικρότερες προτάσεις γενικά επεξεργάζονται καλύτερα από τα συστήματα NLP, θα μπορούσε να χρησιμοποιηθεί ως βήμα προεπεξεργασίας που διευκολύνει και βελτιώνει την απόδοση των αναλυτών, των σημασιολογικών ετικετών ρόλων και των συστημάτων μηχανικής μετάφρασης. Θα πρέπει επίσης να είναι χρήσιμο για άτομα με αναγνωστικές δυσκολίες επειδή επιτρέπει τη μετατροπή μεγαλύτερων προτάσεων σε μικρότερες. Σε αυτό το άρθρο, οι συγγραφείς δημιουργούν και καθιστούν διαθέσιμο ένα σημείο αναφοράς που αποτελείται από 1.066.115 πλειάδες που αντιστοιχίζουν μια σύνθετη πρόταση σε μια ακολουθία προτάσεων που εκφράζουν το ίδιο νόημα. Δεύτερον, προτείνουν πέντε μοντέλα (vanilla sequence-to-sequence σε μοντέλα με σημασιολογικά κίνητρα) για να κατανοήσουν τη δυσκολία της προτεινόμενης εργασίας.

Για την εκπαίδευση μοντέλων Split-and-Rephrase, αυτό το σύνολο δεδομένων είναι αναμφισβήτητα ακατάλληλο καθώς αποτελείται από 108.016 σύνθετες και 114.924 απλοποιημένες προτάσεις, αποδίδοντας έτσι έναν μέσο αριθμό απλών προτάσεων ανά σύνθετη πρόταση 1,06. Κατά μέσο όρο, αυτό το νέο σύνολο δεδομένων συσχετίζει 4,99 απλές προτάσεις με κάθε σύνθετη πρόταση. Στη έρευνα αυτή αντλούν ένα σύνολο δεδομένων Split-and-Rephrase από το σώμα WEBNLG που παρουσιάζεται στους Gardent et al. (2017).

Για όλα τα νευρωνικά μοντέλα, χρησιμοποιήθηκαν RNN με μονάδες LSTM τριών επιπέδων, 500 κρυφές καταστάσεις και εγκατάλειψη τακτοποίησης με πιθανότητα 0,8. Όλες οι παράμετροι LSTM αρχικοποιήθηκαν τυχαία σε ομοιόμορφη κατανομή εντός $[-0,05, 0,05]$. Οι συγγραφείς εκπαιδύσαν τα μοντέλα με στοχαστική κλίση κατάβασης με αρχικό ρυθμό εκμάθησης 0,5. Πραγματοποίησαν επίσης mini-batch εκπαίδευση με μέγεθος παρτίδας 64 προτάσεων για SEQ2SEQ και MULTISEQ2SEQ και 32 για SPLIT-SEQ2SEQ και SPLIT-MULTISEQ2SEQ. Καθώς το μέγεθος του λεξιλογίου των δεδομένων WEBSPLIT είναι μικρό, εκπαιδύσαν τόσο τον κωδικοποιητή όσο και τον αποκωδικοποιητή με πλήρες λεξιλόγιο. Αρχικοποίησαν τυχαία τις ενσωματώσεις λέξεων στην αρχή και άφησαν το μοντέλο να τις εκπαιδεύσει κατά τη διάρκεια της εκπαίδευσης. Χρησιμοποίησαν το σύστημα των Zoph and Knight (2016) για να εκπαιδεύσουν τόσο απλά μοντέλα ακολουθίας σε ακολουθία όσο και πολλαπλών πηγών ακολουθίας σε ακολουθία και το σύστημα των Narayan και Gardent (2014) για να εκπαιδεύσουν το μοντέλο HYBRIDSIMPL.

Συμπερασματικά, οι συγγραφείς έχουν προτείνει μια νέα εργασία απλοποίησης προτάσεων την οποία ονομάζουν "Split and Rephrase". Έχουν δημιουργήσει ένα νέο σώμα για αυτήν την εργασία, το οποίο βασίζεται σε άμεσα διαθέσιμα δεδομένα που χρησιμοποιούνται για την αξιολόγηση NLG (Natural Language Generation). Τα αρχικά πειράματα υποδεικνύουν ότι η ικανότητα διαχωρισμού είναι ένας βασικός παράγοντας για τη δημιουργία ρευστού και νοήματος διατήρησης των αναδιατυπώσεων, επειδή επιτρέπει τη μείωση μιας σύνθετης εργασίας δημιουργίας (δημιουργία κειμένου που αποτελείται από τουλάχιστον δύο προτάσεις) σε μια σειρά απλούστερων εργασιών (δημιουργία σύντομων προτάσεων). Ενώ τα αποτελέσματα που παρουσιάζονται σε αυτό το έγγραφο χρησιμοποιούν μια έκδοση του σώματος WebNLG που αποτελείται από 13.308 ζεύγη

MR-Text, 7049 διακριτά MR και 8 κατηγορίες DBpedia, το τρέχον σώμα WebNLG περιλαμβάνει 43.056 ζεύγη MR-Text, 16.138 διακριτές κατηγορίες DBpedia.

19. Synonym Replacement based on a Study of Basic-level Nouns in Swedish Texts of Different Complexity

Evelina Rennes, Arne Jonsson

Σε αυτό το άρθρο, διερευνάται η χρήση ουσιαστικών βασικού επιπέδου σε κείμενα διαφορετικής πολυπλοκότητας και υποθέτουμε ότι τα υπερώνυμα με χαρακτηριστικά λέξεων βασικού επιπέδου θα μπορούσαν να είναι χρήσιμα για το έργο της λεξικής απλοποίησης. Πραγματοποιήθηκαν δύο μελέτες σωμάτων χρησιμοποιώντας τέσσερα διαφορετικά σώματα, δύο σώματα τυπικών σουηδικών και δύο σώματα απλών σουηδικών, και διερευνήθηκαν εάν τα σώματα απλών κειμένων περιέχουν υψηλότερη αναλογία ουσιαστικών βασικού επιπέδου από τα σώματα των τυπικών σουηδικών. Με βάση πληροφορίες από τις μελέτες corpus, αναπτύχθηκε ένας νέος αλγόριθμος για την επιλογή του καλύτερου συνωνύμου επιβραβεύοντας τις υψηλές σχετικές συχνότητες και τη μονολεξιμότητα και περιορίζοντας την άνοδο στην ιεραρχία των λέξεων ώστε να μην προτείνουμε συνώνυμα πολύ υψηλού επιπέδου περιεκτικότητας.

Διερευνήθηκε η χρήση ουσιαστικών βασικού επιπέδου μάθησης σε κείμενα διαφορετικής πολυπλοκότητας καθώς εκτιμήθηκε ότι τα υπερώνυμα με χαρακτηριστικά λέξεων βασικού επιπέδου θα μπορούσαν να είναι χρήσιμα για το έργο της λεξικής απλοποίησης. Στη συνέχεια χρησιμοποιήθηκε αυτή η εκτίμηση για τη δημιουργία ενός αλγόριθμου για την αντικατάσταση συνωνύμων.

Οι συγγραφείς συνέκριναν την απόδοση του συνδυασμένου αλγόριθμου συχνότητας/μονολεξίας (εδώ και στη συνέχεια: FM) με δύο βασικούς αλγόριθμους. Η πρώτη γραμμή βάσης (OneLevel) επέλεγε πάντα τη λέξη ένα επίπεδο ψηλότερα στην ιεραρχία ως το καλύτερο συνώνυμο. Αν δεν υπήρχε υπερέχουσα λέξη, η λέξη παρέμενε αμετάβλητη. Η δεύτερη γραμμή βάσης (Freq) επέλεγε πάντα τη λέξη με τη συνολική υψηλότερη σχετική συχνότητα ως το καλύτερο συνώνυμο, αγνοώντας έτσι τις πληροφορίες για την ύπαρξη μίας μόνο λέξης ως κατάλληλης.

Ο αλγόριθμος για την εύρεση συνωνύμων που προτείνεται σε αυτό το άρθρο βασίζεται στη θεωρία και τις μελέτες corpus. Δηλώνουν ότι θα ήταν πολύτιμο να συγκριθεί ο αλγόριθμος με συνώνυμα από, για παράδειγμα, το λεξικό SynLex, και να αξιολογηθεί εάν τα συνώνυμα που ανταλλάσσονται είναι απλούστερα, όταν συμβουλευόμαστε λεξικά βασικών λεξιλογίων, καθώς και φυσικούς ομιλητές. Μπορεί επίσης να ενισχυθεί με τεχνικές για τη χρήση σημασιολογικής και συνωνυμικής ομοιότητας

Παρατηρήθηκε ότι το σώμα του απλού κειμένου ειδήσεων περιλάμβανε πράγματι περισσότερα ουσιαστικά βασικού επιπέδου από το σώμα των τυπικών ειδήσεων. Αυτό με τη σειρά του δείχνει ότι η λεξιλογική απλοποίηση, μέσω της χρήσης ουσιαστικών βασικού επιπέδου, μπορεί να ωφεληθεί από τη διέλευση μιας ιεραρχίας λέξεων προς τα πάνω. Αυτό θα μπορούσε να χρησιμεύσει ως συμπλήρωμα στις συχνά χρησιμοποιούμενες μεθόδους αντικατάστασης που βασίζονται σε μετρήσεις μήκους και συχνότητας λέξης. Παρουσιάστηκαν επίσης τεχνικές για την εύρεση του καλύτερου υποψηφίου συνωνύμου σε μια δεδομένη ιεραρχία λέξεων, με βάση πληροφορίες σχετικά με τις σχετικές συχνότητες. Διαπιστώθηκε ότι όλες οι τεχνικές αντικατάστασης συνωνύμων, συμπεριλαμβανομένων των βασικών μεθόδων, αύξησαν τις σχετικές συχνότητες. Ο αλγόριθμος FM στόχευε να ανταμείψει τις υψηλές σχετικές συχνότητες ενώ δεν ανέβαινε πολύ ψηλά τη λέξη ιεραρχία και φαίνεται να αποδίδει καλά σε σχέση με αυτά τα κριτήρια.

20. Unsupervised Neural Text Simplification

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, Karthik Sankaranarayanan

Η εργασία παρουσιάζει μια πρώτη προσπάθεια για την απλούστευση του κειμένου χωρίς επίβλεψη που βασίζεται μόνο σε σώματα κειμένου με τη χρήση νευρωνικών δικτύων. Το βασικό πλαίσιο αποτελείται από έναν κοινό κωδικοποιητή και ένα ζεύγος αποκωδικοποιητών προσοχής, που υποβοηθούνται σημαντικά από απώλειες και αποθρομβοποίηση που βασίζονται σε διακρίσεις. Το πλαίσιο εκπαιδεύεται χρησιμοποιώντας κείμενο χωρίς ετικέτα που συλλέγεται από την ένδειξη en-Wikipedia. Η ανάλυσή (τόσο ποσοτική όσο και ποιοτική που περιλαμβάνει ανθρώπινους αξιολογητές) σε δεδομένα δημοσίων δοκιμών δείχνει ότι το προτεινόμενο μοντέλο μπορεί να

εκτελέσει απλοποίηση κειμένου τόσο σε λεξιλογικό όσο και σε συντακτικό επίπεδο, ανταγωνιστικό με τις υπάρχουσες εποπτευόμενες μεθόδους. Επίσης, ξεπερνά τις βιώσιμες μη εποπτευόμενες βασικές γραμμές. Η προσθήκη μερικών ζευγών με ετικέτα συμβάλλει στη βελτίωση της απόδοσης περαιτέρω.

Το σύστημα είναι χτισμένο με βάση την αρχιτεκτονική του στυλ κωδικοποίησης-παρακολούθησης-αποκωδικοποίησης (Bahdanau et al., 2014) με αλγοριθμικές και αρχιτεκτονικές αλλαγές που εφαρμόζονται στο τυπικό μοντέλο. Συγκεντρώθηκαν σώματα που περιέχουν απλές και σύνθετες προτάσεις και χρησιμοποιήθηκαν για να εκπαιδευτεί ένα σύστημα που βασίζεται σε έναν κοινόχρηστο κωδικοποιητή και δύο αποκωδικοποιητές. Προτείνεται ένα νέο εκπαιδευτικό σχήμα που επιτρέπει στο μοντέλο να εκτελεί ταυτόχρονη μείωση περιεχομένου και λεξιλογική απλοποίηση μέσω των προτεινόμενων απωλειών και αποθορυβοποίησης. Πραγματοποιήθηκαν πειράματα για πολλές παραλλαγές του συστήματος, καθώς και για γνωστές μη εποπτευόμενες γραμμές βάσης και εποπτευόμενα συστήματα. Η ποιοτική και ποσοτική ανάλυση των αποτελεσμάτων για δεδομένα δοκιμών που είναι διαθέσιμα στο κοινό καταδεικνύει ότι τα μοντέλα, αν και χωρίς επίβλεψη, μπορούν να αποδώσουν καλύτερα από ή ανταγωνιστικά σε αυτές τις γραμμές βάσης.

21. A Multilingual Simplified Language News Corpus

Renate Hauser, Jannis Vamvas, Sarah Ebling, Martin Volk

Σε αυτό το άρθρο, παρουσιάζουμε το SNIML, ένα μεγάλο πολυγλωσσικό σώμα ειδήσεων σε απλοποιημένη γλώσσα. Το σώμα περιέχει 13 χιλιάδες απλοποιημένα άρθρα ειδήσεων γραμμένα σε μία από τις έξι γλώσσες: φινλανδικά, γαλλικά, ιταλικά, σουηδικά, αγγλικά και γερμανικά. Περιλαμβάνει συνολικά 13.400 δημοσιεύματα που δημοσιεύθηκαν μεταξύ 2003 και 2022. Όλα τα άρθρα μοιράζονται με ανοιχτές άδειες που επιτρέπουν την ακαδημαϊκή χρήση. Το επίπεδο απλοποίησης του κειμένου ποικίλλει ανάλογα με την πύλη ειδήσεων. Παρόλο που το SNIML δεν είναι μια παράλληλη βάση, μπορεί να είναι χρήσιμο ως συμπλήρωμα στα πιο ομοιογενή αλλά συχνά μικρότερες βάσεις ειδήσεων στην απλοποιημένη ποικιλία μιας γλώσσας που χρησιμοποιούνται αυτήν τη στιγμή.

Για να αποκτήσουν όλα τα δημοσιευμένα άρθρα, οι συγγραφείς ανέπτυξαν web scrapers για να εξετάσουν τις σελίδες αρχείων των παρόχων. Για αυτό, συλλέχθηκαν όλες οι διευθύνσεις URL των άρθρων. Εκτελώντας αιτήματα στις συλλεγμένες διευθύνσεις URL, έλαβαν τα αρχεία HTML των σελίδων των άρθρων, τα οποία στη συνέχεια αναλύθηκαν. Στην περίπτωση του Informazione Facile, ενήργησαν εξαγωγή βάσης δεδομένων σε μορφή RSS των συντακτών, επομένως δεν χρειαζόταν καμία ανάλυση ιστού. Το σώμα ήταν δομημένο σε XML. Κάθε άρθρο είναι αντιπροσωπευτικό και μπορούσε να αναγνωριστεί με ένα μοναδικό αναγνωριστικό.

Η μελέτη έδειξε ότι το σώμα SNIML είναι ικανό να χρησιμεύσει ως χρήσιμο συμπλήρωμα σε άλλους πόρους για απλοποιημένη γλώσσα. Για παράδειγμα, η Simple English Wikipedia έχει μεγαλώσει πολύ, αλλά περιορίζεται σε μία μόνο γλώσσα, στυλ κειμένου και επίπεδο απλοποίησης. Συμπληρώνοντας την Απλή Αγγλική Wikipedia, το σώμα SNIML θα μπορούσε έτσι να βελτιώσει την ποικιλομορφία και την πολυγλωσσία των μοντέλων γλώσσας για απλοποιημένη γλώσσα, καθώς και συστήματα αναγνώρισης για απλοποιημένη γλώσσα.

Επιπλέον, η πολύγλωσση σύνθεση του SNIML ανοίγει δυνατότητες για την αξιολόγηση της διαγλωσσικής μεταφοράς. Για παράδειγμα, ένα σύστημα για την αναγνώριση της απλοποιημένης γλώσσας θα μπορούσε να εκπαιδευτεί στο αγγλικό τμήμα και να αξιολογηθεί στις άλλες πέντε γλώσσες του σώματος. Επιπλέον, η χρονική διαστρωμάτωση του σώματος καθιστά δυνατή την αξιολόγηση ενός μοντέλου για απλοποιημένη γλώσσα σε έννοιες και θέματα που δεν φαίνονται κατά τη διάρκεια της εκπαίδευσης. Έχει αποδειχθεί στο παράδειγμα των τυπικών αγγλικών ότι η χρονική γενίκευση είναι μια πρόκληση για τα γλωσσικά μοντέλα και οι συγγραφείς πιστεύουν ότι θα μπορούσε ακόμη περισσότερο να είναι μια πρόκληση για την απλοποιημένη γλώσσα.

22.Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL

M. Zakaria Kurdi

Ο στόχος αυτής της εργασίας είναι να δημιουργήσει έναν ταξινομητή που μπορεί να προσδιορίσει την πολυπλοκότητα του κειμένου στο πλαίσιο της διδασκαλίας της ανάγνωσης σε μαθητές της αγγλικής ως δεύτερης γλώσσας (ESL). Για να παρουσιαστούν στους μαθητές γλωσσών κείμενα

κατάλληλα για το επίπεδο της αγγλικής τους, ένα σύνολο χαρακτηριστικών που μπορούν να περιγράψουν τη φωνολογική, μορφολογική, λεξιλογική, συντακτική, λεκτική και ψυχολογική πολυπλοκότητα ενός κειμένου εντοπίστηκαν και χρησιμοποιήθηκαν για την ταξινόμηση κειμένων. Χρησιμοποιώντας ένα σύνολο 6171 κειμένων, τα οποία είχαν ήδη ταξινομηθεί σε τρία διαφορετικά επίπεδα δυσκολίας από ειδικούς της ESL, διεξήχθησαν διαφορετικά πειράματα με πέντε αλγόριθμους Μηχανικής Μάθησης. Τα αποτελέσματα έδειξαν ότι τα υιοθετημένα γλωσσικά χαρακτηριστικά παρέχουν καλή συνολική απόδοση ταξινόμησης (F-Score = 0,97). Διεξήχθη μια αξιολόγηση επεκτασιμότητας για να ελεγχθεί εάν ένας τέτοιος ταξινομητής θα μπορούσε να χρησιμοποιηθεί σε πραγματικές εφαρμογές, όπου μπορεί να συνδεθεί σε μια μηχανή αναζήτησης ή σε μια ενότητα απόξεσης ιστού. Σε αυτήν την αξιολόγηση, τα κείμενα στο σετ δοκιμών δεν διαφέρουν μόνο από αυτά του σετ εκπαίδευσης, αλλά και διαφορετικών τύπων (κείμενα ESL σε σύγκριση με την αναγνωστική αποτελεσματικότητα παιδιών που διαβάζουν κείμενα). Αν και η συνολική απόδοση του ταξινομητή μειώθηκε σημαντικά (F-Score = 0,65), ο πίνακας σύγκρισης δείχνει ότι τα περισσότερα από τα σφάλματα ταξινόμησης είναι μεταξύ των κατηγοριών δύο και τρία (οι τάξεις μεσαίου επιπέδου). Ομοίως, ο πίνακας σύγκρισης δείχνει επίσης ότι το σύστημα έχει μια ισχυρή απόδοση στην κατηγοριοποίηση κειμένων της κατηγορίας 1 και τέταρτης τάξης. Αυτή η συμπεριφορά μπορεί να εξηγηθεί από τη διαφορά στα κριτήρια ταξινόμησης μεταξύ των δύο σωμάτων. Ως εκ τούτου, τα παρατηρούμενα αποτελέσματα επιβεβαιώνουν τη χρηστικότητα ενός τέτοιου ταξινομητή σε μια εφαρμογή σε πραγματική τάξη.

23. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?

Zarah Weiss, Detmar Meurers

Η εργασία παρουσιάζει ένα νέο μοντέλο αξιολόγησης αναγνωσιμότητας προτάσεων τελευταίας τεχνολογίας για Γερμανούς αναγνώστες L2. Δημιουργήθηκε ένα μοντέλο μηχανικής μάθησης με ευρεία πληροφόρηση γλωσσικά και συγκριθεί η απόδοσή του με τέσσερις τύπους αναγνωσιμότητας που χρησιμοποιούνται συνήθως. Για να κατανοηθεί πότε οι γλωσσικές γνώσεις που χρησιμοποιούνται για την ενημέρωση του μοντέλου κάνουν τη διαφορά για την αξιολόγηση αναγνωσιμότητας και πότε αρκούν απλοί τύποι αναγνωσιμότητας, έγινε σύγκριση της απόδοσής

τους με βάση δύο κοινές εργασίες αξιολόγησης αυτόματης αναγνωσιμότητας, την προγνωστική παλινδρόμηση και την κατάταξη ζευγών προτάσεων. Διαπιστώθηκε ότι η αξιοποίηση γλωσσικών γνώσεων αποφέρει κορυφαίες επιδόσεις σε όλες τις εργασίες, αλλά ότι για τον προσδιορισμό απλοποιημένων προτάσεων, οι τύποι αναγνωσιμότητας – που είναι ευκολότεροι στον υπολογισμό και πιο προσβάσιμοι – μπορούν να είναι αρκετά ακριβείς. Η γλωσσικά ενημερωμένη μοντελοποίηση, ωστόσο, είναι η μόνη βιώσιμη επιλογή για αποτελέσματα υψηλής ποιότητας σε εργασίες πρόβλεψης. Στη συνέχεια, διερευνήθηκε το προφίλ αναγνωσιμότητας από άποψη προτάσεων των κειμένων που γράφτηκαν για μαθητές ξένων γλωσσών σε αρχικό, μεσαίο και προχωρημένο επίπεδο της γερμανικής γλώσσας.

Το σώμα TextCoplexityDE αποτελείται από 1.119 προτάσεις. Εξήχθησαν 1.019 προτάσεις από 23 άρθρα της Wikipedia που σχετίζονται με την ιστορία, την κοινωνία ή την επιστήμη και 100 προτάσεις από δύο άρθρα στο Leichte Sprache (αγγλ. «απλή γλώσσα»). Όλα βαθμολογήθηκαν από 267 Γερμανούς μαθητές L2 σε τρεις ξεχωριστές διαστάσεις που ορίζονται από τους Naderi et al. (2019a): αναγνωσιμότητα, κατανοητή διατύπωση και λεξιλογική δυσκολία. Για κάθε διάσταση, οι προτάσεις βαθμολογήθηκαν από έως και δέκα μαθητές σε μια κλίμακα Likert 7 βαθμών. Αυτές οι αξιολογήσεις συγκεντρώθηκαν σε μια ενιαία Μέση βαθμολογία γνώμης (MOS).

Για την εξαγωγή χαρακτηριστικών, οι συγγραφείς χρησιμοποίησαν το σύστημα CTAP το οποίο έχει επεκταθεί για να διευκολύνει την ανάλυση της γερμανικής γλώσσας από τους Weiss et al. (2021). Οι συγγραφείς επέλεξαν αυτό το σύστημα, επειδή είναι εξ όσων γνωρίζουν το πιο εκτεταμένο διαθέσιμο σύστημα ανάλυσης για τα γερμανικά. Η υποκείμενη μηχανή εξαγωγής χαρακτηριστικών για τα γερμανικά έχει αποδειχθεί εξαιρετικά επιτυχημένη και ισχυρή σε μια ποικιλία εργασιών που σχετίζονται με την εκπαίδευση, όπως αξιολόγηση αναγνωσιμότητας και εργασία που συνδέεται με αξιολόγηση ποιότητας γραφής. Επίσης, η χρήση ενός δημοσίως διαθέσιμου συστήματος βασισμένου στο διαδίκτυο αυξάνει την επαναχρησιμοποίηση οποιουδήποτε μοντέλου που χρησιμοποιεί αυτές τις δυνατότητες στην πράξη.

Οι συγγραφείς προγραμματίσαν και συνέκριναν αρκετούς αλγόριθμους μηχανικής μάθησης χρησιμοποιώντας 10πλάσια διασταυρούμενη επικύρωση (10 CV) και τους μετασχηματισμούς z των 373 χαρακτηριστικών. Επέλεξαν αυτούς τους αλγόριθμους με βάση τη χρήση τους σε προηγούμενη έρευνα ή την ανθεκτικότητά τους έναντι μεγάλων συνόλων χαρακτηριστικών με

πολλαπλή συγγραμμικότητα. Το Bayesian Ridge Regression ξεπέρασε τις επιδόσεις των άλλων μοντέλων. Για να αξιολογηθεί η συνολική απόδοση αυτού του μοντέλου που βασίζεται στην πολυπλοκότητα, υπολογίσαμε τη συσχέτιση κατάταξης RMSE και Spearman (rs) κατά τη διάρκεια 10 CV, τέθηκε σε σύγκριση με την τρέχουσα απόδοση SOTA στα δεδομένα (RMSE= 0,847).

Και στα δύο βήματα αξιολόγησης, συγκρίθηκε η απόδοση του CBM με πέντε εναλλακτικά μοντέλα. Εκπαιδεύθηκε ένα μοντέλο παλινδρόμησης Bayesian Ridge χρησιμοποιώντας μόνο μέτρα μήκους επιφάνειας ως προγνωστικά ως βάση (εφεξής: μοντέλο βάσει μήκους ή LBM): ο δείκτης αναγνωσιμότητας Amstad, ο δείκτης αναγνωσιμότητας Erste Wiener Sachtextformel, ο δείκτης αναγνωσιμότητας LIX, ο δείκτης αναγνωσιμότητας EFL Miyazaki. Τα ευρήματά υπογραμμίζουν ότι η αναγνωσιμότητα ενός κειμένου καθορίζεται από τη μέγιστη και όχι από τη συνολική αναγνωσιμότητα των προτάσεων. Αυτό έχει άμεσες συνέπειες για την προσαρμογή του εκπαιδευτικού υλικού και δείχνει τη σημασία της μελέτης της αναγνωσιμότητας με βάση το επίπεδο του εγγράφου.

24. Subjective Text Complexity Assessment for German

Laura Seiffe, Fares Kallel, Babak Naderi, Sebastian Moller, Roland Roller

Σε αυτό το άρθρο, οι συγγραφείς περιγράφουν υποκειμενικά πειράματα για να αξιολογήσουν την αναγνωσιμότητα του γερμανικού κειμένου. Οι συγγραφείς συνθέτουν ένα νέο σύνολο προτάσεων που παρέχονται από έναν γερμανικό πάροχο υπηρεσιών πληροφορικής. Οι προτάσεις σχολιάζονται με τις υποκειμενικές αξιολογήσεις πολυπλοκότητας από δύο ομάδες συμμετεχόντων, δηλαδή ειδικούς και μη ειδικούς για αυτόν τον τομέα κειμένου. Στη συνέχεια, οι συγγραφείς εξάγουν ένα εκτεταμένο σύνολο χαρακτηριστικών με γλωσσικά κίνητρα που υποτίθεται ότι αλληλεπιδρούν με την αντίληψη της πολυπλοκότητας. Οι συγγραφείς δείχνουν ότι ένα μοντέλο γραμμικής παλινδρόμησης με ένα υποσύνολο αυτών των χαρακτηριστικών μπορεί να είναι ένας πολύ καλός προγνωστικός δείκτης της πολυπλοκότητας του κειμένου.

Το σύνολο δεδομένων αποτελείται από κείμενο που παρέχεται από τον συνεργάτη του έργου μας, έναν Γερμανό πάροχο υπηρεσιών πληροφορικής στο πλαίσιο Γερμανών φορολογικών συμβούλων, ελεγκτών και δικηγόρων. Αυτό το κείμενο χωρίζεται σε προτάσεις και προέρχεται

από οδηγίες, σχόλια και περιγραφές που απευθύνονται σε υπαλλήλους του παρόχου υπηρεσιών, καθώς και σε εξωτερικούς χρήστες του συστήματος. διεξήγαγαν δύο πειράματα για να σχολιάσουν το σώμα με υποκειμενικές αξιολογήσεις αναγνωσιμότητας - ένα από μη ειδικούς που χρησιμοποιούν crowdsourcing και ένα από εξειδικευμένο προσωπικό του παρόχου υπηρεσιών πληροφορικής. διεξήγαγαν επίσης το πείραμα με μη ειδικούς σε τομείς χρησιμοποιώντας την πλατφόρμα crowdsourcing Clickworkers, μια πλατφόρμα με μια αρκετά μεγάλη ομάδα γερμανόφωνων εργαζομένων στο πλήθος. Το crowdsourcing παρέχει πρόσβαση σε μια μεγάλη κλίμακα γεωγραφικά κατανεμημένων ομάδων εργαζομένων που μπορούν να λάβουν μέρος σε ένα υποκειμενικό τεστ.

Ανέπτυξαν μια διοχέτευση βασισμένη σε python για την εξαγωγή όλων των γλωσσικών χαρακτηριστικών για κάθε πρόταση στο σύνολο δεδομένων. Όλα τα χαρακτηριστικά υπολογίζονται και αποθηκεύονται ως αριθμητικά δεδομένα πίνακα για περαιτέρω ανάλυση. Για το σκοπό αυτό χρησιμοποιούν κυρίως τη βιβλιοθήκη επεξεργασίας φυσικής γλώσσας Spacy, η οποία παρέχει έναν αριθμό προεκπαιδευμένων αγωγών. Η επισήμανση των γερμανικών κατηγοριών λέξεων στο Spacy βασίζεται στο σύνολο Tagset της Στουτγάρδης-Tubingen. Χρησιμοποιώντας αυτές τις ετικέτες, κατόρθωσαν να μετρήσουν τις εμφανίσεις διαφορετικών τμημάτων της ομιλίας και να τις χρησιμοποιήσουμε για να εξαγάγουμε σημαντικά χαρακτηριστικά. Τα αποτελέσματα έδειξαν ότι η τεχνογνωσία των συμμετεχόντων έχει σημαντική κύρια επίδραση στις αξιολογήσεις τους. Απροσδόκητα, οι ειδικοί τείνουν να βαθμολογούν τις προτάσεις ως πιο περίπλοκες από την ομάδα μη ειδικών.

25. A Benchmark for Neural Readability Assessment of Texts in Spanish

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Esteban Morales-Esquivel, Fernando Alva-Manchego

Οι συγγραφείς δημοσιεύουν ένα νέο σημείο αναφοράς για την Αυτόματη Αξιολόγηση Αναγνωσιμότητας (ARA) κειμένων στα ισπανικά. Συνδυάσαμε τα υπάρχοντα σώματα με κατάλληλα κείμενα που συλλέγονται από τον Ιστό, δημιουργώντας έτσι το μεγαλύτερο διαθέσιμο σύνολο δεδομένων για το ARA ισπανικών κειμένων. Όλα τα δεδομένα υποβλήθηκαν σε προεπεξεργασία και κατηγοριοποίηση για να επιτρέπεται ο πειραματισμός με μοντέλα ARA που

κάνουν προβλέψεις σε δύο (απλά και σύνθετα) ή τρία (βασικό, ενδιάμεσο και προχωρημένο) επίπεδα αναγνωσιμότητας και σε δύο ευκρίνεια κειμένου (παραγράφους και προτάσεις). Μια ανάλυση που βασίζεται σε δείκτες αναγνωσιμότητας δείχνει ότι οι προτεινόμενες ομαδοποιήσεις συνόλων δεδομένων μας είναι κατάλληλες για το καθορισμένο επίπεδο αναγνωσιμότητας. Χρησιμοποιούμε το σημείο αναφοράς μας για να εκπαιδεύσουμε νευρωνικά μοντέλα ARA που βασίζονται στο BERT σε ρυθμίσεις μηδενικής λήψης, λίγων λήψεων και διαγλωσσικών. Τα αποτελέσματα δείχνουν ότι είτε ένα μονόγλωσσο είτε ένα πολύγλωσσο προεκπαιδευμένο μοντέλο μπορεί να επιτύχει καλά αποτελέσματα όταν συντονίζεται με ακρίβεια σε δεδομένα που αφορούν τη γλώσσα. Επιπλέον, όλα τα μοντέλα μειώνουν την απόδοσή τους όταν προβλέπουν τρεις κατηγορίες αντί για δύο, δείχνοντας ευκαιρίες για την ανάπτυξη καλύτερων μοντέλων ARA για τα Ισπανικά με τους υπάρχοντες πόρους.

26.Text Complexity Classification Data Mining Model Based on Dynamic Quantitative Relationship between Modality and English Context

Dan Zhang

Αυτό το άρθρο ξεκινά αρχικά με τη θεωρητική έρευνα της ανάλυσης πολυπλοκότητας κειμένου και αναλύει την πηγή της πολυπλοκότητας του κειμένου και τα πέντε χαρακτηριστικά του, δυναμική, πολυπλοκότητα, απόκρυψη, συναίσθημα και ασάφεια, σε συνδυασμό με την έκφραση των αναγκών των χρηστών στο περιβάλλον δικτύου. Δεύτερον, με βάση τη συγκεκριμένη διαδικασία εξόρυξης κειμένου, δηλαδή τη συλλογή δεδομένων, την επεξεργασία δεδομένων και την οπτικοποίηση δεδομένων, προτείνεται να υποδιαιρεθεί η ανάλυση της ζήτησης των χρηστών σε τρία στάδια απόκτησης, αναγνώρισης και έκφρασης της πολυπλοκότητας κειμένου, ώστε να επιτευχθεί πολυπλοκότητα κειμένου. ανάλυση βασισμένη στην τεχνολογία εξόρυξης κειμένου. Μετά, με βάση την υπολογιστική γλωσσολογία και τη μαθηματική-στατιστική ανάλυση, σε συνδυασμό με τη μηχανική μάθηση και την τεχνολογία ανάκτησης πληροφοριών, το κείμενο σε οποιαδήποτε μορφή μετατρέπεται σε μορφή περιεχομένου που μπορεί να χρησιμοποιηθεί για μηχανική μάθηση και προκύπτουν μοτίβα ή γνώσεις από αυτήν τη μορφή περιεχομένου. Στη συνέχεια, μέσω της σύγκρισης και της έρευνας της τεχνολογίας εξόρυξης κειμένου, σε συνδυασμό με το μοντέλο ιεραρχικής δομής ανάλυσης πολυπλοκότητας κειμένου, προτείνεται ένα πλαίσιο ανάλυσης ποσοτικής πολυπλοκότητας σχέσεων βασισμένο στην τεχνολογία εξόρυξης κειμένου,

το οποίο ενσωματώνεται στη χρήση τεχνολογίας ανίχνευσης ιστού. Τα πειραματικά αποτελέσματα δείχνουν ότι οι συλλεγόμενες πληροφορίες ποσοτικής σχέσης προσδιορίζονται και εκφράζονται προκειμένου να πραγματοποιηθεί η μετατροπή των ποσοτικών πληροφοριών σχέσεων σε χαρακτηριστικά προϊόντος. Τα δεδομένα αγοράς και τα δεδομένα κειμένου μπορούν να ενσωματωθούν για να βοηθήσουν στη βελτίωση της απόδοσης του μοντέλου και η χρήση δεδομένων κειμένου μπορεί να βελτιώσει περαιτέρω τις προβλέψεις για ακρίβεια.

27. Automatic Classification of Text Complexity

Valentino Santucci, Filippo Santarelli, Luciana Forti and Stefania Spina

Αυτή η εργασία εισάγει ένα αυτόματο σύστημα ταξινόμησης για τη μέτρηση του επιπέδου πολυπλοκότητας ενός δεδομένου ιταλικού κειμένου υπό γλωσσική άποψη. Το έργο της μέτρησης της πολυπλοκότητας ενός κειμένου απορρέει από ένα εποπτευόμενο πρόβλημα ταξινόμησης με την εκμετάλλευση ενός συνόλου κειμένων που παράγονται σκόπιμα από ειδικούς γλωσσολόγους για σκοπούς διδασκαλίας και αξιολόγησης δεύτερης γλώσσας. Τα κοινά υιοθετούμενα επίπεδα Κοινού Ευρωπαϊκού Πλαισίου Αναφοράς για τις Γλώσσες (CEFR) χρησιμοποιήθηκαν ως τάξεις ταξινόμησης στόχου, τα κείμενα επεξεργάστηκαν λαμβάνοντας υπόψη ένα μεγάλο σύνολο αριθμητικών γλωσσικών χαρακτηριστικών και πραγματοποιήθηκε μια πειραματική σύγκριση μεταξύ δέκα ευρέως χρησιμοποιούμενων μοντέλων μηχανικής μάθησης. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη προσέγγιση είναι σε θέση να επιτύχει καλή ακρίβεια πρόβλεψης, ενώ πραγματοποιήθηκε περαιτέρω ανάλυση προκειμένου να εντοπιστούν οι κατηγορίες χαρακτηριστικών που επηρέασαν τις προβλέψεις.

28. Evaluating Factuality in Text Simplification

Ashwin Devaraj, William Sheffield, Byron C. Wallace, Junyi Jessy Li

Τα μοντέλα αυτοματοποιημένης απλοποίησης στοχεύουν να κάνουν τα κείμενα εισόδου πιο ευανάγνωστα. Τέτοιες μέθοδοι έχουν τη δυνατότητα να κάνουν σύνθετες πληροφορίες προσβάσιμες σε ένα ευρύτερο κοινό, π.χ., παρέχοντας πρόσβαση στην πρόσφατη ιατρική βιβλιογραφία, η οποία διαφορετικά θα ήταν αδιαπέραστη για έναν απλό αναγνώστη. Ωστόσο, τέτοια μοντέλα κινδυνεύουν να εισάγουν σφάλματα σε αυτόματα απλουστευμένα κείμενα, για

παράδειγμα με την εισαγωγή δηλώσεων που δεν υποστηρίζονται από το αντίστοιχο αρχικό κείμενο ή με την παράλειψη βασικών πληροφοριών. Η παροχή πιο ευανάγνωστων αλλά ανακριβών εκδόσεων κειμένων μπορεί σε πολλές περιπτώσεις να είναι χειρότερη από το να μην παρέχεται καθόλου τέτοια πρόσβαση. Το πρόβλημα της πραγματικής ακρίβειας (και η έλλειψή του) έχει λάβει μεγαλύτερη προσοχή στο πλαίσιο των μοντέλων περίληψης, αλλά η πραγματικότητα των αυτόματα απλουστευμένων κειμένων δεν έχει διερευνηθεί. Η εν λόγω έρευνα εισάγει μια ταξινόμηση σφαλμάτων που χρησιμοποιήθηκε για την ανάλυση τόσο των παραπομπών που προέρχονται από τυπικά σύνολα δεδομένων απλοποίησης όσο και των εξερχόμενων μοντέλων τελευταίας τεχνολογίας. Διαπιστώθηκε ότι συχνά εμφανίζονται σφάλματα και στα δύο που δεν καταγράφονται από τις υπάρχουσες μετρήσεις αξιολόγησης, παρακινώντας την ανάγκη για έρευνα για τη διασφάλιση της πραγματικής ακρίβειας των αυτοματοποιημένων μοντέλων
απλοποίησης.

Οι συγγραφείς σχολίασαν παραδείγματα από τα σύνολα δεδομένων Wikilarge και Newsela. Αυτά χρησιμοποιούνται συνήθως στη βιβλιογραφία, και έτσι έχουν αναφερθεί αποτελέσματα σε αυτά τα σώματα για μια διαφορετική συλλογή μοντέλων. Το Wikilarge περιλαμβάνει 296.000 χονδρικά ευθυγραμμισμένα ζεύγη προτάσεων από την αγγλική Wikipedia και την απλή αγγλική Wikipedia. Το Newsela αποτελείται από 96.000 ζεύγη προτάσεων που εξάγονται από ένα σύνολο δεδομένων ειδήσεων που έχουν ξαναγραφτεί σε 4 επίπεδα ανάγνωσης από επαγγελματίες. Για να κάνουν την ανάλυση εφικτή σε αυτό το έργο, οι συγγραφείς εξετάζουν το απλούστερο επίπεδο για το Newsela.

Οι συγγραφείς σχολίασαν τα αποτελέσματα που δημιουργήθηκαν από μια συλλογή μοντέλων στα ίδια παραδείγματα επικύρωσης και δοκιμής από το Wikilarge και το Newsela, αντίστοιχα. Οι συγγραφείς επέλεξαν ένα σύνολο μοντέλων που προορίζονται να είναι αντιπροσωπευτικά διαφορετικών αρχιτεκτονικών και μεθόδων εκπαίδευσης. Πιο συγκεκριμένα, για μοντέλα που βασίζονται σε RNN Οι συγγραφείς εξέτασαν το Dress και EditNTS. Το Dress είναι ένα μοντέλο LSTM που εκπαιδεύεται χρησιμοποιώντας το REINFORCE για την ελαχιστοποίηση μιας συνάρτησης ανταμοιβής που αποτελείται από τη διατήρηση του νοήματος, απλότητα και όρους ευχέρειας. Το EditNTS αντιπροσωπεύει κάθε ζεύγος προτάσεων ως μια ακολουθία πράξεων επεξεργασίας και μαθαίνει απευθείας αυτές τις πράξεις για να κάνει απλοποίηση. Για

αρχιτεκτονικές που βασίζονται σε μετασχηματιστές αξιολογήθηκαν δύο μοντέλα που προτάθηκαν προηγουμένως:

Πρόσβαση και ControlTS. Η Access εκπαιδεύει έναν τυχαία αρχικοποιημένο μετασχηματιστή για να δημιουργήσει απλουστεύσεις παραμετροποιημένες από διακριτικά ελέγχου που επηρεάζουν χαρακτηριστικά όπως η λεξιλογική πολυπλοκότητα και η συμπίεση μήκους. Το ControlTS είναι μια υβριδική μέθοδος που δημιουργεί υποψηφίους απλοποίησης χρησιμοποιώντας γραμματικούς κανόνες και στη συνέχεια εφαρμόζει ένα μοντέλο παράφρασης που βασίζεται στο BERT. Επιπλέον, βελτιστοποιήθηκε επίσης το T5 για την εργασία απλούστευσης, που περιγράφεται λεπτομερώς στο Παράρτημα Α. Το T5 είναι ένα μοντέλο που βασίζεται σε μετασχηματιστή, προεκπαιδευμένο από κοινού τόσο σε στόχους μοντελοποίησης γλώσσας χωρίς επίβλεψη όσο και σε μια σειρά εποπτευόμενων εργασιών, συμπεριλαμβανομένης της σύνοψης και της μετάφρασης, όλα πλαισιωμένα ως κείμενο- προβλήματα σε κείμενο. Οι συγγραφείς διαπίστωσαν ότι τα σφάλματα εμφανίζονται συχνά τόσο στις αναφορές όσο και στις παραγόμενες εξόδους. Στα σύνολα δεδομένων, τα σφάλματα διαγραφής είναι αρκετά συχνά, με το Newsela να περιέχει περισσότερα από το Wikilarge. Οι εξοδοί του συστήματος υποδεικνύουν ότι τα μοντέλα τείνουν επίσης να διαγράφουν πληροφορίες, κάτι που είναι πιθανώς μια συμπεριφορά που μαθαίνεται από τα δεδομένα εκπαίδευσης. Τα αποτελέσματα του μοντέλου περιέχουν περισσότερα σφάλματα υποκατάστασης από τα σύνολα δεδομένων, έτσι ώστε η συμπεριφορά είναι πιθανώς μια μεροληψία μοντέλου και όχι κάτι που προκύπτει από τα δεδομένα.

4.2. Μεθοδολογία Έρευνας

Σκοπός της εργασίας είναι να διερευνήσει τον τρόπο με τον οποίο οι αλγόριθμοι μηχανικής μάθησης απλοποιούν επιστημονικά κείμενα, ώστε απλοί φυσικοί ομιλητές που δεν έχουν γνώση του αντικειμένου να διαβάζουν και κατανοούν επιστημονικά κείμενα διαφορετικών επιπέδων αναγνωσιμότητας σε ψηφιακή μορφή.

4.2.1 Ερευνητική στρατηγική

Οι βιβλιογραφικές ανασκοπήσεις συνιστούν ερευνητικές μεθόδους που συμβάλλουν στην ανάπτυξη κάθε επιστημονικού κλάδου, ανεξαρτήτως αντικειμένου, καθώς παρέχει μια κριτική πρόσβαση σε προηγούμενα επιστημονικά δεδομένα πάνω σε μια συγκεκριμένη θεματική. Σε σύγκριση με τις παραδοσιακές επισκοπήσεις της βιβλιογραφίας, οι οποίες συχνά αφήνουν πολλά στην τεχνογνωσία των συγγραφέων, οι συστηματικές βιβλιογραφικές ανασκοπήσεις αντιμετωπίζουν τη διαδικασία ανασκόπησης της βιβλιογραφίας σαν μια επιστημονική διαδικασία και εφαρμόζουν έννοιες εμπειρικής έρευνας προκειμένου να καταστήσουν τη διαδικασία ανασκόπησης πιο διαφανή και πιο έγκυρη και να μειώσουν την πιθανότητα μεροληψίας. Οι συστηματικές βιβλιογραφικές ανασκοπήσεις έχουν καταστεί ένα καίριο εργαλείο ερευνητικής μεθοδολογίας, έχοντας αναπτύξει μια συγκεκριμένη υποδομή για τη διεξαγωγή αυτών των ανασκοπήσεων και τη συνεχή βελτίωση της μεθόδου για την αντιμετώπιση νέων ερευνητικών ερωτημάτων. Ορισμένοι συγγραφείς στο κίνημα της «επιστήμης σχεδίασης» στην έρευνα διαχείρισης προτείνουν ότι οι επιστήμονες σχεδιασμού θα πρέπει να χρησιμοποιήσουν αυτήν την προσέγγιση για να αναπτύξουν προτάσεις σχεδίασης βασισμένες σε συστηματικές ανασκοπήσεις εμπειρικών στοιχείων (Tranfield et al., 2003 van Aken and Romme, 2009).

Άλλοι συγγραφείς έχουν παραπονεθεί για την περιορισμένη χρήση της μεθόδου στην έρευνα σχεδιασμού, καθώς εμποδίζει την ικανότητά τους να εξελίξουν περαιτέρω την έρευνά τους συγκεντρώνοντας και συνθέτοντας τα αποτελέσματά μας. Ωστόσο, δεν υπάρχει καθοδήγηση σχετικά με τον τρόπο εκτέλεσης αυτών των αναθεωρήσεων και η μέθοδος δεν αποτελεί μέρος της παραδοσιακής εργαλειοθήκης έρευνας σχεδιασμού (Cash, 2018). Τούτο δε καθώς η συστηματική βιβλιογραφική επισκόπηση είναι ένας τρόπος σύνθεσης επιστημονικών στοιχείων για την απάντηση σε ένα συγκεκριμένο ερευνητικό ερώτημα με τρόπο που να είναι διαφανής και δυνάμενος να αναπαραχθεί, επιδιώκοντας ταυτόχρονα τη συμπερίληψη στην επιστημονική βάση δεδομένων του, όλων των δημοσιευμένων στοιχείων για το θέμα καθώς και μια αξιολόγηση της ποιότητας αυτών των στοιχείων. Ο κύριος στόχος της προσέγγισης των συστηματικών βιβλιογραφικών επισκοπήσεων είναι να μειώσει τον κίνδυνο μεροληψίας και να αυξήσει τη διαφάνεια σε κάθε στάδιο της διαδικασίας αναθεώρησης, επιτρέποντας στον εκάστοτε ερευνητή να βασιστεί για την έκδοση των επιστημονικών του αποτελεσμάτων σε σαφείς, συστηματικές

μεθόδους αποσκοπώντας στη μείωση της μεροληψίας κατά την επιλογή και τη συμπερίληψη μελετών, στην αξιολόγηση της ποιότητας των περιεχομένων μελετών και σε μια αντικειμενική σύνοψη των διαθέσιμων ερευνητικών δεδομένων (Liberati et al., 2009).

4.2.2 Καθορισμός κριτηρίων εισόδου και αποκλεισμού

Με τον καθορισμό των κριτηρίων εισόδου και αποκλεισμού των ερευνών που έχουν διενεργηθεί και δημοσιευθεί παγκοσμίως επιτυγχάνεται η ομαλή και η μεθοδολογική αρτιότητα της παρούσας διπλωματικής εργασίας, καθώς έτσι είναι δυνατή η αντιπαράβολή των ευρημάτων αυτών με τα ερευνητικά ερωτήματα που έχουν οδηγήσει στην εκπόνηση της παρούσας και τα ευρήματα των μελετητών παγκοσμίως. Για το λόγο αυτό και για να εξασφαλιστεί η ποιοτική ολοκλήρωση της διαδικασίας είναι ζωτικής σημασίας η ανάγκη καθορισμού εκ των προτέρων των κριτηρίων με βάση τα οποία θα επιλέγονται ή θα απορρίπτονται οι μελέτες. Ειδικότερα, η ποιοτική αναζήτηση και αξιοποίηση των σύγχρονων μεθόδων στοχεύει στην ανεύρεση επιστημονικών δημοσιεύσεων στην αγγλική γλώσσα, καθώς αυτά είναι κατανοητά από το σύνολο της επιστημονικής κοινότητας και με αυτό τον τρόπο αξιολογούνται και ανανεώνονται τάχιστα. Έτσι, αυτό θα αποτελεί και κριτήριο επιλογής των αποτελεσμάτων ενώ εάν εξάγονται από την αναζήτηση κείμενα συνεδρίων ή οποιασδήποτε μορφή δημοσίευσης διαφορετικής από την αποδεκτή καθώς και σε γλώσσα διαφορετική από τα αγγλικά θα αποκλείονται κατά τη διαλογή. Το χρονικό εύρος το οποίο οριοθετεί την αναζήτηση πρόκειται να είναι τα τελευταία 15 έτη, περιλαμβάνοντας άρθρα και βιβλία στα αγγλικά ώστε να διαφυλαχθεί κατά το δυνατόν ότι η ισχύς και η αξιοπιστία των αποτελεσμάτων.

Επίσης, αναφορικά με τη θεματολογία, για την είσοδο μίας μελέτης-άρθρου στην παρούσα ανασκόπηση χρειάζεται να πραγματεύεται αυστηρά και να καταλήγει σε συμπέρασμα που να σχετίζεται και να απαντά στο καθορισμένο ερευνητικό ερώτημα. Πιο συγκεκριμένα, θα επιλέγονται άρθρα και βιβλία τα οποία ερευνούν την εργασιακή αποκατάσταση και απασχόληση ατόμων με νοητικές και αναπτυξιακές αναπηρίες και θα καταλήγουν σε συμπέρασμα σχετικό με την ένταξη και την αξιοποίηση τους στο εργασιακό περιβάλλον. Τέλος, δεν έχει τεθεί κάποιο κριτήριο για την περιοχή – χώρα που αναφέρεται το άρθρο ή το βιβλίο, με σκοπό να εξεταστεί το φαινόμενο αυτό σε παγκόσμια κλίμακα.

4.2.3 Διαδικασία Αναζήτηση άρθρων

Για την εντόπιση κατάλληλων άρθρων πρόκειται να ερευνηθούν οι βάσεις δεδομένων Google Scholar και Researchgate. Η αναζήτηση στις βάσεις δεδομένων οργανώθηκε χρησιμοποιώντας σχετικές λέξεις-κλειδιά στην αγγλική γλώσσα και πραγματοποιήθηκε κατά το χρονικό διάστημα μεταξύ 15 Νοεμβρίου 2023 και 20 Φεβρουαρίου 2024, ενώ ενημερώθηκε οριστικά με τελευταία εφαρμογή στις 13 Μαρτίου 2024. Προκειμένου να συμπεριληφθούν στην αναζήτηση οι διαφορετικοί αλλά παρεμφερείς σημασιολογικά όροι που χρησιμοποιήθηκαν για το συγκεκριμένο θέμα, αλλά και να επιτευχθεί αποτελεσματικότερα η έρευνα των σχετικών άρθρων αξιοποιήθηκαν οι τελεστές συσχετισμού (AND / OR / NOT).

Ειδικότερα, ο αρχικός όρος του ερευνητικού ερωτήματος που συμπεριλήφθηκε στην αναζήτηση αφορούσε τα ιδιαίτερα χαρακτηριστικά των ατόμων με ειδικές νοητικές και αναπτυξιακές αναπηρίες και στα αγγλικά οι συνηθέστερες λέξεις που χρησιμοποιούνται για αυτές τις έννοιες είναι «text simplification of scientific research». Προκειμένου να συμπεριληφθούν οι διαφορετικές διατυπώσεις, αριθμοί αλλά και τα απαρέμφατα και οι μετοχές με τις οποίες μπορεί να εμφανίζονται στον τίτλο ενός άρθρου προκύπτουν οι εξής λέξεις κλειδιά: *intell** και *develop**, καθώς με το σύμβολο «*» αξιοποιείται στην αναζήτηση η εμφάνιση όλων των λέξεων αυτής της ρίζας. Στη συνέχεια, αξιοποιώντας τους τελεστές συσχέτισης AND ΚΑΙ OR συνδυάζονται οι ανωτέρω λέξεις κλειδιά κατάλληλα ώστε να ληφθούν τα περισσότερα σχετικά αποτελέσματα. Προκύπτει, λοιπόν, η εξής στρατηγική αναζήτησης:

(text) AND (simplification) AND (scientific research)

Εφαρμόζοντας το ανωτέρω κείμενο στο χωρίο αναζήτησης των Google Scholar και Researchgate, με την επιλογή για αναζήτηση στον τίτλο των δημοσιεύσεων καθώς και ορισμό του χρονικού διαστήματος δημοσίευσης μεταξύ των 2014 και 2023, εξάγονται 55 αποτελέσματα. Στη συνέχεια, με την εφαρμογή από το χωρίο των φίλτρων του περιορισμού για εμφάνιση μόνο άρθρων και βιβλίων – Articles and Books - καθώς και μόνο στην αγγλική γλώσσα τα αποτελέσματα περιορίζονται σε 30 άρθρα από τις δύο βάσεις δεδομένων.

4.2.4. Το εργαλείο του QuillBot

Το QuillBot είναι ένα προηγμένο εργαλείο γραφής τεχνητής νοημοσύνης (AI), το οποίο προσφέρει δυνατότητες όπως αναδιατύπωση/απλοποίηση προτάσεων, επανεγγραφή άρθρων και δημιουργία περιεχομένου που δημιουργείται από AI. Το QuillBot χρησιμεύει ως ένα πλήρες εργαλείο επεξεργασίας γλώσσας και παραγωγικότητας με τις λειτουργίες βελτίωσης γραμματικού ελέγχου και γραφής. Στοχεύει στο να εκπληρώσει μια ποικιλία απαιτήσεων γραφής, συμπεριλαμβανομένης της ακαδημαϊκής γραφής, του blogging και της διόρθωσης. Ως μοντέλο γλώσσας AI, το QuillBot βοηθά τους χρήστες να δημιουργούν καλά δομημένες προτάσεις και βελτιώνει τη συνολική ποιότητα γραφής.

Το QuillBot είναι ένα προηγμένο διαδικτυακό εργαλείο γραφής και παράφρασης που υποστηρίζεται από AI. Χρησιμοποιεί τεχνικές επεξεργασίας φυσικής γλώσσας αιχμής (NLP) για να βοηθήσει τους χρήστες να γράφουν καλύτερα, να βρίσκουν νέες φράσεις και να βελτιώνουν τη σαφήνεια και την ποιότητα του περιεχομένου τους. Η ικανότητα του QuillBot να παραφράζει αποτελεσματικά το κείμενο είναι ένα από τα βασικά χαρακτηριστικά του. Οι χρήστες μπορούν να εισαγάγουν μια λέξη, μια παράγραφο ή ακόμα και ένα ολόκληρο κείμενο και το QuillBot θα παράγει εναλλακτικές εκδόσεις διατηρώντας παράλληλα την αίσθηση του πρωτοτύπου.

Για να διασφαλιστεί ότι η έξοδος της παραφρασμένης πρότασης είναι συνεκτική και διατηρεί το επιδιωκόμενο νόημα, το QuillBot υπερβαίνει την αντικατάσταση βασικών συνωνύμων και χρησιμοποιεί τη γνώση των συμφραζομένων. Είναι ένα χρήσιμο εργαλείο για την αύξηση της ευχέρειας και της σαφήνειας της γραφής επειδή λαμβάνει υπόψη τη γραμματική, τη δομή της πρότασης και το πλαίσιο. Το QuillBot περιλαμβάνει πολλές λειτουργίες και δυνατότητες γραφής, όπως περίληψη, έλεγχο γραμματικής και προτάσεις λέξεων, εκτός από την παράφραση. Αυτά τα εργαλεία έχουν σχεδιαστεί για να βελτιώνουν τη διαδικασία γραφής συνολικά και να προσφέρουν στους χρήστες εποικοδομητική κριτική και προτάσεις για να βελτιώσουν τη δουλειά τους.

Για τη βοήθεια της παράφρασης και της γραφής, το QuillBot συνδυάζει μεθόδους NLP αιχμής με αλγόριθμους βαθιάς μάθησης. Πιο συγκεκριμένα:

Κατανόηση εισροών

Όταν οι χρήστες παρέχουν κείμενο στο QuillBot, αυτό αναλύει και κατανοεί την είσοδο χρησιμοποιώντας πολλαπλές τεχνικές NLP. Αποδομεί το κείμενο στις λέξεις που το αποτελούν, αποκαλύπτει τις συνδέσεις μεταξύ τους και αποτυπώνει το συνολικό πλαίσιο.

Αντικατάσταση λέξεων και φράσεων

Το QuillBot δημιουργεί εναλλακτικές επιλογές λέξεων και δομές φράσεων χρησιμοποιώντας την εκτενή συλλογή γλωσσικών μοτίβων και συνωνύμων του. Για να διατηρήσει η έξοδος το αρχικό νόημα, λαμβάνει υπόψη τη συμφραζόμενη και σημασιολογική ομοιότητα εκτός από τα άμεσα συνώνυμα.

Γραμματική και συντακτική προσαρμογή

Κατά την παράφραση, το QuillBot δίνει μεγάλη προσοχή στην κατασκευή φράσεων και στη γραμματική. Για να δημιουργήσει γραμματικά ακριβείς και πειστικές παραφράσεις, αναδιατάσσει προτάσεις και τροποποιεί τις επιλογές λέξεων.

Συμφραζόμενη κατανόηση

Το QuillBot προσπαθεί να κατανοήσει το πλαίσιο και το νόημα του αρχικού κειμένου προκειμένου να παράγει αποτελεσματικές παραφράσεις. Προκειμένου να παράγει πιο ακριβή και συναφή με τα συμφραζόμενα αποτελέσματα, λαμβάνει υπόψη τις λέξεις, τη δομή της πρότασης και το συνολικό πλαίσιο του περιεχομένου.

Ανατροφοδότηση και επανάληψη

Το QuillBot μαθαίνει και αναπτύσσεται συνεχώς μέσω των σχολίων των χρηστών. Η ομάδα QuillBot μπορεί να χρησιμοποιήσει τα σχόλια των χρηστών για να βελτιώσει το σύστημα εάν ανακαλύψει σφάλματα ή ασυνέπειες στην έξοδο που δημιουργεί.

Είναι ζωτικής σημασίας το γεγονός ότι οι αλγόριθμοι του QuillBot έχουν βελτιστοποιηθεί για να μεγιστοποιούν την ποιότητα των παραφρασμένων εξόδων με την εκπαίδευση σε τεράστιους

όγκους δεδομένων. Οι χρήστες θα πρέπει να αξιολογήσουν το παραφρασμένο κείμενο και να κάνουν οποιεσδήποτε απαραίτητες τροποποιήσεις για να βεβαιωθούν ότι ικανοποιεί τις μοναδικές απαιτήσεις τους, επειδή, όπως με κάθε τεχνολογία που βασίζεται σε τεχνητή νοημοσύνη, τα αποτελέσματα δεν είναι πάντα εγγυημένα ότι θα είναι τέλεια.

Κεφάλαιο 5. Αποτελέσματα Έρευνας

Εισάγοντας κάθε μία από τις προτάσεις του σώματος κειμένων που συντάχθηκε στην εφαρμογή του Quilbot, εξάγονται τα ακόλουθα:

1. Sentence simplification is a **standard** NLP **task** of **reducing reading complexity** for people who have **limited linguistic skills to read**.

1a. Sentence simplification is a **common** NLP **activity** to **make reading easier** for those with little **language proficiency**.

2. **In particular**, children, non-native speakers and individuals with language impairments such as **Dyslexia** and **Autistic** (Evans et al., 2014), would benefit from the task which makes sentences easier to understand.

2a. The exercise that simplifies sentences would be especially helpful for kids, non-native speakers, and people with language disabilities like **dyslexia and autism**.

3. There are two categories for the task: lexical simplification and sentence simplification. Both categories enable a paraphrasing process, turning a normal input into a simpler output while maintaining the same/similar semantics between the input and the output.

3a. The work is divided into two categories: phrase simplification and lexical simplification. Both types allow for the process of paraphrase, which reduces a typical input to a more straightforward output while preserving the same or comparable meanings.

4. **Inspired** by the great achievements of machine translation, several studies treat the sentence simplification problem as monolingual translation task and achieve promising results.

4a. **Motivated** by the remarkable accomplishments in machine translation, a number of studies approach the problem of phrase simplification as a single-language translation challenge and obtain encouraging outcomes.

5. **We observe that** sentence simplification **usually means** the simple output sentence is **derived** from the **normal** input sentence with parts of terms changed.

5a. **It is common** for sentence simplification **to result** in a simple output sentence that is **modified in sections** from the **regular** input sentence.

6. Machine translation for **under-resourced languages** is facing **a number of problems**.

6a. **There are several issues** with machine translation **for languages with limited resources**.

7. **First**, there is not enough parallel data to **build** robust **statistical machine translation (SMT)** systems.

7a. **First off**, **statistical machine translation (SMT)** systems cannot be **constructed** with the amount of parallel data available.

8. Second, **most of these languages** (including Serbian) have a very rich morphology and suffer from data sparsity when it comes to less frequently used cases, tenses, etc.

8a. Second, when it comes to less commonly used cases, tenses, etc., **the majority of these languages**—including Serbian—have extremely rich morphologies and experience data sparsity.

9. Third, there is a number of syntactic differences which are difficult to capture.

9a. Third, there are certain difficult to capture syntactic differences.

10. For English-to-Serbian SMT, a number of language related problems has been identified so far (Popovic and Ar'can, 2015).

10a. Numerous language-related issues have been found thus far for English-to-Serbian SMT (Popovic and Ar'ac, 2015).

11. Most of them are related to syntactic differences, e.g. missing verb parts due to distinct structure of certain verb tenses, incorrect prepositions, or incorrect translations of English sequences of nouns.

11a. The majority have to do with syntactic variations, such as missing verb parts because of the unique way that some verb tenses are structured, improper prepositions, or mistranslated English noun sequences.

12. However, in these twenty years, the motivation for building ATS systems has shifted from improving text processing systems to making texts more accessible to wider audiences (e.g. children, non-native speakers, people with low literacy levels, and people with various language or learning disabilities).

12a. The goal of developing ATS systems, however, has changed over the past 20 years from developing text processing systems to bettering text accessibility for a larger range of users, including children, non-native speakers, individuals with low literacy levels, and those with different language or learning disabilities.

13. Therefore, ATS systems have only been evaluated for the quality of the generated output, its readability levels, and usefulness in making texts more accessible to target populations (reducing reading speed and improving comprehension).

13a. As a result, the only factors used to assess ATS systems are the output's quality, readability, and usefulness to improve understanding and cutting reading speeds for those whom it is intended for.

14. To the best of our knowledge, there has been no evaluation of the state-of-the-art ATS systems in terms of how much (if at all) they can help improve MT systems (which was, as previously mentioned, their first intended goal and main motivation).

14a. As far as we know, no evaluation of the ability of the most advanced ATS systems to enhance MT systems, if at all, has been conducted (this was, after all, its main goal and original intended goal).

15. Most sentence compression approaches focus on deleting words (the words appearing in the compression are words occurring in the input) and therefore only perform limited paraphrasing.

15a. The majority of phrase compression techniques only achieve a limited amount of paraphrasing because they focus on removing words (the words that appear in the reduction are terms that exist in the input).

16. As noted by Pitler (2010) and Toutanova et al. (2016) however, the ability to paraphrase is key for the development of abstractive summarisation systems since summaries written by humans often rephrase the original content using paraphrases or synonyms or alternative syntactic constructions.

16a. Pitler (2010) and Toutanova et al. (2016) have found that building of abstractive summarization systems requires the ability to paraphrase, as human-written summaries often use synonyms, paraphrases, or other grammatical structures to rephrase the initial data.

17. Recent proposals by Rush et al. (2015) and Bingel and Søgaard (2016) address this issue.

17a. This issue appears in recent works by Bingel and Søgaard (2016) and Rush et al. (2015).

18. Rush et al. (2015) proposed a neural model for abstractive compression and summarisation, and Bingel and Søgaard (2016) proposed a structured approach to text simplification which jointly predicts possible compressions and paraphrases.

18a. In 2015, Rush et al. presented a neural model for abstractive compression and summarization, while in 2016, Bingel and Søgaard presented a structured method for text simplification that anticipates potential compressions and paraphrases simultaneously.

Το TS έχει πραγματοποιηθεί με πολλούς διαφορετικούς τρόπους. Πολλά συστήματα χρησιμοποιούν έναν συνδυασμό προσεγγίσεων για την απλοποίηση του κειμένου με διαφορετικούς τρόπους. Αυτές οι διαφορετικές μέθοδοι TS είναι σε μεγάλο βαθμό ανεξάρτητες και μεθοδολογικά διακριτές η μία από την άλλη.

Η λεξιλογική απλοποίηση που είδαμε αρκετά ανωτέρω (π.χ. 18-18a proposed-presented, 15-15a deleting-removing) σχετίζεται με τον εντοπισμό και την αντικατάσταση σύνθετων λέξεων με απλούστερα υποκατάστατα. Αυτό δεν περιλαμβάνει καμία προσπάθεια απλοποίησης της γραμματικής ενός κειμένου, αλλά εστιάζει στην απλοποίηση σύνθετων πτυχών του λεξιλογίου. Η λεξική απλοποίηση μπορεί να διατυπωθεί ως σύστημα υποκατάστασης βασισμένο σε φράσεις, το οποίο λαμβάνει υπόψη περιορισμένες συντακτικές πληροφορίες. Υπάρχουν συνήθως 4 βήματα για τη λεξιλογική απλοποίηση. Πρώτον, πρέπει να προσδιοριστούν οι περίπλοκοι όροι σε ένα έγγραφο, όπως εν προκειμένω τα χρωματισμένα σημεία στις αρχικές προτάσεις του σώματος κειμένου. Δεύτερον, πρέπει να δημιουργηθεί μια λίστα αντικαταστάσεων για κάθε μία (οι επιλογές αυτές δόθηκαν κατά τη λειτουργία του Quilbot). Τρίτον, αυτές οι αντικαταστάσεις θα πρέπει να βελτιωθούν ώστε να διατηρηθούν εκείνες που έχουν νόημα στο δεδομένο πλαίσιο. Εν προκειμένω κατά τη λειτουργία του Quilbot επιλέχθηκαν εκείνες οι λέξεις και οι φράσεις που απέδιδαν απλούστερα το νόημα της κάθε αρχικής φράσης. Τέλος, οι υπόλοιπες αντικαταστάσεις πρέπει να ταξινομηθούν με τη σειρά της απλότητάς τους. Το πιο απλό συνώνυμο χρησιμοποιείται ως αντικατάσταση της αρχικής λέξης.

Από την άλλη, η συντακτική απλοποίηση είναι επίσης τεχνική που συναντάνται στην απλοποίηση κειμένου και συνιστά την τεχνική αναγνώρισης πολύπλοκων γραμματικών φαινομένων σε ένα κείμενο και επανεγγραφής αυτών σε απλούστερες δομές. Στο σώμα κειμένων που δόθηκε ανωτέρω (π.χ. 14-14a **To the best of our knowledge**, -As far as we know, 7-7a 7. **First**, there is not enough parallel data to **build** robust **statistical machine translation (SMT)** systems-. **First off**, **statistical machine translation (SMT)** systems cannot be **constructed** with the amount of parallel data available). Υπάρχουν πολλοί τύποι συντακτικής πολυπλοκότητας στους οποίους αυτό μπορεί

να ισχύει: Οι μεγάλες προτάσεις μπορούν να χωριστούν στις συνιστώσες τους. Οι προτάσεις που χρησιμοποιούν την παθητική φωνή μπορούν να ξαναγραφτούν και η αναφορά μπορεί να επιλυθεί.

Η συντακτική απλοποίηση γίνεται συνήθως σε τρεις φάσεις. Αρχικά, το κείμενο αναλύεται για να προσδιοριστεί η δομή του και να αναλυθεί το δέντρο. Αυτό μπορεί να γίνει με διαφορετική ευαισθησία, αλλά έχει αποδειχθεί ότι λειτουργεί σε μάλλον ουσιαστικότερο επίπεδο. Σε αυτό το επίπεδο, οι λέξεις και οι φράσεις ομαδοποιούνται σε «υπερ-ετικέτες» που αντιπροσωπεύουν ένα κομμάτι της υποκείμενης πρότασης. Αυτές οι σούπερ ετικέτες μπορούν να ενωθούν με συμβατικούς γραμματικούς κανόνες για να παρέχουν μια δομημένη έκδοση του κειμένου. Κατά τη φάση της ανάλυσης, η πολυπλοκότητα μιας πρότασης καθορίζεται για να αποφασιστεί εάν θα χρειαστεί απλοποίηση. Αυτό μπορεί να γίνει με την αυτόματη αντιστοίχιση κανόνων, αλλά έχει γίνει επίσης χρησιμοποιώντας έναν δυαδικό ταξινομητή μηχανής διανυσμάτων υποστήριξης. Η δεύτερη φάση είναι ο μετασχηματισμός, στον οποίο γίνονται τροποποιήσεις στο δέντρο ανάλυσης σύμφωνα με ένα σύνολο κανόνων επανεγγραφής. Αυτοί οι κανόνες επανεγγραφής εκτελούν λειτουργίες απλοποίησης όπως διαχωρισμός προτάσεων και αναδιάταξη ρήτρας. Αν και υπάρχουν τεχνικές για την αυτόματη επαγωγή αυτών των κανόνων, τα περισσότερα άλλα συστήματα που εφαρμόζουν συντακτική απλοποίηση χρησιμοποιούν χειρόγραφους κανόνες επανεγγραφής.

6.1. Συζήτηση των αποτελεσμάτων

Οι ερμηνείες των αποτελεσμάτων των μελετών που συμπεριλαμβάνονται στο ερευνητικό σκέλος, επιβεβαιώνουν την πολυπλοκότητα και την ποικιλομορφία των προσεγγίσεων στην απλοποίηση κειμένου. Η λεξιλογική αφομοίωση, η χρήση περιορισμένων αλληλουχιών και η αξιολόγηση της αναγνωσιμότητας αποτελούν σημαντικούς παράγοντες που συμβάλλουν στην αποτελεσματική απλοποίηση και κατανόηση των κειμένων. Τα ευρήματα υπογραμμίζουν την ανάγκη για περαιτέρω έρευνα και ανάπτυξη νέων μεθόδων και εργαλείων που θα βελτιώσουν την ακρίβεια και την αποδοτικότητα αυτών των διεργασιών. Πιο συγκεκριμένα:

1. Ενσωματώσεις λέξεων και χαρακτηριστικών για ανίχνευση λεξιλογικής πολυπλοκότητας

Τα ευρήματα της μελέτης δείχνουν ότι οι συχνότερες λέξεις και οι συνδυασμοί χαρακτήρων και φωνημάτων είναι πιο αποτελεσματικοί στην πρόβλεψη της λεξιλογικής πολυπλοκότητας σε σχέση με την αφομοίωση λέξεων. Αυτή η γνώση μπορεί να εφαρμοστεί στην ανάπτυξη εκπαιδευτικών εργαλείων που στοχεύουν στη βελτίωση της γλωσσικής εκπαίδευσης. Για παράδειγμα, οι εκπαιδευτικές πλατφόρμες μπορούν να χρησιμοποιούν αυτά τα συστήματα για να προσαρμόζουν το επίπεδο δυσκολίας των κειμένων ανάλογα με τις ανάγκες των μαθητών, προκειμένου να βελτιώσουν την κατανόηση και την εκμάθηση νέων λέξεων.

2. Μοντέλο αλληλουχίας για την απλοποίηση προτάσεων

Το μοντέλο αλληλουχίας με περιορισμούς (Constrained Seq2Seq) για την απλοποίηση προτάσεων μπορεί να έχει σημαντικές εφαρμογές στην εκπαίδευση και στην επικοινωνία με άτομα με γνωστικές ή γλωσσικές δυσκολίες. Η δυνατότητα απλοποίησης σύνθετων προτάσεων χωρίς να χάνεται το νόημα μπορεί να βοηθήσει στη δημιουργία εκπαιδευτικών υλικών που είναι πιο προσβάσιμα σε μαθητές με μαθησιακές δυσκολίες ή άτομα που μαθαίνουν μια δεύτερη γλώσσα. Επιπλέον, τέτοιες τεχνολογίες μπορούν να ενσωματωθούν σε λογισμικά επεξεργασίας κειμένου για να διευκολύνουν τη δημιουργία κειμένων που είναι εύκολα κατανοητά από το ευρύ κοινό.

3. Σώμα για την αυτόματη αξιολόγηση αναγνωσιμότητας και απλοποίησης κειμένου στα γερμανικά

Η ανάπτυξη ενός τέτοιου σώματος δεδομένων προσφέρει νέες ευκαιρίες για την έρευνα στην αναγνωσιμότητα και την απλοποίηση κειμένου στα γερμανικά. Τα ευρήματα μπορούν να

εφαρμοστούν στην ανάπτυξη εκπαιδευτικών λογισμικών και εργαλείων που βοηθούν στη διδασκαλία της γερμανικής γλώσσας, καθώς και σε προγράμματα που στοχεύουν στη βελτίωση της αναγνωσιμότητας για άτομα με προβλήματα όρασης ή μαθησιακές δυσκολίες. Επίσης, μπορούν να χρησιμοποιηθούν για τη βελτίωση της ποιότητας των μεταφραστικών εργαλείων, καθιστώντας τα πιο ακριβή και αξιόπιστα.

4. Σύνολο δεδομένων και μετρήσεις αξιολόγησης για αφηρημένη συμπίεση προτάσεων και σύντομων παραγράφων

Το σύνολο δεδομένων και οι μετρήσεις αξιολόγησης για την αφηρημένη συμπίεση προτάσεων προσφέρουν μια νέα προσέγγιση στη δημιουργία συνοπτικών και κατανοητών κειμένων. Αυτές οι τεχνολογίες μπορούν να εφαρμοστούν σε ειδήσεις, επιστημονικά άρθρα και άλλα πληροφοριακά κείμενα για τη δημιουργία περιλήψεων που διατηρούν την ουσία του πρωτοτύπου. Επίσης, μπορούν να βοηθήσουν στη βελτίωση των εργαλείων επεξεργασίας κειμένου, καθιστώντας τα πιο αποτελεσματικά στην παραγωγή σύντομων και σαφών περιλήψεων.

5. Εξερεύνηση της προσέγγισης PB-SMT για την απλοποίηση κειμένου

Η ανάλυση της ποιότητας των δεδομένων εκπαίδευσης στην προσέγγιση PB-SMT για την απλοποίηση κειμένου αναδεικνύει τη σημασία της ποιότητας έναντι της ποσότητας. Αυτό έχει σημαντικές επιπτώσεις για την ανάπτυξη μελλοντικών μοντέλων απλοποίησης κειμένου, καθώς υπογραμμίζει την ανάγκη για υψηλής ποιότητας δεδομένα εκπαίδευσης. Οι εφαρμογές αυτών των ευρημάτων μπορούν να ενσωματωθούν σε λογισμικά εκπαίδευσης και επικοινωνίας, προσφέροντας καλύτερες και πιο αποδοτικές λύσεις για την απλοποίηση κειμένων.

Οι επιπτώσεις των παραπάνω ευρημάτων τονίζουν την αναγκαιότητα για τη συνεχή βελτίωση και ανάπτυξη εργαλείων που θα κάνουν τα κείμενα πιο προσιτά και κατανοητά. Η εφαρμογή αυτών των τεχνολογιών μπορεί να προσφέρει σημαντική βοήθεια σε εκπαιδευτικά περιβάλλοντα, επικοινωνιακά εργαλεία και πλατφόρμες ψηφιακής πρόσβασης, ενισχύοντας την κατανόηση και τη συμμετοχή όλων των χρηστών, ανεξαρτήτως των γνωστικών ή γλωσσικών τους ικανοτήτων.

6.2 Προτάσεις για Μελλοντική Έρευνα

Συνολικά, το TS είναι ένας τομέας που έχει προκύψει ως η απάντηση στην ανάγκη κατανόησης επιστημονικών, αλλόγλωσσων και απλά δύσκολων κειμένων. Αυτό έχει συμβεί για διαφορετικές εφαρμογές, όπως η προεπεξεργασία για αυτόματη μετάφραση και η υποστηρικτική τεχνολογία για άτομα με προβλήματα ανάγνωσης ή αλλοδαπούς. Αυτές οι εφαρμογές υπόσχονται να μειώσουν την πολυπλοκότητα του κειμένου βελτιώνοντας παράλληλα την αναγνωσιμότητα και την κατανόηση. Αυτή είναι μια εξαιρετικά χρήσιμη εργασία και είναι εξαιρετικά εφαρμόσιμη σε πολλές ρυθμίσεις, όπως οι μαθητές δεύτερης γλώσσας και οι απλοί αναγνώστες τεχνικών εγγράφων. Η απλοποίηση κειμένου βέβαια δεν περιορίζεται αποκλειστικά στον αναγνώστη, μπορεί επίσης να εφαρμοστεί από τον συγγραφέα σε ένα κείμενο προκειμένου να διασφαλίσει ότι η άποψή του μεταδίδεται με σαφήνεια, ή ακόμη και σε μια γραμμή επεξεργασίας φυσικής γλώσσας για τη βελτίωση της απόδοσης μεταγενέστερων στοιχείων.

Υπάρχουν επίσης πολλές προσεγγίσεις στις οποίες μπορεί να βοηθήσει η απλοποίηση κειμένου. Ορισμένοι επικεντρώνονται στο λεξιλογικό επίπεδο, αντικαθιστώντας σύνθετες λέξεις με απλούστερες συνώνυμες. Ορισμένοι τροποποιούν τη σύνταξη ενός κειμένου για να αφαιρέσουν πολύπλοκες γραμματικές δομές. Ωστόσο, άλλοι εκτελούν μηχανική μετάφραση βασισμένη σε φράσεις σε μια προσπάθεια να μάθουν αυτόματα έγκυρες μεθόδους απλοποίησης. Το πεδίο αυτή τη στιγμή βλέπει ένα κύμα ανάπτυξης με πολλά νέα ερευνητικά έργα και νέες προσεγγίσεις να αναπτύσσονται. Καθώς το πεδίο προχωρά, περισσότερες τεχνικές θα είναι χρήσιμες καθιστώντας την απλοποίηση κειμένου να διανεμηθεί ευρέως. Παρακάτω καταγράφονται κάποιες περιπτώσεις σημαντικής εφαρμογής του TS:

1. Ενσωματώσεις λέξεων και λεξιλογική πολυπλοκότητα

- Διεύρυνση Λεξιλογικών Πόρων: Επεκτείνετε τους λεξιλογικούς πόρους όπως τα Manulex και FLElex με περισσότερα δεδομένα από διαφορετικές γλωσσικές κοινότητες και θεματικές ενότητες.
- Διαπολιτισμικές Συγκρίσεις: Διερευνήστε πώς η λεξιλογική πολυπλοκότητα διαφοροποιείται μεταξύ διαφορετικών γλωσσών και πολιτισμών, χρησιμοποιώντας παρόμοιες μεθόδους ανάλυσης.

2. Μοντέλα αλληλουχίας για απλοποίηση προτάσεων

- Βελτίωση Μοντέλων Αλληλουχίας: Ανάπτυξη πιο εξελιγμένων μοντέλων Seq2Seq με περιορισμούς, ενσωματώνοντας μηχανισμούς προσοχής για βελτίωση της ακρίβειας και της φυσικότητας της απλοποίησης.
- Δοκιμή σε Πραγματικά Σενάρια: Εφαρμογή των μοντέλων σε πραγματικά εκπαιδευτικά ή επαγγελματικά σενάρια για να αξιολογηθεί η αποτελεσματικότητα και η χρηστικότητά τους στην καθημερινή χρήση.

3. Σώματα δεδομένων για αναγνωσιμότητα και απλοποίηση κειμένου

- Πολυτροπικά Δεδομένα: Ανάπτυξη σωμάτων δεδομένων που περιλαμβάνουν πολυτροπικά στοιχεία (εικόνες, βίντεο, ήχο) για καλύτερη αξιολόγηση της αναγνωσιμότητας και της απλοποίησης σε διάφορα μέσα.
- Ανάλυση Χρηστών: Εξετάστε πώς διαφορετικές ομάδες χρηστών (π.χ. παιδιά, ηλικιωμένοι, άτομα με αναπηρίες) αντιδρούν σε απλοποιημένα κείμενα για να κατανοήσετε καλύτερα τις ανάγκες τους.

4. Συμπίεση προτάσεων και παραγράφων

- Εξερεύνηση Μεθόδων Συμπίεσης: Διερευνήστε νέες μεθόδους συμπίεσης που συνδυάζουν αφηρημένα και εξαγωγική προσέγγιση για βελτίωση της συνοπτικότητας και της αναγνωσιμότητας.
- Προσαρμοστική Συμπίεση: Ανάπτυξη συστημάτων που προσαρμόζουν τη συμπίεση βάσει του προφίλ του χρήστη και των προτιμήσεών του.

5. PB-SMT για απλοποίηση κειμένου

- Ανάπτυξη Νέων Μετρήσεων Αξιολόγησης: Δημιουργία νέων μετρήσεων που θα είναι πιο κατάλληλες για την αξιολόγηση της ποιότητας της απλοποίησης κειμένου, πέρα από το BLEU.
- Εκπαίδευση με Ποιοτικά Δεδομένα: Συνεχίστε την έρευνα για την επίδραση της ποιότητας των δεδομένων εκπαίδευσης, δοκιμάζοντας διαφορετικές πηγές δεδομένων και προσεγγίσεις για την εξασφάλιση υψηλής ποιότητας δεδομένων.

Η μελλοντική έρευνα σε αυτά τα πεδία μπορεί να προσφέρει σημαντικές προόδους στη γλωσσική τεχνολογία και στην κατανόηση της αναγνωσιμότητας και της απλοποίησης κειμένου. Οι προτάσεις αυτές στοχεύουν στη βελτίωση των μεθόδων και των εργαλείων που χρησιμοποιούνται, ώστε να καταστούν πιο αποτελεσματικά και χρήσιμα σε διάφορα πλαίσια εφαρμογής.

Βιβλιογραφία

- A.I. Tweissi The effects of the amount and the type of simplification on foreign language reading comprehension *Reading in a Foreign Language*, 11 (1998)
- Carbonell, J.G., Michalski, R.S., Mitchell, T.M. (1983). An Overview of Machine Learning. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds) *Machine Learning. Symbolic Computation*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-12405-5_1
- Long, M. & Ross, S. (1993). Modifications that preserve language and content. SEAMEO Regional Language Center
- Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994) *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York. <http://www1.maths.leeds.ac.uk/~charles/statlog/whole.pdf>
- Russell, S. & Norvig, Peter. (2003). *Artificial Intelligence, A Modern Approach*. Second Edition.
- Aina Garí, Marianna Apidianaki, Alexandre Allauzen. A comparative study of word embeddings and other features for lexical complexity detection in French. *Conférence sur le Traitement Automatique des Langues Naturelles*, May 2018, Rennes, France. fihal-01838512ff
- Zhang, Y., Ye, Z., Feng, Y., Zhao, D., & Yan, R. (2017). A Constrained Sequence-to-Sequence Neural Model for Sentence Simplification. *ArXiv*, *abs/1704.02312*.
- Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. *A Corpus for Automatic Readability Assessment and Text Simplification of German*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.
- Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A Dataset and Evaluation Metrics for Abstractive Compression of Sentences and Short Paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.

- Stajner, Sanja & Béchara, Hanna & Saggion, Horacio. (2015). A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. 10.3115/v1/P15-2135.
- Santucci V, Santarelli F, Forti L, Spina S. Automatic Classification of Text Complexity. *Applied Sciences*. 2020; 10(20):7285. <https://doi.org/10.3390/app10207285>
- Alonzo, Oliver, "Automatic Text Simplification Tools to Provide Reading Assistance for Deaf and Hard-of-Hearing Individuals" (2023). Thesis. Rochester Institute of Technology.
- Stajner, Sanja & Popovic, Maja. (2016). Can Text Simplification Help Machine Translation?
- Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An Exploration into Neural Text Simplification Models. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 5588–5594, Marseille, France. European Language Resources Association.
- Brunato D, Dell’Orletta F and Venturi G (2022) Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian. *Front. Psychol.* 13:707630. doi: 10.3389/fpsyg.2022.707630
- Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*. Volume 1 : conférence principale, pages 116–125, Avignon, France. ATALA.
- Yaushian Wang and Hung-Yi Lee. 2018. Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4187–4195, Brussels, Belgium. Association for Computational Linguistics.
- Phatak A, Savage DW, Ohle R, Smith J, Mago V. Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning-Based Text Simplification Approach. *JMIR Med Inform.* 2022 Nov 18;10(11):e38095. doi: 10.2196/38095. PMID: 36399375; PMCID: PMC9719064.
- Nassar, I., Ananda-Rajah, M.R., & Haffari, G. (2019). Neural Versus Non-Neural Text Simplification: A Case Study. *Australasian Language Technology Association Workshop*.

- Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating Factuality in Text Simplification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics, 4:401–415.
- Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4972–4984, Online. Association for Computational Linguistics.
- Elinor Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic Structural Evaluation for Text Simplification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics
- Xingxing Zhang and Mirella Lapata. 2017. Sentence Simplification with Deep Reinforcement Learning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.
- Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and Rephrase. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.
- Evelina Rennes and Arne Jönsson. 2021. Synonym Replacement based on a Study of Basic-level Nouns in Swedish Texts of Different Complexity. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 259–267, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.
- Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised Neural Text Simplification. In Proceedings of the 57th Annual Meeting of

the Association for Computational Linguistics, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

- Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.
- Kurdi, Mohamed. (2020). Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL.
- Zarah Weiss and Detmar Meurers. 2022. Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?. In Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 141–153, Seattle, Washington. Association for Computational Linguistics.
- Laura Seiffe, Fares Kallel, Sebastian Möller, Babak Naderi, and Roland Roller. 2022. Subjective Text Complexity Assessment for German. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 707–714, Marseille, France. European Language Resources Association.
- Renate Hauser, Jannis Vamvas, Sarah Ebling, and Martin Volk. 2022. A Multilingual Simplified Language News Corpus. In Proceedings of the 2nd Workshop on Tools and Resources to Empower People with READING Difficulties (READI) within the 13th Language Resources and Evaluation Conference, pages 25–30, Marseille, France. European Language Resources Association.
- Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A Benchmark for Neural Readability Assessment of Texts in Spanish. In Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Dan Zhang, 2021. "Text Complexity Classification Data Mining Model Based on Dynamic Quantitative Relationship between Modality and English Context," Mathematical Problems in Engineering, Hindawi, vol. 2021, pages 1-10, December.