TR•AID: ΟΡΟΛΟΓΙΑ ΚΑΙ ΣΥΓΧΡΟΝΗ ΜΕΤΑΦΡΑΣΤΙΚΗ ΠΡΑΚΤΙΚΗ

Ιωάννης Τριανταφύλλου, Χρήστος Μαλαβάζος, Στέλιος Πιπερίδης

ΠΕΡΙΛΗΨΗ

Το άρθρο αυτό περιγράφει το σύστημα **TR•AID,** μια πολυεπίπεδη αρχιτεκτονική που υποστηρίζει την πλατφόρμα ενός συστήματος μετάφρασης με τη βοήθεια Η/Υ. Το σύστημα χρησιμοποιεί διαφορετικά επίπεδα πληροφορίας και διαδικασιών σε μια προσπάθεια να μεγιστοποιήσει την επαναχρησιμοποίηση παλιότερων μεταφράσεων και να βελτιστοποιήσει τη χρήση ορολογικών δεδομένων, διατηρώντας παράλληλα τη συνέπεια της μετάφρασης σε συγκεκριμένους τύπους κειμένου.

TR·AID: TERMINOLOGY AND MODERN TRANSLATION PRACTICE

Ioannis Triantafyllou, Christos Malavazos, Stelios Piperidis

ABSTRACT

This paper describes **TR•AID**, a multi-level architecture for a computer-aided translation (CAT) system platform. The system employs different levels of information and processing in an attempt to maximize past translation reuse as well as terminology and style consistency in the translation of specific types of text.

Λέξεις-Κλειδιά: Μηχανική Μετάφραση-(Machine Translation/**MT**), Μηχανική Μετάφραση Βασισμένη σε Παραδείγματα (Example Based Machine Translation/**EBMT**), Εντοπισμός και Μετάφραση Όρων (Term Spotting and Translation).

0 ΕΙΣΑΓΩΓΗ

0.1 ΓΕΝΙΚΑ

Η χρήση μεθόδων μάθησης και ταυτοποίησης προτύπων στον τομέα της μηχανικής μετάφρασης, που υιοθετήθηκαν για πρώτη φορά στις αρχές της δεκαετίας του 80 [11]-(Nagao 84) με τον τίτλο "Μετάφραση κατ' Αναλογία" ("*Translation by Analogy*") και η επιστροφή σε στατιστικές μεθόδους στις αρχές τις δεκαετίας του 90 [1]-(Brown et al. 93) προετοίμασαν το έδαφος για έντονους προβληματισμούς και αναθεωρήσεις για την αρχιτεκτονική των σύγχρονων συστημάτων μηχανικής μετάφρασης. Η επεξεργασία δίγλωσσων σωμάτων κειμένων και συγκεκριμένα η διαδικασία της στοίχισης παραλλήλων κειμένων (text alignment) και η περαιτέρω εκμετάλλευση της πληροφορίας των

παραδειγμάτων μετάφρασης που προκύπτουν από αυτήν, σηματοδότησαν την αρχή ενός νέου ρεύματος στη μηχανική μετάφραση.

Τα παραδοσιακά συστήματα μηχανικής μετάφρασης βάσει κανόνων (Rule Based Machine Translation/**RBMT**) πάσχουν από προβλήματα ανίχνευσης των λαθών, προσαρμοστικότητας, καθώς και προβλήματα ποιότητας της μετάφρασης. Τα συστήματα μηχανικής μετάφρασης βάσει παραδειγμάτων (Example Based Machine Translation/**EBMT**) προσπάθησαν να φέρουν στην επιφάνεια και να εισάγουν εναλλακτικούς τρόπους αντιμετώπισης του προβλήματος της μετάφρασης, σημειώνοντας ενδιαφέροντα αποτελέσματα.

0.2 ΑΝΑΣΚΟΠΗΣΗ

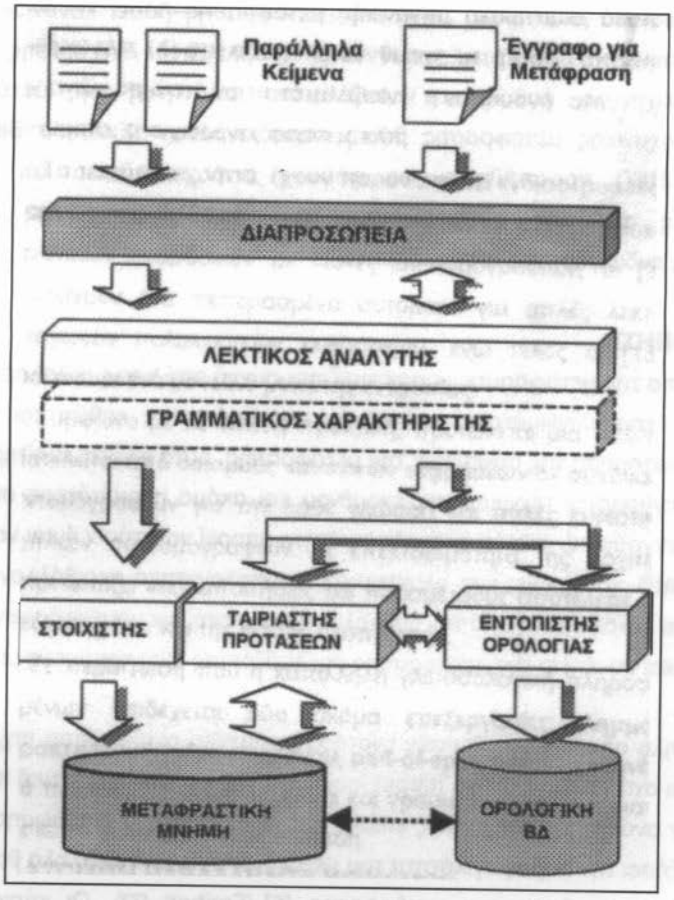
Η διαδικασία της μετάφρασης χαρακτηρίζεται συχνά από τρεις αλληλοσυγκρουόμενες παραμέτρους: επαναληψιμότητα, ανάγκη άμεσης απόκρισης καθώς και ποιότητας, με έμφαση στη διατήρηση της συνέπειας της μετάφρασης. Αυτό γίνεται ιδιαίτερα εμφανές σε περιπτώσεις εγγράφων τεχνικού περιεχομένου και ακόμα περισσότερο σε περιπτώσεις νομοθετικών εγγράφων, όπου η επαναληψιμότητα μπορεί να φτάσει ή και να ξεπεράσει το 70%. Το **TR•AID** προσφέρει ένα ολοκληρωμένο υπολογιστικό περιβάλλον το οποίο: (α) απαλλάσσει το μεταφραστή από την επανάληψη παλαιοτέρων μεταφράσεων, (β) βελτιώνει την ποιότητα και συνέπεια της μετάφρασης συνδυάζοντας διαφορετικούς μεταφραστικούς μηχανισμούς.

Η κατάλληλη αποθήκευση ζευγών ενοτήτων κειμένου στη γλώσσα πηγή (source language/SL) και στη γλώσσα στόχο (target language/TL) ως "μεταφραστικά παραδείγματα", με σκοπό την ανάκτηση τους καθώς επίσης και τη δυνατότητα διόρθωσης των κειμένων αυτών θα αυξήσει την παραγωγικότητα του μεταφραστή ενώ παράλληλα θα βελτιώσει την ποιότητα και τη συνέπεια της μετάφρασης [5]-(Freibott 92). Οι κύριοι άξονες της μεθοδολογίας που υιοθετείται είναι τέσσερις: (1) "αυτόματη" στοίχιση παράλληλων κειμένων, δηλαδή εύρεση επιμέρους αντιστοιχιών μεταξύ τμημάτων παραλλήλων κειμένων, (2) οργάνωση πολύγλωσσων σωμάτων κειμένων, δηλαδή κείμενων σε διαφορετικές γλώσσες, που το ένα αποτελεί μετάφραση του άλλου, επιτρέποντας την αποτελεσματική αποθήκευση και εξαγωγή μεταφραστικών παραδειγμάτων, όπως επίσης και ορολογικής πληροφορίας, (3) έξυπνες τεχνικές ταυτοποίησης προτύπων για γρήγορη ανάκτηση των περισσοτέρων προτύπων μετάφρασης, (4) έξυπνες τεχνικές εντοπισμού και μετάφρασης όρων.

Εναλλακτικές τεχνικές για το κάθε λειτουργικό κομμάτι της προτεινόμενης αρχιτεκτονικής μελετήθηκαν με σκοπό την υιοθέτηση των πιο πρακτικών αλλά και λιγότερο απαιτητικών λύσεων για την ανάπτυξη του συστήματος **TR•AID (Translation Aid).**

1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

1.1 ΓΕΝΙΚΑ



Εικόνα 1: Αρχιτεκτονική του Tr-AID

Στην εικόνα 1 παρουσιάζεται η γενική αρχιτε-κτονική του **TR•AID** συστήματος καθώς επίσης και τα υποσυστήματά του. Μια λεπτομερής περιγραφή για το καθένα από αυτά δίνεται στις επόμενες παραγράφους.

1.2 ΛΕΚΤΙΚΗ ΑΝΑΛΥΣΗ

Προκειμένου να μπορεί να γίνει πλήρης χρήση των σωμάτων κειμένων, τα τελευταία πρέπει να αποδοθούν σε μια κατάλληλη μορφή. Με βάση το σκεπτικό αυτό τα σώματα κειμένων πρέπει να κανονικοποιηθούν και να αναλυθούν λεκτικά πριν τη διαδικασία της στοίχισης. Η **κανονικοποίηση** συνίσταται στην εξαγωγή από τα σώματα κειμένων όλης

εκείνης της πληροφορίας που δεν είναι απαραίτητη στις υπόλοιπες διαδικασίες που θα ακολουθήσουν (λεκτική ανάλυση, στοίχιση).

Η **λεκτική ανάλυση** μπορεί να θεωρηθεί σαν τη διασύνδεση μεταξύ του κειμένου εισόδου και των διαφόρων υποσυστημάτων επεξεργασίας του κειμένου. Οι κύριες λειτουργίες που αφορούν το λεκτικό αναλυτή περιλαμβάνουν: (1) ανάλυση της **μορφής** του κειμένου εισόδου (έντονα ή πλάγια γράμματα κλπ) και κωδικοποίηση σε μια γενικά αποδεκτή μορφή αναγνωρίσιμη από την εφαρμογή, (2) αναγνώριση των **μονάδων κειμένου** στο επίπεδο της παραγράφου και της πρότασης, (3) αναγνώριση των **ειδικών λεκτικών μονάδων,** όπως ημερομηνίες, ακρωνύμια, συντομογραφίες, αριθμημένες λίστες, αριθμοί κλπ, (4) στη φάση της σύνθεσης, μετατροπή από τη μορφή που έχει το κείμενο μετά την ανάλυση ώστε να το καταλαβαίνει η εφαρμογή, στην αρχική μορφή εισόδου π.χ. πλάγια γράμματα, στοιχισμένες φράσεις κλπ.

Τα τελευταία χρόνια, έχουν παρουσιασθεί ενδιαφέρουσες τεχνικές λεκτικής ανάλυσης και αναγνώρισης προτάσεων. Οι [8]-(Grefenstette & Tapanainen 94) εφαρμόζουν γραμματικές κανονικών εκφράσεων με λίστες συντομογραφιών, οι [3]-(Chanod & Tapanainen 96) προτείνουν ένα πεπερασμένο αυτόματο για απλές λεκτικές μονάδες και ένα λεκτικό μεταγωγέα πολυλεκτικών εκφράσεων, ενώ τέλος οι [13]-(Reynar & Ratnaparkhi 97) προτείνουν ένα μοντέλο μέγιστης εντροπίας.

Σύμφωνα με τη διεθνή πρακτική, προτείνεται μια πολυεπίπεδη αρχιτεκτονική, η οποία αποτελείται από ορισμούς κανονικών εκφράσεων σε συνδυασμό με προκαθορισμένες λίστες συντομογραφιών για την κάθε γλώσσα και απλές ευριστικές τεχνικές για τον εντοπισμό επιπλέον συντομογραφιών. Η επεκτασιμότητα της αρχιτεκτονικής ώστε να μπορεί να αντιμετωπίζει νέες απαιτήσεις και παραμέτρους αποτέλεσε βασικό κριτήριο τόσο της φάσης σχεδίασης όσο και της φάσης υλοποίησης.

Σε περίπτωση που η κατάλληλη γλωσσολογική πληροφορία είναι διαθέσιμη, το σώμα κειμένων επιδέχεται δύο ακόμα επεξεργασίες: **λημματοποίηση** και **γραμματικό χαρακτηρισμό** (μέρη του λόγου, part-of-speech/pos). Πιθανές αμφισημίες που προέρχονται από πολλαπλά πιθανά λήμματα και γραμματικά χαρακτηριστικά αποθηκεύονται στη μνήμη με σκοπό μελλοντική τους επίλυση.

1.3 ΣΤΟΙΧΙΣΗ ΠΑΡΑΛΛΗΛΩΝ ΚΕΙΜΕΝΩΝ

Σημαντικό παράγοντα στη μεθοδολογία στοίχισης παραλλήλων κειμένων αποτελεί η μονάδα κειμένου (πρόταση, επίπεδο μικρότερο της πρότασης, φράσεις, λέξεις). Η επιλογή της μονάδας κειμένου επηρεάζει κυρίως το μηχανισμό ανάκτησης της πιο ταιριαστής πρότασης. Επειδή η επιλογή του επιπέδου της πρότασης εξασφαλίζει την άρση πιθανών μεταφραστικών σημασιολογικών αμφισημιών, σαν μονάδα κειμένου στο

περιβάλλον του **TR•AID,** όπως επίσης και στη διαδικασία της στοίχισης, έχει επιλεχθεί η πρόταση.

Διάφορες προσεγγίσεις έχουν προταθεί για το πρόβλημα της στοίχισης σε διάφορα επίπεδα. Η τεχνική των [2]-(Catizone et al. 89) βασίζεται στις συνεμφανίσεις λέξεων στα δύο παράλληλα κείμενα. Οι [7]-(Gale & Church 91) προτείνουν ένα απλό στατιστικό μοντέλο με μήκη χαρακτήρων. Το μοντέλο βασίζεται στην παρατήρηση ότι τα μήκη αντίστοιχων προτάσεων μεταξύ δύο γλωσσών σχετίζονται σε αρκετά μεγάλο βαθμό. Οι [10]-(Kay & Roescheisen 91) παρουσιάζουν έναν αλγόριθμο που συνδυάζει αναδρομικά αντιστοιχίες τόσο μεταξύ λέξεων όσο και προτάσεων.

Το προτεινόμενο σχήμα συνίσταται σε μια πολυεπίπεδη αρχιτεκτονική με κύριο άξονα το μηχανισμό των [7]-(Gale & Church 91). Ιδιαίτερη προσπάθεια καταβλήθηκε με σκοπό να ενισχυθεί ο βασικός μηχανισμός δια μέσου ισχυρών σημείων στοίχισης ή σημείων αγκίστρωσης, με βάση μόνο επισκόπηση της πληροφορίας που περιέχουν τα δύο κείμενα. Υποψήφιες στοιχίσεις λέξεων υπολογίζονται με βάση τις συνεμφανίσεις και τις κατανομές απλών και σύνθετων λεκτικών μονάδων κειμένου. Κατόπιν η πληροφορία αυτή χρησιμοποιείται για να εντοπισθούν οι πιο πιθανές στοιχίσεις προτάσεων, οι οποίες και υποδηλώνουν τα όρια εντός των οποίων θα εφαρμοσθεί ο βασικός αλγόριθμος. Σημαντική βελτίωση των αποτελεσμάτων μπορεί να επιτευχθεί με τη χρήση δίγλωσσων λεξικών.

1.4 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Η πολυπλοκότητα που χαρακτηρίζει τη μεταφραστική διαδικασία στα πλαίσια ενός τυπικού **EBMT** περιβάλλοντος καθιστά αναγκαία τη βέλτιστη οργάνωση της πληροφορίας σε μια βάση δεδομένων (ΒΔ). Η ανάγκη για βέλτιστη διαχείριση διαφορετικών επιπέδων πληροφορίας καθώς και για άμεση απόκριση του συστήματος, προϋποθέτει ένα αποδοτικό, καλά οργανωμένο και ισχυρό αρχιτεκτονικό σχήμα της ΒΔ.

Σαν μετα-δεδομένα ορίζουμε όλες τις αυτόνομες (διακριτές) μονάδες κειμένου που χειρίζεται το σύστημα, οι οποίες προέρχονται από το αρχικό κείμενο δια μέσου της λεκτικής ανάλυσης και της στοίχισης παραλλήλων κειμένων όπως περιγράφηκαν προηγουμένως. Η προτεινόμενη αρχιτεκτονική εκτός από την αποθήκευση των μετα-δεδομένων του μονόγλωσσου σώματος κειμένου, χρειάζεται επιπλέον να μεριμνήσει για την αποθήκευση των δίγλωσσων μετα-δεδομένων ως αποτέλεσμα της διαδικασίας στοίχισης.

Τα μετα-δεδομένα που αποθηκεύονται στο προτεινόμενο σχήμα της ΒΔ, αποτελούνται από τις ακόλουθες λογικές μονάδες πληροφορίας: (α) **Λέξεις:** όλες οι μορφές λέξεων όπως εμφανίζονται στα κείμενα, (β) **Λήμματα:** όλες οι μορφές λημμάτων

που προέρχονται από τις μορφές των λέξεων που υπάρχουν στα κείμενα, (γ) **Γραμματικά Χαρακτηριστικά:** τα μέρη του λόγου (Part-Of-Speech/POS) για την κάθε λέξη του κειμένου, (δ) **Προτάσεις:** βασικές δομικές μονάδες, (ε) **Κείμενα:** τα αρχεία που αποτελούν ένα σώμα κειμένου, (στ) **Σώματα Κειμένων:** συλλογή από κείμενα, (ζ) **Μεταφραστική Μνήμη:** σύνολο που σχετίζεται με μια συγκεκριμένη θεματική περιοχή και πιθανά ένα συγκεκριμένο χρήστη. Περιλαμβάνει όλες τις υπόλοιπες δομές και μπορεί να θεωρηθεί η υπερδομή του σχήματος της ΒΔ.

Ο χρήστης έχει τη δυνατότητα να παρέμβει τόσο με μαζικές όσο και με ελεγχόμενες (διαλογικές) διαδικασίες για την εισαγωγή νέων μεταφραστικών παραδειγμάτων στη ΒΔ καθώς και για τη διαχείριση της δομής της (δημιουργία, διαγραφή, εισαγωγή, ενημέρωση).

1.5 ΜΗΧΑΝΙΣΜΟΣ ΤΑΙΡΙΑΣΜΑΤΟΣ

Δύο είναι τα κυριότερα στοιχεία που λαμβάνονται υπόψη κατά το σχεδιασμό του μηχανισμού αναζήτησης και εξαγωγής του βέλτιστου ταιριάσματος: (1) Το μήκος των υπό σύγκριση τμημάτων κειμένου, ο καθορισμός, δηλαδή, του επιπέδου στο οποίο γίνεται η αναζήτηση και το ταίριασμα (επίπεδο μεγαλύτερο, ίσο ή μικρότερο από αυτό της πρότασης), (2) Η μετρική σύγκρισης και ομοιότητας δύο τμημάτων κειμένου.

Οι προτάσεις αποτελούν τη βασική μονάδα κειμένου κατά τη μεταφραστική διαδικασία. Αυτό γιατί, όχι μόνο τα όρια της πρότασης είναι σαφή, αλλά και γιατί η μετάφραση στο επίπεδο της πρότασης συνήθως ανταποκρίνεται στις ανάγκες του ανθρώπου-μεταφραστή. Βέβαια, καθώς το μήκος των προτάσεων αυξάνεται, η πιθανότητα να βρεθεί πλήρες ταίριασμα μειώνεται σημαντικά.

Εάν η μετάφραση πραγματοποιείται σε επίπεδο μικρότερο από αυτό της πρότασης (φράση, λέξη), είναι πολύ πιθανό ότι το τελικό αποτέλεσμα θα είναι χαμηλής ποιότητας. Τα βασικότερα προβλήματα που παρουσιάζονται σε μια τέτοια προσέγγιση είναι (α) η κατάλληλη επιλογή των τμημάτων υπό μετάφραση, (β) η επίλυση πιθανών αμφισημιών ιδιαίτερα στην περίπτωση μικρών τμημάτων κειμένου και τέλος, (γ) ο κατάλληλος συνδυασμός των επιμέρους μεταφράσεων με σκοπό την παραγωγή της τελικής μετάφρασης του κειμένου [15]-(Sato & Nagao 90), [12]-(Nirenburg et al. 93), [9]-(Kaji et al. 92).

Σχετικά με τη μετρική σύγκρισης των προτάσεων οι βασικές απαιτήσεις είναι δύο. (1) Η μετρική σύγκρισης μεταξύ δύο προτάσεων θα πρέπει να υποδηλώνει το βαθμό ομοιότητας αυτών, καθώς επίσης, (2) είναι επιθυμητό να σχετίζεται και να υποδεικνύει τις ομοιότητες και διαφορές οι οποίες οδήγησαν στον εν λόγω βαθμό ομοιότητας. Το τελευταίο αποτελεί κρίσιμο χαρακτηριστικό ενός **EBMT** συστήματος όπως επίσης και σημαντική βοήθεια προς το μεταφραστή.

Οι μετρικές που έχουν παρουσιασθεί στη βιβλιογραφία μπορούν να διαχωρισθούν με βάση το είδος των μονάδων κειμένου στα οποία εφαρμόζονται. Έτσι λοιπόν, οι μετρικές βασισμένες σε λέξεις συγκρίνουν μεμονωμένες λέξεις των υπό σύγκριση προτάσεων σε σχέση με τη μορφολογία τους, μέρος του λόγου, σημασιολογία (συνώνυμα, υπερώνυμα κλπ) [12]-(Nirenburg et al. 93), [16]-(Sumita & Iida 91). Ο τελικός βαθμός ομοιότητας δύο προτάσεων υπολογίζεται συνδυάζοντας κατάλληλα τις επιμέρους ομοιότητες των λέξεων που τις αποτελούν.

Οι μετρικές βασισμένες σε λέξεις είναι οι πιο διαδεδομένες. Παρόλα αυτά και άλλες εξίσου σημαντικές προσεγγίσεις έχουν παρουσιασθεί και οι οποίες βασίζονται σε άλλου επιπέδου πληροφορία, όπως χαρακτήρες [14]-(Sato 92), σύνταξη [17]-(Sumita & Tsutsumi 88), καθώς επίσης και ορισμένες υβριδικές τεχνικές που συνδυάζουν πληροφορία διαφορετικών επιπέδων [6]-(Furuse & Iida 92). Οι τεχνικές που βασίζονται στη σύνταξη των προτάσεων αν και δίνουν βέλτιστα αποτελέσματα (ιδιαίτερα σε συνδυασμό με τεχνικές ομοιότητας βασισμένες σε λέξεις), προϋποθέτουν συντακτική ανάλυση του κειμένου με σχετικά αμφίβολα αποτελέσματα.

Το τρίτο πρόβλημα που πρέπει να αντιμετωπίσει ένα **EBMT** σύστημα είναι αυτό της κατάλληλης εκμετάλλευσης των μεταφραστικών παραδειγμάτων της βάσης τα οποία έχουν κριθεί σχετικά με την πρόταση εισόδου από τον προηγούμενο μηχανισμό ταιριάσματος. Αυτό επιτυγχάνεται με υιοθέτηση τεχνικών δανεισμένων από το χώρο της μηχανικής μετάφρασης (**MT**) [16]-(Sumita & Iida 91), [9]-(Kaji et al. 92). Απλή τροποποίηση της προτεινόμενης μετάφρασης βασισμένη σε αντικατάσταση λέξεων ή όρων θα μπορούσε να αποτελέσει "μερική" λύση του προβλήματος, με την προϋπόθεση ότι οι μεταφράσεις αυτών των λέξεων ή όρων είναι διαθέσιμες είτε δια μέσου κατάλληλης στοίχισης των διαθέσιμων σωμάτων κειμένων, είτε δια μέσου των απαραίτητων ηλεκτρονικών λεξικών.

Ο μηχανισμός ταιριάσματος αποτελεί τον πυρήνα του συστήματος **TR•AID**. Έπειτα από κατάλληλη επεξεργασία των διαθέσιμων παραλλήλων σωμάτων κειμένων (λεκτική ανάλυση, αναγνώριση προτάσεων, στοίχιση), ο μηχανισμός ταιριάσματος αναλαμβάνει να εντοπίσει προτάσεις που ταυτίζονται ή τουλάχιστον μοιάζουν με την πρόταση εισόδου, καθώς επίσης και να εξάγει την αντίστοιχη μετάφραση.

Το ταίριασμα των προτάσεων αποτελείται από δυο επιμέρους λειτουργίες: (1) το πλήρες ταίριασμα κατά το οποίο το σύστημα εντοπίζει προτάσεις της βάσης δεδομένων που ταυτίζονται με την πρόταση εισόδου, και (2) το μερικό ταίριασμα, κατά το οποίο το σύστημα εντοπίζει όλες τις προτάσεις της ΒΔ οι οποίες φέρουν βαθμό ομοιότητας με την πρόταση εισόδου μεγαλύτερο του κατωφλίου ομοιότητας που έχει καθορίσει ο χρήστης.

1.5.1 ΜΗΧΑΝΙΣΜΟΣ ΠΛΗΡΟΥΣ ΤΑΙΡΙΑΣΜΑΤΟΣ

Η διαδικασία αυτή αφορά τον εντοπισμό των προτάσεων της ΒΔ που αποτελούν πλήρη ταιριάσματα της πρότασης εισόδου. Προκειμένου να επιτευχθεί αυτό γρήγορα και αποτελεσματικά χρησιμοποιείται στατιστική πληροφορία με σκοπό να εντοπισθεί ένα αρχικό μικρό σύνολο από υποψήφιες προτάσεις της ΒΔ μέσα στο οποίο θα βρίσκονται τελικά όλα τα πλήρη ταιριάσματα της πρότασης εισόδου, εάν βέβαια υπάρχουν. Επιπλέον, για να αντιμετωπιστούν μικρές διαφορές στο ταίριασμα και να ξεπεραστούν κατά κάποιο τρόπο προβλήματα ευελιξίας του συστήματος, η διαδικασία του τέλειου ταιριάσματος δεν λαμβάνει υπόψη της διαφορές σε λεκτικές μονάδες ορισμένου τύπου (ημερομηνίες, αριθμοί) έτσι ώστε τα γλωσσικά τέλεια ταιριάσματα να μην παραβλέπονται εξαιτίας μικρών διαφοροποιήσεων τέτοιου είδους.

Εάν δεν βρεθεί ένα τέλειο ταίριασμα, ο μηχανισμός ταιριάσματος ψάχνει τη ΒΔ με σκοπό να εντοπίσει προτάσεις παρόμοιες με την πρόταση εισόδου (μερικό ταίριασμα).

1.5.2 ΜΗΧΑΝΙΣΜΟΣ ΜΕΡΙΚΟΥ ΤΑΙΡΙΑΣΜΑΤΟΣ

Η συμβολή της δεύτερης φάσης του μηχανισμού ταιριάσματος έγκειται στην εύρεση μιας ή ενός συνόλου προτάσεων στη μεταφραστική μνήμη που να είναι όσο το δυνατό όμοιες με την πρόταση εισόδου. Η προσέγγιση που υιοθετείται βασίζεται στον υπολογισμό των κοινών στοιχείων μεταξύ των προτάσεων όπως επίσης και τον υπολογισμό κοινών περιοχών μεταξύ τους. Οι υπολογισμοί βασίζονται είτε στις λέξεις και τις θέσεις τους μέσα στις προτάσεις ή στα ζευγάρια λήμματα-γραμματικά χαρακτηριστικά για την κάθε λέξη και βέβαια τη θέση τους μέσα στις προτάσεις.

Για λόγους αποδοτικότητας, ο μηχανισμός του μερικού ταιριάσματος πραγματοποιείται σε δύο διαδοχικά στάδια: **(α)** εξαγωγή ενός μικρού συνόλου υποψηφίων προτάσεων από τη ΒΔ (μείωση του χώρου αναζήτησης), **(β)** διαδικασία μερικού ταιριάσματος.

Το πρώτο στάδιο αποβλέπει στον περιορισμό του τελικού χώρου αναζήτησης σε ένα μικρότερο σύνολο από υποψήφιες προτάσεις που μοιράζονται κάποια κοινά στοιχεία με την πρόταση εισόδου, βελτιώνοντας έτσι τη χρονική απόκριση του συστήματος. Χαρακτηριστικά όπως το μήκος της πρότασης, καθώς επίσης και κοινές λέξεις ή ομάδες λέξεων μεταβλητού μήκους, μεταξύ των προτάσεων λαμβάνονται υπόψη σε αυτό το στάδιο με σκοπό τη βελτίωση των αποτελεσμάτων.

Η συνεισφορά της διαδικασίας του μερικού ταιριάσματος έγκειται στην εξαγωγή του καλύτερου μερικού ταιριάσματος μέσα από το προηγούμενο σύνολο υποψηφίων προτάσεων. Κάθε πρόταση κωδικοποιείται σε ένα διάνυσμα και βασίζεται στα στοιχεία που η πρόταση περιέχει. Στη συνέχεια μια τεχνική ταυτοποίησης προτύπων δυναμικού

προγραμματισμού υπολογίζει ένα ποσοστό ομοιότητας (ταιριάσματος) για την κάθε πρόταση, βασιζόμενη στα κοινά σημεία, τις κοινές περιοχές και το μήκος των προτάσεων υπό εξέταση. Τα κοινά όπως επίσης και τα διαφορετικά στοιχεία των δύο προτάσεων που συνέβαλαν στη διαμόρφωση του ποσοστού ομοιότητας, εντοπίζονται και παρουσιάζονται στο χρήστη ώστε να μπορέσει να προσαρμόσει εύκολα την προτεινόμενη μετάφραση. Στην πιο απλή περίπτωση τα στοιχεία είναι οι λέξεις. Η διαδικασία μπορεί επίσης να συμπεριλάβει και μικρό βαθμό γλωσσολογικής πληροφορίας, όπου τότε τα στοιχεία είναι ζευγάρια λέξεων και λημμάτων (ή/και γραμματικών χαρακτηριστικών).

Στη συνέχεια παρουσιάζονται μερικά παραδείγματα μερικού ταιριάσματος μεταξύ δυο προτάσεων Π1,Π2 όπου τα Πα, Πβ, Πγ, Πδ δηλώνουν περιοχές των προτάσεων που αποτελούνται από τυχαίο αριθμό λέξεων.

Π1: Πα Πβ	**Π1:** Πα Πγ	**Π1:** Πα Πβ ***Πγ***	**Π1:** Πα ***Πβ Πγ***
Π2: Πα Πβ ***Πγ***	**Π2:** Πα **Πβ** Πγ	**Π2:** Πα Πβ ***Πδ***	**Π2:** Πα ***Πγ Πβ***

Εικόνα 2: Παραδείγματα περιπτώσεων μερικού ταιριάσματος

Επιπλέον πειραματισμοί έχουν πραγματοποιηθεί με σκοπό την ανάδειξη του καλύτερου μηχανισμού εύρεσης του ποσοστού ομοιότητας καθώς και του αλγορίθμου ταιριάσματος που χρησιμοποιείται. Ενδιαφέροντα αποτελέσματα παρατηρήθηκαν με τη χρήση ενός βελτιωμένου αλγόριθμου "απόστασης" συμβολοσειρών. Ο αλγόριθμος υπολογίζει το μικρότερο αριθμό αλλαγών (εισαγωγές, διαγραφές, αντικαταστάσεις, μετακινήσεις και αντιστροφές) προκειμένου να οδηγηθούμε, με μια αναστρέψιμη διαδικασία, από τη μια πρόταση στην άλλη. Το τελικό ποσοστό ομοιότητας υπολογίζεται αντιστοιχίζοντας κάποιους βαθμούς ποινής (κόστους) σε καθεμιά από αυτές τις αλλαγές. Παρόλο που η συγκεκριμένη μέθοδος πετυχαίνει μια πιο ολοκληρωμένη και λεπτομερή συγκριτική διαδικασία μεταξύ των δύο προτάσεων, είναι αμφίβολο το αν αποτελεί πιο αποδοτική λύση.

Σε περίπτωση που βρεθούν μερικά ταιριάσματα αποδεκτά από το χρήστη, ο χρήστης καλείται να μεταφράσει στη γλώσσα στόχο, τις περιοχές της πρότασης εισόδου οι οποίες δεν έχουν ταίριασμα. Το νέο ζευγάρι μετάφρασης μπορεί, αν αυτό είναι επιθυμητό, να αποθηκευτεί στη ΒΔ για μελλοντική χρήση. Σε περιπτώσεις όπου δεν έχει βρεθεί μερικό ταίριασμα, συμπεριλαμβανομένης και της περίπτωσης όπου μερικά ταιριάσματα έχουν βρεθεί αλλά το ποσοστό ομοιότητας δεν είναι αποδεκτό με βάση το όριο που έχει καθορίσει ο χρήστης, ο χρήστης καλείται να μεταφράσει εξ ολοκλήρου την προς μετάφραση πρόταση. Και σε αυτήν την περίπτωση το νέο ζευγάρι μπορεί να

αποθηκευτεί στη ΒΔ. Με αυτό τον τρόπο το σύστημα της μεταφραστικής μνήμης μαθαίνει συνεχώς νέα μεταφραστικά παραδείγματα, με ελεγχόμενο (διαλογικό) τρόπο.

1.6 ΕΝΤΟΠΙΣΜΟΣ ΚΑΙ ΜΕΤΑΦΡΑΣΗ ΟΡΟΛΟΓΙΑΣ

Εντοπισμός και μετάφραση όρων είναι δύο λειτουργίες που έχουν συμπεριληφθεί στο γενικό περιβάλλον του **TR•AID** σαν ένα ενδιάμεσο βήμα πριν την τελική μετάφραση ενός εγγράφου. Το εργαλείο εντοπίζει υποψήφιους όρους και τους αντικαθιστά με τις μεταφράσεις τους (αν υπάρχουν) στην επιθυμητή γλώσσα στόχο. Και στις δύο περιπτώσεις το σύστημα χρησιμοποιεί μια πολυγλωσσική ορολογική ΒΔ, για να αναγνωρίσει έναν όρο και στη συνέχεια να ανακτήσει τη μετάφρασή του. Το υπάρχον σχήμα της ΒΔ στοχεύει κυρίως στην αποδοτική αποθήκευση τόσο των μονολεκτικών όσο και των πολυλεκτικών όρων με άμεσο στόχο την αποτελεσματική ανάκτηση της πληροφορίας.

Το σύστημα συμβάλλει στο γρήγορο εντοπισμό μορφολογικών παραλλαγών των όρων που βρίσκονται αποθηκευμένοι στη ΒΔ, με τη βοήθεια μιας διαδικασίας "συσχέτισης όρων" (term conflation) [4]-(Frakes 84). Η "συσχέτιση" όρων πραγματοποιείται κατά το χρόνο αναζήτησης, επιτρέποντας έτσι την αποθήκευση μόνο των βασικών μορφών των όρων στη ΒΔ. Για λόγους απόδοσης, η διαδικασία εντοπισμού όρων πραγματοποιείται σε δυο διαδοχικά στάδια. Το πρώτο στάδιο, όπως και στο μηχανισμό μερικού ταιριάσματος των προτάσεων, αποβλέπει στην ελαχιστοποίηση του χώρου αναζήτησης, βελτιώνοντας έτσι το σύστημα σε θέματα απαιτήσεων μνήμης καθώς και χρόνου απόκρισης. Κατά τη διάρκεια της πρώτης φάσης το σύστημα εξάγει ένα μικρό σύνολο από υποψήφιους όρους βασιζόμενο σε στατιστική πληροφορία. Διαδοχικά, και κατά τη διάρκεια της δεύτερης φάσης, μια πιο πολύπλοκη διαδικασία αναλαμβάνει να ταξινομήσει με βάση το βαθμό ομοιότητας τους όρους που έχουν εντοπισθεί παράγοντας έτσι μια λίστα με όρους και ποσοστά ομοιότητας για τον κάθε πιθανό όρο του κειμένου εισόδου. Ο μηχανισμός βαθμολόγησης είναι βασισμένος σε ένα περιβάλλον δυναμικού προγραμματισμού ειδικά σχεδιασμένο να αναθέτει υψηλότερους βαθμούς ομοιότητας σε μορφολογικές διαφοροποιήσεις που έχουν την ίδια ρίζα-θέμα. Το σύστημα μπορεί να εντοπίσει τόσο μονολεκτικούς όσο και πολυλεκτικούς όρους, αγνοώντας κατά τη διαδικασία του ταιριάσματος ορισμένες κατηγορίες λέξεων (όπως άρθρα, συνδέσμους κλπ), εάν αυτές είναι διαθέσιμες.

Μια ενδιαφέρουσα πτυχή της διαδικασίας εντοπισμού-μετάφρασης της ορολογίας, η οποία βρίσκεται σε πειραματικά στάδια, αφορά την ενσωμάτωση και εναρμόνιση της με τη διαδικασία ταιριάσματος και μετάφρασης προτάσεων, δηλαδή τη χρήση της πληροφορίας ύπαρξης ενός όρου τόσο κατά τη φάση του ταιριάσματος όσο και κατά τη φάση παραγωγής της μετάφρασης.

2 ΣΥΜΠΕΡΑΣΜΑΤΙΚΑ ΣΧΟΛΙΑ

Το πραγματικό προτέρημα ενός συστήματος μετάφρασης είναι η δυνατότητα που παρέχει στο χρήστη να αυξάνει την αποδοτικότητα της μεταφραστικής διαδικασίας ελαττώνοντας το κόστος και το χρόνο, διατηρώντας παράλληλα την ποιότητα και τη συνέπεια της μετάφρασης. Πλήρως αυτοματοποιημένη μεταφραστική διαδικασία από Η/Υ δεν είναι ακόμα εφικτή. Ο στόχος είναι η ανάπτυξη ενός συστήματος το οποίο θα συνδυάζει διαφορετικά επίπεδα επεξεργασίας και πληροφορίας, θα είναι αρκετά αποδοτικό και θα προσαρμόζεται εύκολα και γρήγορα σε διαφορετικές φυσικές γλώσσες και διαφορετικές θεματικές περιοχές.

ΑΝΑΦΟΡΕΣ

[1] (Brown et al. 93) P. F Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer, *The Mathematics of Statistical Machine Translation: Parameter Estimation,* Computational Linguistics, June 1993.

[2] (Catizone et al. 89) R. Catizone, G. Russell, S. Warwick, *Deriving translation data from bilingual texts,* Proc. of the First Lexical Acquisition Workshop, Detroit 1989

[3] (Chanod & Tapanainen 96) J. P. Chanod and P. Tapanainen. *A non-deterministic tokenizer for finite-state parsing,* Proceedings of the ECAI 96 Workshop, 1996.

[4] (Frakes 84) W. B. Frakes *Term Conflation for Information Retrieval.* Research and Development in Information Retrieval, New York: Cambridge University Press, 1984.

[5] (Freibott 92) G.P. Freibott, *Computer Aided Translation in an Integrated Document Production Process:*Tools and Applications, Translating and the Computer 14, pp 45-66, 1992.

[6] (Furuse & Iida 92) O. Furuse and H. Iida, *Cooperation between Transfer and Analysis in Example-Based Framework.* Proc. Coling, pp 645-651, 1992.

[7] (Gale & Church 91) W. A. Gale and K. W. Church *A Program for Aligning Sentences in Bilingual Corpora.* Proc. of the 29th Annual Meeting of the ACL., pp 177-184, 1991.

[8] (Grefenstette & Tapanainen 94) G. Grefenstette and P. Tapanainen *What is a word, What is a sentence? Problems of tokenization,* COMPLEX 94.

[9] (Kaji et al. 92) H. Kaji, Y. Kida and Y. Morimoto, *Learning Translation Templates from Bilingual Text.* Proc. Coling., pp 672-678, 1992.

[10] (Kay & Roscheisen 91) M. Kay, M. Roscheisen, *Text-Translation Alignment,* Computational Linguistics Vol. 19, No 1, 1991.

[11] (Nagao 84) M. Nagao, *A framework of a mechanical translation between Japanese and English by analogy principle.* Artificial and Human Intelligence, ed. Elithorn A. and Banerji R., North-Holland, pp 173-180, 1984.

[12] (Nirenburg et al. 93) S. Nirenburg, C. Domashnev D. J. Grannes. *Two Approaches to Matching in Example-Based Machine Translation.* Proc. of TMI-93, Kyoto, Japan, 1993.

[13] (Reynar & Ratnaparkhi 97) J. C. Reynar and A. Ratnaparkhi, *A maximum entropy approach to identifying sentence boundaries,* Computational Linguistics Archive cmp-lg/9704002, 1997.

[14] (Sato 92) S. Sato, *CTM: An Example-Based Translation Aid System.* Proc. of Coling, pp 1259-1263, 1992.

[15] (Sato & Nagao 90) S. Sato and M. Nagao, *Toward Memory-based Translation.* Proc. of Coling, pp 247-252, 1990.

[16] (Sumita & Iida 91) E. Sumita and H. Iida, *Experiments and Prospects of Example-based Machine Translation.* Proc. of the 29th Annual Meeting of the Association for Computational Linguistics, pp 185-192, 1991.

[17] (Sumita & Tsutsumi 88), E. Sumita and Y. Tsutsumi, *A Translation Aid System Using Flexible Text Retrieval Based on Syntax-Matching.* TRL Research Report, Tokyo Research Laboratory, IBM, 1988.

Ιωάννης Τριανταφύλλου, Χρήστος Μαλαβάζος, Στέλιος Πιπερίδης - Ερευνητές ΙΕΛ

Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ)
Τμήμα Γλωσσικών Εφαρμογών
Αρτέμιδος & Επιδαύρου, 15125, Μαρούσι
yiannis, christos, spip@ilsp.gr