ΕΛΕΤΟ – 1° Συνέδριο «Ελληνική Γλώσσα και Ορολογία»
Αθήνα, 30, 31 Οκτωβρίου και 1 Νοεμβρίου 1997

ΑΥΤΟΜΑΤΗ ΚΑΤΑΣΚΕΥΗ ΔΙΓΛΩΣΣΩΝ ΛΕΞΙΚΩΝ

Σ. Πιπερίδης, Σ. Μπούτσης, Ι. Δεμοίρος

Περίληψη

Το άρθρο αυτό περιγράφει μία μέθοδο για την αυτόματη κατασκευή δίγλωσσων λεξικών από παράλληλα δίγλωσσα κείμενα με στατιστικές τεχνικές. Εχει καταβληθεί προσπάθεια η χρησιμοποιούμενη γλωσσική πληροφορία να διατηρηθεί στο ελάχιστο, προκειμένου η μέθοδος να είναι όσο το δυνατόν ανεξάρτητη από τα γλωσσικά ζεύγη για τα οποία αρχικά αναπτύχθηκε. Η μέθοδος δέχεται στην είσοδο δίγλωσσα παράλληλα κείμενα και εξάγει ισοδυναμίες μεταξύ λέξεων ή ομάδων λέξεων των κειμένων αυτών. Τα παράλληλα κείμενα παραλληλοποιούνται στο επίπεδο της πρότασης με τη βοήθεια στατιστικών αλγορίθμων και στη συνέχεια αναλύονται γραμματικά και συντακτικά με σκοπό την εξαγωγή ονοματικών φράσεων από αυτά και την αναγνώριση των πλέον στατιστικά και γλωσσικά έγκυρων πολυλεκτικών μονάδων. Οι πιθανές μεταφράσεις μεταξύ των δύο γλωσσών διερευνούνται και αξιολογούνται από μία μετρική που συνδυάζει συχνότητες εμφάνισης και συνεμφάνισης των σχετικών μεταφραστικών μονάδων. Η μέθοδος έχει εφαρμοστεί σε ένα μικρό σώμα κειμένων σχετικό με συστήματα λογισμικού δίνοντας αποτελέσματα που είναι σωστά κατά 94% περίπου, προτείνοντας μεταφράσεις για το 50% των λέξεων του Ελληνικού κειμένου και το 40% του Αγγλικού.

AUTOMATIC BILINGUAL LEXICON BUILDING

S.Piperidis, S. Boutsis, J. Demiros

Abstract

This paper describes a method for the automatic construction of a translation lexicon from hybrid parallel texts. The method features mainly statistical techniques, while effort has been made to invoke the least required language specific information in an attempt to implement as language independent a method as possible. The method presupposes parallel texts and identifies translational equivalences at the word or multi-word unit level for those cases that such an equivalence holds true. Parallel texts are first statistically aligned at sentence level and then lemmatised and tagged with their part-of-speech categories. Noun phrase grammars operating on tag sequences extract noun phrases on both sides of the parallel corpus and statistical evaluation yields the most coherent, statistically relevant multi-word units on either side. Translation candidates of word- or multi-word units are then evaluated by a similarity metric defined by the co-occurrence frequency and independent frequencies of the units. The method has been tested on a small English-Greek corpus consisting of texts relevant to software systems, yielding approximately 94% accurate translations, while proposing translations for approximately 50% of all occurrences of content words of the English side and 40% of the Greek side of the corpus.

0 ΕΙΣΑΓΩΓΗ

Η ύπαρξη δίγλωσσων κειμένων σε ηλεκτρονική μορφή έχει δώσει τη δυνατότητα να αναπτυχθούν μία σειρά από εφαρμογές που στοχεύουν στην αυτόματη ή ημιαυτόματη εξαγωγή από τα κείμενα αυτά γλωσσικών πόρων κατάλληλων για τη σύνταξη πολύγλωσσων λεξικών, γραμματικών και συντακτικών κανόνων μεταφοράς για συστήματα αυτόματης μετάφρασης και μεταφραστικών παραδειγμάτων, [4], [10], [3].

Στόχος αυτής της εργασίας είναι η περιγραφή μιας τεχνικής για την κατασκευή ενός δίγλωσσου λεξικού μεταφραστικών ισοδυναμιών μεταξύ λέξεων και ομάδων λέξεων. Οι τεχνικές που χρησιμοποιούνται για τον σκοπό αυτό βασίζονται σε στατιστική ανάλυση σε συνδυασμό με τεχνικές γλωσσολογικής επεξεργασίας που συμβάλλουν στην αναγνώριση των πιθανών μεταφραστικών μονάδων σε κάθε γλώσσα του παράλληλου κειμένου. Η στατιστική επεξεργασία έχει αποδειχτεί κατάλληλη για την εξαγωγή μεταφραστικών ισοδυναμιών στο επίπεδο της πρότασης, [5]. Στο [2] χρησιμοποιείται μία πιθανοτική μετρική για την εκτίμηση της συσχέτισης των λέξεων μεταξύ δύο γλωσσών στο περιβάλλον της στατιστικής μετάφρασης. Στο [7] παρουσιάζεται ένας αλγόριθμος για την παραλληλοποίηση κειμένων με βάση μόνο την εσωτερική τους μαρτυρία και ο οποίος παράγει αντιστοιχίες τόσο μεταξύ προτάσεων όσο και λέξεων. Η επεξεργασία λαμβάνει χώρα σε πολλές επαναλήψεις και κάθε νέα επανάληψη υπολογίζει καλύτερες προσεγγίσεις αυτών των αντιστοιχιών χρησιμοποιώντας τα αποτελέσματα των προηγουμένων. Στο [8] προτείνεται η μετρική "dice" για να εκφράσει την ομοιότητα μεταξύ λέξεων με σκοπό τη βελτίωση της παραλληλοποίησης στο επίπεδο της πρότασης. Η ίδια μετρική χρησιμοποιείται επίσης στο [9] για την εξαγωγή ισοδυναμιών μεταξύ αγγλικών και ιαπωνικών λέξεων. Στο [1] περιγράφεται μία μέθοδος για την εξαγωγή μονολεκτικών ισοδυναμιών με βάση μία μετρική με παρόμοια απόδοση με τη μετρική "dice" από κείμενα που έχουν παραλληλοποιηθεί στο επίπεδο της πρότασης με στατιστικές μεθόδους.

Στην παρούσα εργασία προτείνουμε μία μέθοδο που επεξεργάζεται δίγλωσσα κείμενα τα οποία είναι παραλληλοποιημένα στο επίπεδο της πρότασης. Με τη χρήση στατιστικών και γλωσσολογικών τεχνικών η μέθοδος εξετάζει και αναγνωρίζει πρότυπα και στις δύο γλώσσες του δίγλωσσου σώματος κειμένων τα οποία εκφράζουν αντιστοιχίες μεταξύ λέξεων και πολυλεκτικών μονάδων.

Η βασική υπόθεση στην οποία βασίζεται η μέθοδος είναι ότι ζεύγη λέξεων, οι οποίες είναι μετάφραση η μία της άλλης, συχνά εμφανίζονται σε αντίστοιχες θέσεις στο κείμενο πηγή και στο κείμενο στόχο, δηλαδή εμφανίζονται σε αντίστοιχες

προτάσεις ή ομάδες προτάσεων του παραλληλοποιημένου δίγλωσσου κειμένου. Ασφαλώς, κατά τη διαδικασία της μετάφρασης η σειρά των λέξεων δεν παραμένει η ίδια και η μετάφραση δε νοείται να είναι "ένα προς ένα". Επίσης, η επίδραση των συμφραζομένων είναι αρκετά ισχυρή και πολλές λέξεις αλλάζουν σημασία ανάλογα με τα συμφραζόμενά τους. Η μέθοδος που προτείνεται μπορεί να αντιμετωπίσει αποτελεσματικά μόνο περιπτώσεις λέξεων που μεταφράζονται με συνεπή τρόπο από τη μία γλώσσα στην άλλη. Επειδή η ορολογία (όροι) μεταφράζεται(ονται) συνήθως με τρόπο συνεπή, κείμενα που περιέχουν ορολογία από το ίδιο γνωστικό πεδίο αναμένεται να προσφέρονται για επεξεργασία με αυτή τη μέθοδο.

Είναι προφανές ότι οι συσχετίσεις των λέξεων που παράγονται από τη μέθοδο εξαρτώνται από το είδος του κειμένου από το οποίο προκύπτουν. Μπορούμε να αναμένουμε ότι η μετάφραση που προκύπτει για μία λέξη από ένα κείμενο δε θα είναι ίδια με τη μετάφραση που ίσως προκύψει από ένα άλλο, όταν τα γνωστικά πεδία των δύο κειμένων είναι διαφορετικά.

1 ΣΧΕΔΙΑΓΡΑΜΜΑ ΤΗΣ ΜΕΘΟΔΟΥ

Η εξαγωγή των δίγλωσσων ισοδυναμιών πραγματοποιείται από μια ακολουθία αυτόνομων διεργασιών. Η έξοδος μιας διεργασίας οδηγείται στην είσοδο της επόμενης. Η μέθοδος παίρνει στην είσοδο παράλληλα κείμενα και δίνει στην έξοδο τις δίγλωσσες ισοδυναμίες που έχει αναγνωρίσει. Στο πρώτο στάδιο, τα δίγλωσσα κείμενα παραλληλοποιούνται στο επίπεδο της πρότασης με τη χρήση στατιστικών τεχνικών. Στη συνέχεια μέσω γραμματικής ανάλυσης υπολογίζεται το μέρος του λόγου κάθε λέξης όπως και το λήμμα της. Μέσω επιφανειακής συντακτικής ανάλυσης, εξάγονται οι ονοματικές φράσεις του κειμένου και αξιολογούνται στη συνέχεια στατιστικά. Κατά το στάδιο της συνδυαστικής επεξεργασίας οι λεκτικές μονάδες συσχετίζονται με τις πιθανές μεταφράσεις τους στην άλλη γλώσσα και μία μετρική υπολογίζεται για την αξιολόγηση όλων των πιθανών συσχετίσεων. Στο τελευταίο στάδιο λαμβάνει χώρα μία επιλογή βασισμένη σε κανόνες δίνοντας στην έξοδο αντιστοιχίες που μπορεί να είναι '1-1' ή 'n-m' στην γενική περίπτωση.

2 ΣΩΜΑ ΚΕΙΜΕΝΩΝ

Το σώμα των κειμένων που χρησιμοποιήθηκε για την ανάπτυξη της προτεινόμενης μεθόδου αποτελείται από τα έγγραφα τεκμηρίωσης της πλατφόρμας λογισμικού HP-VUE της εταιρείας HP.

Το Ελληνικό κείμενο περιέχει 35726 λέξεις και το Αγγλικό 28872 λέξεις. Ο αριθμός των διαφορετικών λέξεων για το Ελληνικό κείμενο είναι 4512 και για το

Αγγλικό 3219. Το γεγονός ότι η μορφολογία της Ελληνικής είναι πιο πλούσια από της Αγγλικής εξηγεί την κατά 30% διαφορά των δύο τελευταίων μεγεθών. Επιπλέον το Ελληνικό κείμενο περιέχει 2588 λήμματα και το Αγγλικό 2111.

3 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΜΕΘΟΔΟΥ

3.1 ΠΑΡΑΛΛΗΛΟΠΟΙΗΣΗ ΚΕΙΜΕΝΩΝ

Η παραλληλοποίηση των κειμένων είναι το πρώτο βήμα προς την εξαγωγή του δίγλωσσου λεξικού. Το δίγλωσσο σώμα κειμένων αναδιοργανώνεται σε μία ακολουθία δίγλωσσων αντιστοιχιών μεταξύ των προτάσεων των δύο γλωσσών.

Η παραλληλοποίηση στο επίπεδο της πρότασης βασίζεται σε στατιστική επεξεργασία. Ο στατιστικός παραλληλοποιητής κειμένων, όπως λέγεται το σχετικό προγράμμα, βασίζεται σε απλά στατιστικά μοντέλα του μήκους του κειμένου σε χαρακτήρες ή λέξεις. Το μοντέλο στηρίζεται στην παρατήρηση ότι μεγαλύτερες προτάσεις του κειμένου πηγή τείνουν να μεταφράζονται σε μεγαλύτερες προτάσεις του κειμένου στόχος και αντίστροφα [6]. Ενα πιθανοτικό σκορ υπολογίζεται για κάθε προτεινόμενο ζεύγος παράλληλων προτάσεων και ένα μοντέλο δυναμικού προγραμματισμού επιλέγει την καλύτερη παραλληλοποίηση για τις προτάσεις των δύο κειμένων.

3.2 ΓΡΑΜΜΑΤΙΚΗ ΑΝΑΛΥΣΗ

Τόσο το Ελληνικό όσο και το Αγγλικό κείμενο αναλύονται γραμματικά. Οι λέξεις χαρακτηρίζονται με το αντίστοιχο λήμμα και με το μέρος του λόγου στο οποίο ανήκουν.

Η γραμματική ανάλυση για την Ελληνική γλώσσα βασίζεται σε ένα μορφολογικό λεξικό και σε ένα σύστημα κανόνων για άρση της αμφισημίας. Κάθε λέξη λαμβάνει γραμματικά χαρακτηριστικά από την αντίστοιχη καταχώρηση του λεξικού. Σε περίπτωση που υπάρχουν περισσότερες της μιας γραμματικές αναλύσεις για μία λέξη, ο γραμματικός αναλυτής επιλέγει μία ανάλυση με βάση τα συμφραζόμενα.

Ο Αγγλικός γραμματικός αναλυτής χρησιμοποιεί επίσης ένα λεξικό για να αποδώσει σε κάθε λέξη έναν ή περισσότερους γραμματικούς χαρακτηρισμούς. Στην περίπτωση που η λέξη δεν υπάρχει στο λεξικό, μία διαδικασία μορφολογικής ανάλυσης παράγει μία εκτίμηση των πιθανών χαρακτηρισμών. Η περίπτωση αμφισημίας αντιμετωπίζεται από ένα σύστημα άρσης της αμφισημίας που κάνει χρήση στατιστικών κανόνων.

3.3 ΕΞΑΓΩΓΗ ΟΝΟΜΑΤΙΚΩΝ ΦΡΑΣΕΩΝ

Ο γραμματικός χαρακτηρισμός που προκύπτει από το προηγούμενο στάδιο, δίνει τη δυνατότητα αναγνώρισης ονοματικών φράσεων που ακολουθούν συγκεκριμένα συντακτικά πρότυπα τα οποία είναι χαρακτηριστικά του σχηματισμού όρων. Οι ονοματικές φράσεις που εξάγονται τυγχάνουν στατιστικής επεξεργασίας προκειμένου να εξεταστεί το κατά πόσο αποτελούν όρους του τεχνικού πεδίου από το οποίο προέρχεται το σώμα κειμένων. Η επεξεργασία σε αυτό το στάδιο είναι μονόγλωσση και το κείμενο κάθε γλώσσας εξετάζεται ανεξάρτητα από το παράλληλό του στην άλλη γλώσσα. Η γραμματική για την εξαγωγή Αγγλικών ονοματικών φράσεων έχει ως εξής:

ng -> [n]+ , adjg -> [adj]+ , (1)

np -> (adjg) ng , np -> np prep np

Οι Ελληνικές ονοματικές φράσεις περιγράφονται από τους κανόνες:

ng -> [n]+ , adjg -> [adj]+ , (2)

np -> (adjg) ng ,

np -> np (det(case:genitive)) np(case:genitive) .

Με παρενθέσεις "()" υποδηλώνεται ότι το σχετικό συστατικό είναι προαιρετικό ενώ το "+" δείχνει ότι το σχετικό συστατικό μπορεί να επαναληφθεί μία ή περισσότερες φορές. Στην περίπτωση που περισσότερες από μία ονοματικές φράσεις διαφορετικού μήκους μπορούν να εξαχθούν από το ίδιο τμήμα κειμένου, όλες οι εναλλακτικές φράσεις λαμβάνονται υπόψη, δηλαδή όλες οι ονοματικές φράσεις εξάγονται, από τα μεμονωμένα ουσιαστικά και τις μικρότερες σε μήκος ονοματικές φράσεις ως και τις μεγαλύτερες σε μήκος φράσεις.

Αφού εξαχθούν οι ονοματικές φράσεις από τα κείμενα των δύο γλωσσών με βάση τα παραπάνω κριτήρια, με επιπρόσθετη στατιστική επεξεργασία εξετάζεται η εγκυρότητα των όρων και των πολυλεκτικών μονάδων καθώς και η συνοχή τους.

Για κάθε πολυλεκτική μονάδα $w_1^n = w_1..w_n$, υπολογίζεται η πιθανότητα $P(w_1^n)$ με βάση την ακόλουθη φόρμουλα:

$$P(w_1^n) = P(w_n|w_1^{n-1}) \cdot P(w_1^{n-1}) = \frac{C(w_1^{n-1}w_n)}{\sum_w C(w_1^{n-1}w)} \cdot P(w_1^{n-1}), \qquad (3)$$

όπου C(w) είναι η συχνότητα εμφάνισης της λέξης w στο κείμενο. Οι πλέον έγκυρες πολυλεκτικές μονάδες οδηγούνται στο επόμενο στάδιο.

3.4 ΣΥΝΔΥΑΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ

Η κύρια υπόθεση στην οποία βασίζεται η μέθοδος είναι ότι ζεύγη μεταφραστικών μονάδων, οι οποίες είναι μετάφραση η μία της άλλης, εμφανίζονται συχνά σε αντίστοιχες προτάσεις ή ομάδες προτάσεων του παραλληλοποιημένου δίγλωσσου κειμένου. Για παράδειγμα, ας υποθέσουμε ότι η μεταφραστική μονάδα X μεταφράζεται με συνέπεια στη μεταφραστική μονάδα Y. Ως εκ τούτου, όταν συναντάται η μεταφραστική μονάδα X σε μία πρόταση του κειμένου πηγή, αναμένουμε, στις περισσότερες των περιπτώσεων, να υπάρχει η μονάδα Y στην αντίστοιχη πρόταση του κειμένου στόχος. Οι αντιστοιχίες μεταξύ των προτάσεων των δύο κειμένων είναι ήδη γνωστές, καθώς τα κείμενα έχουν παραλληλοποιηθεί σε προηγούμενο στάδιο με στατιστικές μεθόδους (παράγραφος 3).

Ας υποθέσουμε ότι μία πρόταση του κειμένου πηγή είναι η (S1, S2, X, S3) και η αντίστοιχη πρόταση του κειμένου στόχος (T1, Y, T2). Για αυτές τις δύο προτάσεις σχηματίζουμε τα ζεύγη:
(S1, T1), (S1, Y), (S1, T2), (S2, T1), (S2, Y), (S2, T2), (X, T1), (X, Y), (X, T2), (S3, T1), (S3, Y), (S3, T2)

Δεδομένου ότι τα X και Y δεν εμφανίζονται με τα ίδια συμφραζόμενα (S1, S2, _, S3) και (T1, _, T2) σε όλο το μήκος του κειμένου, αναμένεται ότι τα ζεύγη (X, Y) θα είναι πολυάριθμα σε αντίθεση με τα ζεύγη π.χ. (S1, Y). Με αυτό τον τρόπο αναδεικνύεται η σωστή μετάφραση, υπό την προϋπόθεση ότι τα X και Y εμφανίζονται αρκετά συχνά.

3.5 ΥΠΟΛΟΓΙΣΜΟΣ ΜΕΤΡΙΚΗΣ

Αυτό το στάδιο στοχεύει στον προσδιορισμό των πιο πιθανών μεταφράσεων για κάθε λέξη επιλέγοντας από τα ζεύγη του προηγούμενου βήματος. Η διαθέσιμη πληροφορία για αυτό το σκοπό είναι: οι συχνότητες των λέξεων του κειμένου πηγή, οι συχνότητες των λέξεων του κειμένου στόχος και τα ζεύγη που έχουν σχηματιστεί με τον τρόπο που αναφέρθηκε στην προηγούμενη παράγραφο. Με χρησιμοποίηση αυτής της πληροφορίας υπολογίζουμε μια μετρική που εκφράζει την πιθανότητα ένα ζεύγος μεταφραστικών μονάδων να είναι ορθό μεταφραστικό ζεύγος.

Η μετρική πρέπει να προσδιορίζει με ακρίβεια τις λέξεις που μεταφράζονται αυτόνομα και με συνέπεια σε όλο το μήκος των κειμένων. Σε αυτήν την περίπτωση η ακόλουθη παρατήρηση είναι ορθή και μπορεί να οδηγήσει στη διαμόρφωση της μετρικής:

"Αν το ζεύγος (X, Y) είναι έγκυρη λεκτική ισοδυναμία τότε οι συχνότητες των μονάδων X και Y, f(X) και f(Y) αντίστοιχα, είναι της ίδιας τάξης μεγέθους. Της ίδιας τάξης μεγέθους είναι επίσης η συχνότητα f(X, Y) του ζεύγους (X, Y)."

Εισάγουμε λοιπόν τη μετρική:

$$M(x, y) = \frac{\sqrt{\left(fx - m\right)^2 + \left(fy - m\right)^2 + \left(fxy - m\right)^2}}{m} \quad (4), \quad \text{όπου} \quad m = \frac{fx + fy + fxy}{3} \quad (5).$$

Η μετρική αυτή δίνει μικρές τιμές όταν οι τιμές f(X), f(Y) και f(X, Y) είναι "αριθμητικά κοντά" η μία στην άλλη. Οταν οι f(X) και f(Y) είναι αρκετά διαφορετικές ή στην περίπτωση που είναι στην ίδια περιοχή τιμών αλλά η f(X, Y) είναι σημαντικά διαφορετική, η M(X, Y) παίρνει μεγάλες τιμές.

Με τη χρήση αυτής της μετρικής είναι δυνατόν για κάθε όρο να εντοπιστούν οι πιο πιθανές μεταφράσεις, οι οποίες είναι οι μόνες που προωθούνται στο επόμενο στάδιο.

Για παράδειγμα ο ελληνικός όρος "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ" που εμφανίζεται 20 φορές στο Ελληνικό κείμενο συσχετίζεται με τις Αγγλικές λέξεις του ακόλουθου πίνακα:

Score	Score % Variation	f(x,y)	Target Word	Target Word Freq.
0.205	0.000000	15.00	DATABASE-FILE	17
0.690	70.246948	7.00	HP-VUE-SYNTAX	13
0.720	4.162519	8.00	CONVERT	26
0.994	27.614494	3.00	SEARCH	23
1.049	5.238557	2.00	LOCATE	20
1.055	0.519161	2.00	DIALOG-BOX	16
1.065	0.954310	2.00	APPEAR	15
1.135	6.140458	1.00	EASY	20
1.143	0.695868	3.00	EXISTING	9
1.157	0.580395	2.00	NEW-ACTION	11

Η πρώτη στήλη δείχνει τις τιμές της μετρικής που υπολογίζονται για τα ζεύγη που σχηματίζονται από τον όρο "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ" και από τις λέξεις της τέταρτης στήλης. Η δεύτερη στήλη δείχνει τη διαφορά στην τιμή του σκορ που αντιστοιχεί στην τρέχουσα γραμμή και στην προηγούμενη. Η τρίτη στήλη δείχνει τη συχνότητα του τρέχοντος ζεύγους, ενώ η τελευταία στήλη τις συχνότητες των λέξεων που σχετίζονται με τον Ελληνικό όρο.

Στο συγκεκριμένο παράδειγμα παρατηρούμε ότι η τιμή της μετρικής παίρνει τη μικρότερη τιμή της για τον Αγγλικό όρο "DATABASE-FILE", ενώ οι τιμές της μετρικής για τις επόμενες γραμμές του πίνακα είναι μεγαλύτερες κατά 70% περίπου. Το

γεγονός αυτό επιτρέπει να συμπεράνουμε ότι το ζεύγος "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ", "DATABASE-FILE" είναι μία έγκυρη δίγλωσση λεκτική ισοδυναμία.

3.6 ΕΠΙΛΟΓΗ ΜΕΤΑΦΡΑΣΕΩΝ

Στόχος αυτού του σταδίου είναι να επιλεχτεί μία μεταφραστική μονάδα (λέξη, όρος, πολυλεκτική μονάδα κλπ.) του κειμένου στόχος για όσες μεταφραστικές μονάδες του κειμένου πηγή είναι δυνατό, έτσι ώστε τα ζεύγη που θα σχηματιστούν να είναι έγκυρες δίγλωσσες λεκτικές ισοδυναμίες.

Η διαδικασία για την εξαγωγή των αντιστοιχιών είναι η εξής: Για κάθε μεταφραστική μονάδα της γλώσσας πηγής (Αγγλική) ανακαλείται από τη βάση δεδομένων ο πίνακας με τις σχετικές μεταφραστικές μονάδες της γλώσσας στόχου (Ελληνική). Ο πίνακας εξετάζεται από πάνω προς τα κάτω και αποτιμούνται οι ακόλουθες εκφράσεις:

(6) συχνότητα μεταφραστικής μονάδας της γλώσσας πηγής / συχνότητα ζεύγους

(7) συχνότητα μεταφραστικής μονάδας της γλώσσας στόχου / συχνότητα ζεύγους

(8) συχνότητα ζεύγους / συχνότητα επόμενου ζεύγους

Αν οι εκφράσεις αυτές, αφού αποτιμηθούν για το πρώτο ζεύγος του πίνακα, βρίσκονται σε μία συγκεκριμένη περιοχή τιμών, μία αντιστοιχία είναι πιθανό να υπάρχει. Τότε ο πίνακας με τις πιθανές μεταφράσεις της σχετικής μεταφραστικής μονάδας της γλώσσας στόχου ανακαλείται και υπολογίζονται οι (6), (7) και (8). Αν οι τιμές τους βρίσκονται στο προκαθορισμένο όριο τιμών μία έγκυρη μεταφραστική ισοδυναμία τυπώνεται στην έξοδο.

Κατά την επεξεργασία που περιγράφεται ανωτέρω, αν μία από τις (6) - (8) δε βρίσκεται στην κατάλληλη περιοχή τιμών, η επεξεργασία σταματά για τη σχετική λέξη και συνεχίζει με την επόμενη. Μπορεί να σημειωθεί ότι μία έγκυρη ισοδυναμία παράγεται όταν μία λέξη του κειμένου πηγή αντιστοιχίζεται σε μία λέξη του κειμένου στόχος, που αντιστοιχίζεται στην ίδια λέξη του κειμένου πηγή.

Οι περιοχές τιμών, στις οποίες πρέπει να κυμαίνονται οι (6), (7) λήφθηκαν: (6)<2 και (7)<2. Αυτό είναι κατάλληλο επειδή δύο μεταφραστικές μονάδες που είναι μετάφραση η μία της άλλης, μπορεί να αναμένεται να μεταφράζονται κατ' αυτόν τον τρόπο τουλάχιστον τις μισές φορές που εμφανίζονται στο κείμενο.

Για τον υπολογισμό του εύρους στο οποίο κυμαίνεται η μετρική (8), οι σχετικοί πίνακες εξετάστηκαν για έναν αριθμό τυχαία επιλεγμένων όρων και υπολογίστηκαν τα σχετικά μεγέθη για τις περιπτώσεις των σωστών μεταφράσεων. Η τελική τιμή λαμβάνεται όταν η εξέταση επιπλέον παραδειγμάτων δε μεταβάλλει τις παραμέτρους του μοντέλου. Με αυτή τη μέθοδο υπολογίστηκε ότι (8)<0.2.

4 ΑΠΟΤΕΛΕΣΜΑΤΑ

Αφού η μέθοδος εφαρμόστηκε στο σώμα κειμένων που περιγράφτηκε στην παράγραφο (3), προέκυψαν 282 μεταφράσεις. Το ποσοστό επιτυχίας ήταν 93.6%. Οι 282 μεταφράσεις αντιστοιχούν σε 10972 εμφανίσεις στο Ελληνικό κείμενο και 11640 στο Αγγλικό. Μερικές από τις μεταφράσεις είναι οι εξής (οι υπογραμμισμένες μεταφράσεις είναι εσφαλμένες):

ΑΓΓΛΙΚΟΣ <-> ENGLISH
ΑΓΝΟΩ <-> IGNORE
ΑΚΕΡΑΙΟΣ <-> INTEGER
ΑΝΑΓΚΑΙΟΣ <-> NECESSARY
ΑΝΑΓΝΩΡΙΖΩ <-> IDENTIFY
ΑΝΑΖΗΤΗΣΗ <-> SEARCH
ΑΝΑΘΕΤΩ <-> ASSIGN

ΑΝΑΛΥΣΗ <-> RESOLUTION
ΑΝΑΜΕΣΑ <-> BETWEEN
ΑΝΑΦΕΡΩ <-> REFER
ΑΝΑΦΟΡΑ <-> REFERENCE
ΑΝΕΞΑΡΤΗΤΑ <-> REGARDLESS
ΑΝΟΙΓΩ <-> OPEN
ΑΝΤΙΓΡΑΦΗ <-> CUSTOM

...

ΑΛΦΑΡΙΘΜΗΤΙΚΟ-ΕΚΤΕΛΕΣΗ <-> EXECUTION-STRING
ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ <-> DATABASE-FILE
ΑΡΧΕΙΟ-ΔΕΔΟΜΕΝΟ <-> DATA-FILE
ΑΡΧΕΙΟ-ΔΙΑΜΟΡΦΩΣΗ <-> CONFIGURATION-FILE
ΑΡΧΕΙΟ-ΣΤΟΙΧΕΙΟ-ΣΥΜΠΕΡΙΦΟΡΑ <-> RESOURCE-FILE
ΓΕΝΙΚΟΣ-ΕΡΓΑΛΕΙΟΘΗΚΗ <-> GENERAL-TOOLBOX
ΔΙΑΧΕΙΡΙΣΤΗΣ-ΑΡΧΕΙΟ <-> FILE-MANAGER
ΔΙΑΧΕΙΡΙΣΤΗΣ-ΠΕΡΙΟΔΟΣ-ΕΠΙΚΟΙΝΩΝΙΑ <-> SESSION-MANAGER
ΔΙΑΧΕΙΡΙΣΤΗΣ-ΥΦΟΣ <-> STYLE-MANAGER
ΔΙΑΧΕΙΡΙΣΤΗΣ-ΧΩΡΟΣ-ΕΡΓΑΣΙΑ <-> WORKSPACE-MANAGER
ΕΠΙΦΑΝΕΙΑ-ΧΩΡΟΣ-ΕΡΓΑΣΙΑ <-> DESKTOP

5 ΣΥΜΠΕΡΑΣΜΑ

Η εργασία αυτή παρουσίασε μία μέθοδο για την αυτόματη κατασκευή ενός δίγλωσσου λεξικού από παράλληλα κείμενα, με τη βοήθεια τεχνικών που στηρίζονται σε στατιστική και γλωσσολογική επεξεργασία και που επιτρέπουν την αντιμετώπιση του προβλήματος της αναγνώρισης μεταφραστικών μονάδων. Οι τεχνικές αυτές αποδεικνύονται ιδιαίτερα χρήσιμες στην κατασκευή πολυγλωσσικών λεξικών για συγκεκριμένα γνωστικά πεδία με βάση την πραγματική χρήση της σχετικής ορολογίας σε παράλληλα κείμενα αυτών των πεδίων. Η μέθοδος που παρουσιάστηκε εφαρμόστηκε σε ένα μικρό σώμα κειμένων αποτελούμενο από έγγραφα τεκμηρίωσης λογισμικού. Θα μπορούσε εξίσου κατάλληλα να εφαρμοστεί σε κείμενα ενός άλλου

γνωστικού πεδίου με ανάλογα κλειστό λεξιλόγιο. Επίσης, ενώ η παρούσα εργασία δοκιμάστηκε σε κείμενα στην Ελληνική και την Αγγλική, οι θεμελιώδεις αρχές της είναι αρκετά γενικές ώστε να μπορεί να εφαρμοστεί και σε άλλα γλωσσικά ζεύγη.

6 Βιβλιογραφία

[1] Boutsis S and Piperidis S. Automatic Extraction of Bilingual Lexical Equivalences from Parallel Corpora, *Proceedings ECAI/MULSAIC*, pages 27-31, 1996

[2] Brown P. A Statistical Approach to Language Translation, *Proceedings of COLING-88*, volume 1, 71-76, 1988

[3] Brown P., Cocke J., Della Pietra S., Della Pietra V., Jelinek F., Lafferty J., Mercer R., Roosin P. A Statistical Approach to Machine Translation, *Computational Linguistics*, 16:2, 79-85, 1990

[4] Dagan I., Itai A., Schwall U. Two languages are more informative than one. *Proceedings of the 29th Annual Meeting of the ACL*, 130-137, 1991

[5] Gale W.A., Church K.W. Identifying word correspondences in parallel texts, *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, 152-157, 1991

[6] Gale W.A. and Church K.W., A Program for Aligning Sentences in Parallel Corpora, *Proceedings of the 29th Annual Meeting of the ACL.*, 1991. pp.177-184.

[7] Kay M. Text Translation Alignment. *Conference Handbook of ACH/ALLC'91*. Arizona, pp 257, 1991

[8] Kay M., Roescheisen M. Text-translation Alignment. *Computational Linguistics*, 19(1):121-142, 1993

[9] Kitamura M., Matsumoto Y. A Machine Translation System based on Translation Rules Acquired from Parallel Texts. *Recent Advances in Natural Language Processing*, 27-44, 1995

[10] Matsumoto Y., Ishimoto H., Utsuro T. Structural Matching of Parallel Texts" *Proceedings of the 31st Annual Meeting of the ACL*, 23-30, 1993

7 Στοιχεία Συγγραφέων

Σ. Πιπερίδης, Σ. Μπούτσης, Ι. Δεμοίρος
Ινστιτούτο Επεξεργασίας του Λόγου,
Μάργαρη 22, 115-25 Αθήνα
τηλ: 6712250, fax:6741262,
email: {spip, sboutsis, iason}@ilsp.gr