ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ & ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ - ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

"Τεχνολογίες και Διοίκηση Πληροφοριακών και Επικοινωνιακών Συστημάτων"

ΚΑΤΕΥΘΥΝΣΗ

«Διαχείριση Πληροφορίας»

ΕΡΓΑΣΤΗΡΙΟ ΤΕΧΝΟΛΟΓΙΑΣ ΓΝΩΣΕΩΝ & ΛΟΓΙΣΜΙΚΟΥ - ΙΝΣΤΙΤΟΥΤΟ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ - Ε.Κ.Ε.Φ.Ε. «ΔΗΜΟΚΡΙΤΟΣ»

ΤΙΤΛΟΣ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

«Αυτόματη Περίληψη Κειμένου (Automatic Text Summarization) - (Αυτόματη Περίληψη Κειμένου Από Πολλαπλά Έγγραφα Εξελισσόμενων Γεγονότων (Multi-document Summarization Of Evolving Events) - Αναγνώριση και κανονικοποίηση χρονικών εκφράσεων - Συμπλήρωση ορισμάτων μηνυμάτων)»

ΕΚΠΟΝΗΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ
Γεώργιος Κων/νου Σταματίου – icsdm03013

ΥΠΕΥΘΥΝΟΙ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ
Αναπληρωτής Καθηγητής - Δρ. Γεώργιος Βούρος
Β΄ Ερευνητής - Δρ. Ευάγγελος Καρκαλέτσης

Ιούνιος 2005

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή διδασκόντων την 28η Ιουνίου 2005:

Αναπληρωτής Καθηγητής Δρ. Γεώργιος Βούρος, Επιβλέπων
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων
(Υπογραφή)
.....

Λέκτορας Δρ. Κωνσταντίνος Στεργίου, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων
(Υπογραφή)
.....

Λέκτορας Δρ. Ευστάθιος Σταματάτος, Μέλος
Τμήμα Μηχανικών Πληροφοριακών και
Επικοινωνιακών Συστημάτων
(Υπογραφή)
.....

Αφιερώσεις

Αφιερώνεται στη μνήμη του πολυαγαπημένου μου ξάδερφου Φίλιππα Φόρτωμα και του πολυαγαπημένου μου παππού Γεώργιου Σταματίου.

Ευχαριστίες

Θα' θελα να ευχαριστήσω πρώτους απ' όλους τους γονείς μου για την αμέριστη οικονομική και ψυχολογική στήριξή τους όλα αυτά τα χρόνια των σπουδών μου. Δευτερευόντως, θα' θελα να ευχαριστήσω:

Τον αναπληρωτή καθηγητή και πρόεδρο του τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου Δρ. Γεώργιο Βούρο που μου έδωσε την ευκαιρία και με παρότρυνε να ασχοληθώ με τη συγκεκριμένη ερευνητική περιοχή.

Τον Ερευνητή Β΄ και επικεφαλή του εργαστηρίου Τεχνολογίας Γνώσεων & Λογισμικού του Ινστιτούτου Πληροφορικής & Τηλεπικοινωνιών του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» Δρ. Ευάγγελο Καρκαλέτση που με καθοδήγησε και μου προσέφερε πολύτιμη αρωγή στη συγκεκριμένη ερευνητική περιοχή.

Τον υποψήφιο διδάκτορα του εργαστηρίου Τεχνολογίας Γνώσεων & Λογισμικού του Ινστιτούτου Πληροφορικής & Τηλεπικοινωνιών του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» κ. Στέργο Αφαντενό για τη συνεργασία που είχαμε και την αμέριστη βοήθειά του στη συγκεκριμένη ερευνητική περιοχή.

Τον Ερευνητή Γ΄ του εργαστηρίου Τεχνολογίας Γνώσεων & Λογισμικού του Ινστιτούτου Πληροφορικής & Τηλεπικοινωνιών του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» Δρ. Γεώργιο Παλιούρα για το άριστο και φιλικό περιβάλλον εργασίας που μου προσέφερε.

Τον κ. Γεώργιο Πετάση για τις τεχνικές και προγραμματιστικές κατευθύνσεις που μου προσέφερε πάνω στην πλατφόρμα επεξεργασίας φυσικής γλώσσας Ellogon.

Το Ινστιτούτο Πληροφορικής & Τηλεπικοινωνιών του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» που μου έδωσε την ευκαιρία να γνωρίσω το περιβάλλον του και να εργαστώ στο χώρο του.

Τους θείους μου Τάκη και Εβελίνα για τη συμπαράστασή τους κατά τη διάρκεια της εκπόνησης της προκείμενης μεταπτυχιακής εργασίας.

Τον φίλο μου κ. Saied Mohammed Soliman για τη συμπαράστασή του και τις συμβουλές του κατά την περίοδο της εκπόνησης της προκείμενης μεταπτυχιακής εργασίας.

Όλους εσάς που με ιδιαίτερο ζήλο μελετάτε τη μεταπτυχιακή αυτή εργασία.

Σύντομο Βιογραφικό Σημείωμα

Ο Γεώργιος Κων/νου Σταματίου αποφοίτησε από το τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου το 2003 και κατέχει το δίπλωμα του Μηχανικού. Επί του παρόντος είναι μεταπτυχιακός φοιτητής του Προγράμματος Μεταπτυχιακών Σπουδών «Τεχνολογίες και Διοίκηση Πληροφοριακών και Επικοινωνιακών Συστημάτων» του τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου με ειδίκευση στην κατεύθυνση «Διαχείριση Πληροφορίας». Τα ερευνητικά του ενδιαφέροντα προέρχονται κυρίως από το χώρο της Τεχνητής Νοημοσύνης και εστιάζονται στην Επεξεργασία Φυσικής Γλώσσας, στις Οντολογίες και στον Σημασιολογικό Ιστό, στη Μηχανική Μάθηση και στους Ευφυείς Πράκτορες. Κατά τη διάρκεια των μεταπτυχιακών σπουδών του στο τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου προσέφερε επικουρικό έργο στα εργαστήρια των μαθημάτων «Πληροφοριακά και Επικοινωνιακά Συστήματα» και «Τεχνολογία Λογισμικού». Έχει εργαστεί κατά την περίοδο της πρακτικής του άσκησης στην εταιρία “Semantix Information Technologies SA” συμμετέχοντας στην υλοποίηση του project «Data Clearing House Setup Application» ως προγραμματιστής ιστοχώρου (Web Programmer) και κατά την περίοδο των Ολυμπιακών Αγώνων 2004 ως Τεχνικός Σταθμού Εργασίας (Work Station Technician) στα πλαίσια του προγράμματος εθελοντισμού «Αθήνα 2004». Τέλος είναι μέλος του IEEE (Institute of Electrical & Electronics Engineers) από το 2003 καθώς και δόκιμο μέλος της Ελληνικής Εταιρίας Τεχνητής Νοημοσύνης (ΕΕΤΝ) από το 2005.

Περίληψη

Στη σημερινή εποχή, η οποία χαρακτηρίζεται από υπερπληροφόρηση με αποτέλεσμα αυτός ο καταιγισμός των πληροφοριών να περιορίζει και να εξαντλεί τον ελεύθερο χρόνο του χρήστη αλλά και να τον δυσκολεύει στην αφομοίωση τους, η χρήση συστημάτων αυτόματης περίληψης κρίνεται πολλαπλά χρήσιμη και απολύτως αναγκαία. Η αυτόματη περίληψη κειμένου (Automatic Text Summarization) αποτελεί μια από τις βασικές εφαρμογές της επεξεργασίας φυσικής γλώσσας και ουσιαστικά αφορά στην εξαγωγή από ένα μεγάλο πηγαίο κείμενο ενός ίσου ή μικρότερου, με το κεντρικό νόημα του πρώτου. Μια υποπεριοχή αυτής είναι η αυτόματη περίληψη από πολλαπλά έγγραφα (Multi-Document Summarization) και πιο συγκεκριμένα η αυτόματη περίληψη από πολλαπλά έγγραφα εξελισσόμενων γεγονότων (Multi-Document Summarization Of Evolving Events) [5]. Ευελπιστούμε με τη δημιουργία κάποιων αρθρωμάτων και συγκεκριμένα με την επισημείωση των χρονικών εκφράσεων των κειμένων μιας συλλογής και με την κανονικοποίησή τους, καθώς και με μια πρώτη προσέγγιση συμπλήρωσης ορισμάτων μηνυμάτων, τα οποία μηνύματα αναπαριστούν τα περιστατικά (incidents) ενός γεγονότος να προσφέρουμε το λιθαράκι μας στη συγκεκριμένη ερευνητική υποπεριοχή.

Abstract

In nowadays, which is characterized by a lot of information with the result that it limits and exhausts the free time of the user as well as making difficult his ability to digest it, the use of automatic summarization systems seems multiple useful and indispensable. The Automatic Text Summarization constitutes one of the basic applications of natural language processing and substantially concerns the export of a similar or shorter text from a long source text with the central meaning of the first one. A subarea of this area is Multi-Document Summarization and much more specifically Multi-Document Summarization of Evolving Events [5]. We hope to contribute in this particular research subarea with the creation of some modules and specifically with the annotation of time expressions of documents of a corpus and with their normalization, as well as with a first approach of filling the arguments of messages, which messages represents the incidents of an event.

Λεξικό Όρων

Ακροατήριο (Audience)	Παράμετρος ενός συστήματος περίληψης (Summarizer) που σχετίζεται με μια γενική ή εστιασμένη στο χρήστη περίληψη
Ανάκτηση Εγγράφου (Document Retrieval)	Το έργο της λήψης μιας συλλογής εγγράφων και μιας ανάγκης ενός χρήστη και της ανάκτησης εγγράφων σε σχέση με την ανάγκη του χρήστη
Ανάλυση (Analysis)	Φάση της αυτόματης περίληψης, η οποία αναλύει την είσοδο χτίζοντας μια εσωτερική αναπαράσταση αυτής
Ανθρώπινη περίληψη βοηθημένη από μηχάνημα (Machine Assisted Human Summarization – MAHS)	Η περίληψη που εξάγεται από έναν άνθρωπο με το μηχάνημα να επικουρεί
Απάντηση Ερώτησης (Question Answering)	Η λήψη μιας ερώτησης σε φυσική γλώσσα και η κατοχή ενός συστήματος που παρέχει μια απάντηση από μια συλλογή εγγράφων ή μια δομημένη βάση δεδομένων
Αποσπασματική Περίληψη (Fragmentary Summary)	Μια περίληψη η οποία αποτελείται μόνο από μια λίστα λέξεων ή φράσεων
Ασύγχρονη εκπομπή (Asynchronous Emission)	Η μη ταυτόχρονη εκπομπή ειδησεογραφικών αναφορών από τις πηγές
Αυτόματη Περίληψη Από Πολλαπλά Έγγραφα Εξελισσόμενων Γεγονότων (Multi-Document Summarization Of Evolving Events)	Αυτόματη περίληψη από πολλαπλά έγγραφα που περιλαμβάνουν γεγονότα που εξελίσσονται με το χρόνο
Αυτόματη Περίληψη Εγγράφου (Automatic Document Summarization)	Περίληψη εγγράφων που θεωρούνται ως πηγές πληροφοριών, των οποίων το περιεχόμενο αντανακλά πράγματα του κόσμου
Βαθμός συμπίεσης (Compression Rate)	Το μήκος της περίληψης προς το μήκος της πηγής
Βαθμός Συμπύκνωσης (Condensation Rate)	Ό,τι και το Compression Rate
Βαθύτερη Προσέγγιση (Deeper Approach)	Περίληψη η οποία απαιτεί τουλάχιστον ανάλυση στο σημασιολογικό επίπεδο του γλωσσολογικού χώρου και η οποία συνήθως εκτελεί τη σύνθεση περιλαμβάνοντας την παραγωγή γλώσσας από ένα σημασιολογικό ή πραγματολογικό επίπεδο αναπαράστασης
Γενίκευση (Generalization)	Αντικατάσταση στοιχείων κειμένου με πιο γενικά
Γενική Περίληψη (Generic Summary)	Περιλήψεις που στοχεύουν σε ένα συγκεκριμένο (συνήθως ευρύ) αναγνωστικό κοινό
Γραμμική εξέλιξη (Linear Evolution)	Η εξέλιξη όπου τα σημαντικότερα περιστατικά ενός γεγονότος συμβαίνουν σε σταθερά και ενδεχομένως προβλέψιμα κβάντα του χρόνου
Διαχρονική Σχέση (Diachronic Relation)	Η σχέση που προσπαθεί να προσδιορίσει τις ομοιότητες και τις διαφορές, κατά τη διάρκεια του χρόνου, οι οποίες υπάρχουν για ένα γεγονός δεδομένου ότι περιγράφεται από την ίδια πηγή
Έκταση Περίληψης (Summary Span)	Απλή ή πολλαπλών εγγράφων
Ενδεικτική Σύνοψη (Indicative Abstract)	Παρέχει μια συνάρτηση αναφοράς για την επιλογή εγγράφων για περισσότερη εμπεριστατωμένη μελέτη

Εντροπία (Entropy)	Το ποσό της πληροφορίας σε μια τυχαία μεταβλητή, ή το μέσο μήκος ενός μηνύματος που χρειάζεται για να μεταδοθεί ένα αποτέλεσμα αυτής της μεταβλητής. Μπορεί να χρησιμοποιηθεί ως ένα μέτρο της προβλεψιμότητας του πηγαίου κειμένου
Εξαγωγή (Extract)	Μια περίληψη αποτελούμενη ολοκληρωτικά από υλικό που προέρχεται από την είσοδο
Εξαγωγή Κειμένου (Text Mining)	Διαδικασία ανεύρεσης που στοχεύει στην ανίχνευση νέων ή ανωμάλων ή αλλιώς ενδιαφέρουσας πληροφορίας σε μεγάλες αποθήκες κειμένων
Εξαγωγή Πληροφορίας (Information Extraction)	Το έργο της συμπλήρωσης προτύπων ή πινάκων από μια είσοδο σε φυσική γλώσσα
Επίπεδο (Level)	Μορφολογικό (Morphological), Συντακτικό (Syntactic), Σημασιολογικό (Semantic), Πραγματολογικό (Discourse / pragmatic)
Επισημείωση (Annotation)	Μια γλωσσολογική πληροφορία που προστίθεται στο έγγραφο μιας συλλογής και που συνδέει αυθαίρετη πληροφορία (arbitrary information) (με τη μορφή ιδιοτήτων (attributes)) με τμήμα κειμένου που ονομάζεται έκταση (span). Μια επισημείωση αποτελείται από ένα αναγνωριστικό (identification – id), έναν τύπο (type), μία έκταση (span) (δηλαδή ένα ζεύγος από byte offsets) και ένα σύνολο ιδιοτήτων (attributes)
Ετερογλωσσική Περίληψη (Cross-Lingual Summarization)	Επεξεργασία διαφόρων γλωσσών με την περίληψη όμως σε διαφορετική γλώσσα από τη γλώσσα εισόδου
Ευρετηρίαση (Indexing)	Η αναγνώριση κατάλληλων όρων σε ένα έγγραφο, συνήθως για τη διευκόλυνση της ανάκτησης πληροφορίας.
Θεωρητική Πληροφοριακότητα Πληροφορίας (Information Theoretic Informativeness)	Ένα μέτρο του βαθμού με τον οποίο μια περίληψη επιτρέπει σε κάποιον να ανακατασκευάσει το πηγαίο έγγραφο
Κριτική Σύνοψη (Critical Abstract)	Αξιολογεί το θέμα μελέτης της πηγής, εκφράζοντας τις απόψεις του abstractor σχετικά με την ποιότητα της εργασίας του συγγραφέα
Λειτουργία (Function)	Ενδεικτική, πληροφοριακή ή κριτική
Λειτουργία Επιλογής (Selection Operation)	Φιλτράρισμα στοιχείων κειμένου
Λειτουργία Συμπύκνωσης (Condensation Operation)	Επιλογή, Συσσώρευση ή γενίκευση
Μετασχηματισμός (Transformation)	Ό,τι και το ραφινάρισμα (Refinement)
Μη-γραμμική Εξέλιξη (Non-linear Evolution)	Η εξέλιξη όπου τα σημαντικότερα περιστατικά ενός γεγονότος συμβαίνουν σε μη σταθερά κβάντα του χρόνου
Μήνυμα (Message)	Η αναπαράσταση του περιστατικού ενός γεγονότος που αποτελείται από τον τύπο του, ένα σύνολο ορισμάτων, τα οποία παίρνουν τιμές από την οντολογία του υπό εξέταση χώρου, την πηγή από την οποία προέρχεται, τη χρονική στιγμή της έκδοσής του, καθώς και τη χρονική στιγμή στην οποία αυτό πραγματικά αναφέρεται
Μηχανική περίληψη βοηθημένη από άνθρωπο (Human Assisted Machine Summarization – HAMS)	Η περίληψη που εξάγεται από ένα μηχάνημα με τον άνθρωπο να την επεξεργάζεται και να τη διορθώνει.
Μονογλωσσική Περίληψη (Monolingual Summarization)	Η επεξεργασία μόνο μιας γλώσσας με την περίληψη να είναι στην ίδια γλώσσα με την είσοδο
Μορφολογικό Επίπεδο	Ένα συγκεκριμένο επίπεδο του γλωσσολογικού χώρου που

(Morphological Level)	βασίζεται στον προσδιορισμό μορφημάτων
Μορφοποίηση Εξόδου (Output Format)	Μορφοποίηση και διαρρύθμιση (layout) της περίληψης
Περίληψη που εστιάζει σε ερώτηση (Query-Focused Summary)	Περίληψη που εστιάζει στην ερώτηση που τίθεται από κάποιον χρήστη
Περίληψη που εστιάζει σε θέμα (Topic-Focused Summary)	Μια περίληψη που περιέχει πληροφορίες σχετικές με το θέμα
Περίληψη που εστιάζει στον χρήστη (User-Focused Summary)	Μια περίληψη που προσαρμόζεται στις απαιτήσεις ενός συγκεκριμένου χρήστη ή ομάδας χρηστών
Περίληψη Υπογλώσσας (Sublanguage Summary)	Περίληψη σε μια γλώσσα με ένα συγκεκριμένο περιορισμένο λεξιλόγιο και συντακτικό
Πλεονασμός (Redundancy)	Στοιχεία κειμένου τα οποία επαναλαμβάνουν την ίδια ιδέα
Πληροφοριακή Σύνοψη (Informative Abstract)	Καλύπτει όλες τις σημαντικές πληροφορίες της πηγής σε κάποιο επίπεδο λεπτομέρειας
Πλήρως αυτόματη περίληψη (Fully Automated Summarization – FAS)	Η περίληψη που δημιουργείται αποκλειστικά από ένα μηχάνημα
Πολυγλωσσική Περίληψη (Multilingual Summarization)	Η επεξεργασία διαφόρων γλωσσών, με την περίληψη να είναι στην ίδια γλώσσα με την είσοδο
Πολυμεσική περίληψη (Multimedia Summarization)	Περίληψη, όπου η είσοδος και / η έξοδος αποτελούνται από έναν συνδυασμό από διαφορετικούς τύπους μέσων, όπως κείμενο, ήχο, πίνακες, εικόνες και διαγράμματα, ταινίες
Πραγματολογικό Επίπεδο (Discourse / Pragmatic Level)	Ένα συγκεκριμένο επίπεδο του γλωσσολογικού χώρου που βασίζεται σε πραγματική γνώση
Ραφινάρισμα (Refinement)	Μετατρέπει την εσωτερική αναπαράσταση μιας πηγής που κατασκευάζεται από την ανάλυση σε μια αναπαράσταση της περίληψης
Ρηχότερη Προσέγγιση (Shallower Approach)	Περίληψη, η οποία απαιτεί ανάλυση μόνο σ' ένα ρηχό επίπεδο του γλωσσολογικού χώρου και όπου η σύνθεση εκτελείται μόνο από μια αναπαράσταση συντακτικού επιπέδου
Σημαντικότητα (Salience)	Η σημαντικότητα π.χ. μιας πρότασης (σχετίζεται με τη συνάφεια (Relevance))
Σημασιολογική Πληροφοριακότητα (Semantic Informativeness)	Ένα μέτρο του περιεχομένου πληροφορίας της περίληψης
Σημασιολογικό Επίπεδο (Semantic Level)	Ένα συγκεκριμένο επίπεδο του γλωσσολογικού χώρου που βασίζεται σε σημασιολογική γνώση
Σκοπός της περίληψης (Goal of Summarization)	Η λήψη μιας πηγής πληροφορίας, η εξαγωγή περιεχομένου απ' αυτή και η παρουσίαση του πιο σημαντικού περιεχομένου στον χρήστη σε μια συνοπτική μορφή και μ΄ έναν τρόπο συνυφασμένο με τις ανάγκες του χρήστη ή της εφαρμογής
Στοιχείο (Element)	Ένα στοιχείο κειμένου όπως μια λέξη, μια φράση, μια πρόταση, μια παράγραφος ή ένα έγγραφο
Σύγχρονη Εκπομπή (Synchronous Emission)	Η ταυτόχρονη εκπομπή ειδησεογραφικών αναφορών μεταξύ πηγών
Συγχρονική Σχέση (Synchronous Relation)	Η σχέση που προσπαθεί να προσδιορίσει τις ομοιότητες και τις διαφορές που δύο πηγές έχουν, στον ίδιο χρόνο
Συγχώνευση (Aggregation)	Λειτουργία της αυτόματης περίληψης που περιλαμβάνει συγχώνευση των στοιχείων του κειμένου
Συμπίεση Κειμένου (Text	Περίληψη που περιλαμβάνει κείμενο που έχει εξαχθεί από

Compaction)	ένα έγγραφο πηγής και το οποίο έπειτα έχει περικοπεί ή συντομευτεί
Συμπίεση Κειμένου (Text Compression)	Η σύμπτυξη της εισόδου ενός κειμένου με τη μεταχείριση της εισόδου ως κώδικα που προορίζεται για αποτελεσματική αποθήκευση και μετάδοση ανάμεσα στα μηχανήματα
Συνάφεια (Relevance)	Βάρος που προσαρτάται σε πληροφορία σε ένα έγγραφο, αντανακλώντας τόσο το περιεχόμενο του εγγράφου όσο και τη σχέση της πληροφορίας του εγγράφου με την εφαρμογή
Συνέπεια (Coherence)	Ο τρόπος με τον οποίο μέρη του κειμένου συγκεντρώνονται μαζί για να σχηματίσουν ένα ολοκληρωμένο σύνολο
Σύνθεση (Synthesis)	Αποδίδει την αναπαράσταση της περίληψης πίσω σε φυσική γλώσσα
Σύνοψη (Abstract)	Μια περίληψη της οποίας τουλάχιστον μερικό από το περιεχόμενο δεν είναι παρόν στην είσοδο
Συντακτικό Επίπεδο (Syntactic Level)	Ένα συγκεκριμένο επίπεδο του γλωσσολογικού χώρου που βασίζεται σε συντακτική γνώση
Σύστημα Περίληψης (Summarizer)	Ένα σύστημα του οποίου σκοπός είναι να παράγει μια συνοπτική αναπαράσταση του περιεχομένου της εισόδου για χρησιμοποίηση από άνθρωπο
Σχέση με την πηγή (Relation to Source)	Εξαγωγή έναντι Σύνοψης
Τίτλος (Title)	Ο τίτλος ενός εγγράφου που χρησιμοποιείται ως μια ένδειξη σημαντικού πληροφοριακού περιεχομένου. Η παρουσία των τίτλων σε ένα έγγραφο χρησιμοποιείται για την ανίχνευση σημαντικών κομματιών κειμένου στο έγγραφο
Ύφος (Genre)	Διάφορες ποικιλίες κειμένου, συνήθως επιστημονικές ή τεχνικές αναφορές, νέες ιστορίες, μηνύματα ηλεκτρονικού ταχυδρομείου, κύρια άρθρα, βιβλία
Φάση Περίληψης (Summarization Phase)	Ανάλυση, ραφινάρισμα, σύνθεση

Πίνακας Περιεχομένων

1 ΕΙΣΑΓΩΓΙΚΑ 17

1.1 ΑΝΤΙΚΕΙΜΕΝΟ ΤΗΣ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ 17
1.2 ΣΤΟΧΟΙ ΤΗΣ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ 17
1.3 ΔΟΜΗ ΤΗΣ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ 18

2 ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ 19

2.1 ΕΙΣΑΓΩΓΙΚΑ 19
2.2 ΑΥΤΟΜΑΤΗ ΠΕΡΙΛΗΨΗ ΠΟΛΛΑΠΛΩΝ ΕΓΓΡΑΦΩΝ ΕΞΕΛΙΣΣΟΜΕΝΩΝ ΓΕΓΟΝΟΤΩΝ (MULTI – DOCUMENT SUMMARIZATION OF EVOLVING EVENTS) 31

3 ΧΡΟΝΙΚΕΣ ΕΚΦΡΑΣΕΙΣ 39

3.1 ΕΙΣΑΓΩΓΗ 39
3.2 ΔΗΜΙΟΥΡΓΙΑ ΣΩΜΑΤΟΣ ΚΕΙΜΕΝΩΝ ΑΞΙΟΛΟΓΗΣΗΣ 39
3.3 ΣΥΓΓΡΑΦΗ ΚΑΝΟΝΩΝ 41
3.4 ΕΦΑΡΜΟΓΗ ΚΑΝΟΝΩΝ 42
3.5 ΠΕΙΡΑΜΑΤΑ – ΑΠΟΤΕΛΕΣΜΑΤΑ 44

4 ΜΗΝΥΜΑΤΑ 49

4.1 ΕΙΣΑΓΩΓΗ 49
4.2 ΔΗΜΙΟΥΡΓΙΑ ΣΩΜΑΤΟΣ ΚΕΙΜΕΝΩΝ ΑΞΙΟΛΟΓΗΣΗΣ 49
4.3 ΜΕΘΟΔΟΛΟΓΙΑ ΣΥΜΠΛΗΡΩΣΗΣ ΤΩΝ ΟΡΙΣΜΑΤΩΝ ΤΩΝ ΜΗΝΥΜΑΤΩΝ 50
4.4 ΠΕΙΡΑΜΑΤΑ – ΑΠΟΤΕΛΕΣΜΑΤΑ 51

5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ 53

6 ΒΙΒΛΙΟΓΡΑΦΙΑ – ΑΝΑΦΟΡΕΣ 55

7 ΠΑΡΑΡΤΗΜΑ 56

7.1 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΣΥΛΛΟΓΗΣ ΚΕΙΜΕΝΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΕ 56
7.2 ΧΡΗΣΗ ΤΟΥ ELLOGON 57
7.3 ΧΡΟΝΙΚΕΣ ΕΚΦΡΑΣΕΙΣ 70
7.4 ΜΗΝΥΜΑΤΑ 77

Κατάλογος Σχημάτων

Σχήμα 1 - Σχέσεις μεταξύ ενδεικτικών, πληροφοριακών και κριτικών περιλήψεων ..23
Σχήμα 2 – Μια υψηλού επιπέδου αρχιτεκτονική ενός συστήματος περίληψης 27
Σχήμα 3 – Ο γλωσσολογικός χώρος 28
Σχήμα 4 – Αρχιτεκτονική της ρηχής προσέγγισης 29
Σχήμα 5 - Γραμμικές – Μη γραμμικές Εξελίξεις, Σύγχρονη και Ασύγχρονη Εκπομπή 33
Σχήμα 6 – Παραδείγματα σύγχρονων και ασύγχρονων σχέσεων.....35
Σχήμα 7 - Σύστημα περίληψης από πολλαπλά έγγραφα εξελισσόμενων γεγονότων που βασίζεται σε ερώτηση 36
Σχήμα 8 - Υποσύστημα εξαγωγής μηνυμάτων.....37

Κατάλογος Διαγραμμάτων

Διάγραμμα 1 – Total Correct Annotations, Total Key Annotations και Total Response Annotations 45
Διάγραμμα 2 – Total Correct Annotations σε σχέση με τα Total Key Annotations 45
Διάγραμμα 3 – Total Correct Annotations σε σχέση με τα Total Response Annotations 46
Διάγραμμα 4 – Precision, Recall και F-Measure 46

Κατάλογος Πινάκων

Πίνακας 1- Παράμετροι ενός συστήματος περίληψης.....26
Πίνακας 2 – Παραδείγματα Προδιαγραφών Μηνυμάτων 34
Πίνακας 3 – Αποτελέσματα κανονικοποίησης χρονικών εκφράσεων.....45
Πίνακας 4 – Αποτελέσματα αναγνώρισης χρονικών εκφράσεων 47
Πίνακας 5 – Αποτελέσματα αναγνώρισης χρονικών εκφράσεων του συστήματος του project Mitos 48
Πίνακας 6 - Αποτελέσματα σύγκρισης της τιμής των ιδιοτήτων entity1, entity2, from_place & quantity, entity1 & entity2 από κοινού & entity1, entity2, from_place & quantity από κοινού για μηνύματα τύπου free 52

Κατάλογος Εικόνων

Εικόνα 1 - Δημιουργία Συλλογής 57
Εικόνα 2 - Προσθήκη εγγράφων προς επισημείωση 58
Εικόνα 3 - Ενεργοποίηση της Java 59
Εικόνα 4 - Δημιουργία Συστήματος 59
Εικόνα 5 - Αποθήκευση νέου συστήματος 60
Εικόνα 6 – Επιτυχής προσθήκη νέου συστήματος 61
Εικόνα 7 – Το παράθυρο του συστήματος AegeanNcsrAegeanTempExp 61
Εικόνα 8 - Το άρθρωμα TemporalExpressions είναι έτοιμο προς εκτέλεση 62
Εικόνα 9 - Το άρθρωμα TemporalExpressions υπό εκτέλεση 62
Εικόνα 10 - Επιτυχής εκτέλεση του αρθρώματος TemporalExpressions 63
Εικόνα 11 - Το πρόγραμμα παρουσίασης "Exprore Annotations" 64
Εικόνα 12 – Οι χρονικές εκφράσεις 64
Εικόνα 13 – Εμφάνιση επισημειώσεων 65
Εικόνα 14 –Χειρωνακτική εισαγωγή επισημειώσεων 66
Εικόνα 15 – Χειρωνακτική επισημείωση χρονικών εκφράσεων με το πάτημα ενός κουμπιού 66
Εικόνα 16 – Σύγκριση τιμής ιδιότητας χειρωνακτικών επισημειώσεων με τιμή ιδιότητας επισημειώσεων που παράγονται αυτόματα από το σύστημα για το άρθρωμα των χρονικών εκφράσεων 67
Εικόνα 17 – Αποτελέσματα Σύγκρισης 68
Εικόνα 18 – Παράμετροι της σύγκρισης και αποτελέσματα της αξιολόγησης 68
Εικόνα 19 - Σύγκριση τιμής ιδιότητας χειρωνακτικών επισημειώσεων με τιμή ιδιότητας επισημειώσεων που παράγονται αυτόματα από το σύστημα για το άρθρωμα των μηνυμάτων 70

1 Εισαγωγικά

Στη σημερινή εποχή, όπου χαρακτηρίζεται από υπερπληροφόρηση με αποτέλεσμα αυτός ο καταιγισμός των πληροφοριών να περιορίζει και να εξαντλεί τον ελεύθερο χρόνο του χρήστη, η χρήση συστημάτων αυτόματης περίληψης κρίνεται πολλαπλά χρήσιμη. Ας φανταστούμε για παράδειγμα κάποιον ο οποίος θέλει να παρακολουθεί κάποιο γεγονός που περιγράφεται σε διάφορες πηγές ειδήσεων στο διαδίκτυο, όπως εξελίσσεται με το χρόνο. Το πρόβλημα που προκύπτει είναι ότι υπάρχει ένας μεγάλος αριθμός πηγών ειδήσεων που καθιστούν πολύ δύσκολο για κάποιον να συγκρίνει τις διαφορετικές εκδόσεις της είδησης σε κάθε πηγή. Η αυτόματη περίληψη κειμένου (Automatic Text Summarization) που αποτελεί μια από τις βασικές εφαρμογές της επεξεργασίας φυσικής γλώσσας (Natural Language Processing) και ουσιαστικά αφορά στην εξαγωγή από ένα μεγάλο κείμενο ενός μικρότερου, με το κεντρικό νόημα του πρώτου, είναι μια λύση στο πρόβλημα αυτό.

1.1 Αντικείμενο της μεταπτυχιακής εργασίας

Η συγκεκριμένη μεταπτυχιακή εργασία εστιάζει σε μια υποπεριοχή της Αυτόματης Περίληψης Κειμένου, σ' αυτή της Αυτόματης Περίληψης από πολλαπλά έγγραφα (Multi–Document Summarization - MDS) και ειδικότερα σ' αυτή όπου τα έγγραφα αναφέρονται σε ένα γεγονός το οποίο εξελίσσεται σε μια χρονική περίοδο (Multi–Document Summarization Of Evolving Events). Για παράδειγμα, οι ειδήσεις σχετικά με το γεγονός της ομηρίας των δύο Ιταλίδων από ομάδα Ιρακινών.

Στη συγκεκριμένη υποπεριοχή της αυτόματης παραγωγής περιλήψεων, έχει προταθεί μια μεθοδολογία από τον κ. Αφαντενό [4] [5] στα πλαίσια του διδακτορικού του. Σύμφωνα με τη μεθοδολογία αυτή, η παραγωγή περίληψης από πολλαπλά έγγραφα που περιγράφουν το ίδιο γεγονός έτσι όπως αυτό εξελίσσεται στο χρόνο, προϋποθέτει τα ακόλουθα:

- Αναπαράσταση του κάθε εγγράφου ως ένα σύνολο **μηνυμάτων**. Τα μηνύματα περιγράφουν **περιστατικά (incidents)** του γεγονότος και κάθε μήνυμα αντιστοιχεί συνήθως σε μια πρόταση του εγγράφου. Ένα τέτοιο μήνυμα θα μπορούσε για παράδειγμα να αφορά στην απελευθέρωση ενός ομήρου. Θα πρέπει να σημειωθεί ότι ένα μήνυμα χαρακτηρίζεται από τον τύπο του, τα ορίσματά του, την πηγή απ' όπου προέρχεται καθώς και από τον χρόνο του περιστατικού που περιγράφει, ο οποίος υπολογίζεται συνήθως σε σχέση με την ημερομηνία της είδησης, μέσω μιας χρονικής έκφρασης.
- Συσχέτιση των μηνυμάτων διαφορετικών εγγράφων σύμφωνα με κάποιες προκαθορισμένες σχέσεις. Κάθε σχέση καθορίζει τους τύπους μηνυμάτων στους οποίους μπορεί να εφαρμοστεί και τις συνθήκες που πρέπει να ικανοποιούνται έτσι ώστε να εφαρμοστεί

Η αναπαράσταση των εγγράφων με τη χρήση μηνυμάτων και η συσχέτιση των μηνυμάτων επιτρέπει σε ένα επόμενο στάδιο και κατόπιν υποβολής ερώτησης από ένα χρήστη, την επιλογή των σχέσεων που απαντούν στη συγκεκριμένη ερώτηση και την παραγωγή της περίληψης από τις σχέσεις αυτές με χρήση τεχνικών παραγωγής φυσικής γλώσσας,

1.2 Στόχοι της μεταπτυχιακής εργασίας

Στη συγκεκριμένη μεθοδολογία αυτόματης παραγωγής περιλήψεων έχει ιδιαίτερη σημασία ο προσδιορισμός του χρόνου του μηνύματος, καθώς και η συμπλήρωση των ορισμάτων ενός μηνύματος. Η μελέτη των δύο αυτών προβλημάτων αποτέλεσε και το στόχο της συγκεκριμένης εργασίας:

- Πρώτος στόχος ήταν η αυτοματοποιημένη επισημείωση (annotation) των χρονικών εκφράσεων (temporal expressions) που εμφανίζονται στα κείμενα μιας συλλογής καθώς και η κανονικοποίησή (normalization) τους. Συγκεκριμένα, θα πρέπει να αναγνωρίζονται οι χρονικές εκφράσεις, να σημειώνεται η έκτασή τους και να ανάγονται στον πραγματικό χρόνο τον οποίο αντιπροσωπεύουν σε σχέση με την ημερομηνία δημοσίευσης της είδησης
- Δεύτερος στόχος ήταν η συμπλήρωση των ορισμάτων κάποιων ενδεικτικών τύπων μηνυμάτων

Η μεταπτυχιακή εργασίας είχε επίσης ως στόχο την υλοποίηση εργαλείων που μπορούν να αξιοποιηθούν από το σύστημα αυτόματης παραγωγής περιλήψεων που υλοποιείται στο πλαίσιο του διδακτορικού του κ. Αφαντενού [4] [5] [6].

1.3 Δομή της μεταπτυχιακής εργασίας

Το 2ο κεφάλαιο «Θεωρητικό Υπόβαθρο», παρουσιάζει βασικές έννοιες από την περιοχή της Αυτόματης Περίληψης Κειμένου, και την υποπεριοχή της Αυτόματης Περίληψης από πολλαπλά έγγραφα. Στην τελευταία ενότητα του Κεφαλαίου παρουσιάζεται αναλυτικά η μεθοδολογία που προτείνεται στο διδακτορικό του κ. Αφαντενού [4] [5], και το σύστημα που υλοποιεί τη μεθοδολογία αυτή και στο οποίο ενσωματώνονται τα εργαλεία που υλοποιήθηκαν στην προκείμενη μεταπτυχιακή εργασία.

Το 3ο κεφάλαιο «Χρονικές Εκφράσεις», περιγράφει την προσέγγιση και τη μεθοδολογία που ακολουθήθηκε για την επισημείωση των χρονικών εκφράσεων. Δίνονται, πληροφορίες σχετικά με τη δημιουργία του σώματος κειμένων αξιολόγησης, τις κατηγορίες στις οποίες οργανώθηκαν οι χρονικές εκφράσεις τη συγγραφή κανόνων και τρόπων κανονικοποίησης για κάθε μία από τις κατηγορίες αυτές, τη διαδικασία εφαρμογής των κανόνων, και τέλος τα πειράματα που πραγματοποιήθηκαν και τα αποτελέσματα που λήφθηκαν.

Στο 4ο κεφάλαιο «Μηνύματα» περιγράφονται η προσέγγιση και η μεθοδολογία που ακολουθήθηκε για τη συμπλήρωση των ορισμάτων των μηνυμάτων, πληροφορίες σχετικά με τη δημιουργία του σώματος αξιολόγησης, ο τρόπος υλοποίησης και τέλος τα πειράματα που πραγματοποιήθηκαν και τα αποτελέσματα που λήφθηκαν.

Η εργασία περιλαμβάνει επίσης το 5ο κεφάλαιο «Συμπεράσματα και προτάσεις για μελλοντικές κατευθύνσεις» και το Παράρτημα στο οποίο περιλαμβάνονται ενότητες που μπορεί να βρει κάποιος πληροφορίες για τη συλλογή κειμένων (corpus) που χρησιμοποιήθηκε, τη χρήση της πλατφόρμας επεξεργασίας φυσικής γλώσσας Ellogon καθώς και τρόπους και παραδείγματα κανονικοποίησης των χρονικών εκφράσεων κατά κατηγορία, την οντολογία του πεδίου (domain ontology) από την οποία λαμβάνονται οι τιμές για τη συμπλήρωση των ορισμάτων των μηνυμάτων, τις προδιαγραφές των μηνυμάτων και τους τρόπους υλοποίησης σε πολύ αφαιρετικό επίπεδο των αρθρωμάτων των χρονικών εκφράσεων και των μηνυμάτων.

Τονίζουμε για αποφυγή παρερμηνειών ότι οι λέξεις **κείμενο**, **έγγραφο, άρθρο και είδηση** έχουν την ίδια σημασία στην προκείμενη μεταπτυχιακή εργασία και χρησιμοποιούνται εναλλακτικά. Επιπλέον, όπου γίνεται αναφορά στον όρο **περίληψη** πρόκειται για **αυτόματη περίληψη**, εκτός και αν διευκρινίζεται σαφώς ή υπονοείται από τα συμφραζόμενα πως πρόκειται για **χειρωνακτική περίληψη**. Όσον αφορά στις επισημειώσεις, στο όνομα μιας επισημείωσης μπορούμε να αναφερθούμε ως **τύπος μιας επισημείωσης** ή πολύ απλά ως **επισημείωση**. Το ίδιο ισχύει και για τις ιδιότητες μιας επισημείωσης, δηλαδή στο όνομα μιας ιδιότητας μπορούμε να αναφερθούμε ως **τύπος μιας ιδιότητας** ή πολύ απλά ως **ιδιότητα**.

2 Θεωρητικό Υπόβαθρο

2.1 Εισαγωγικά

Με την ταχεία ανάπτυξη του ιστοχώρου και των ηλεκτρονικών υπηρεσιών πληροφοριών, οι πληροφορίες διατίθενται on-line με έναν απίστευτο ρυθμό. Αποτέλεσμα αυτού είναι η συχνά επικριθείσα υπερπληροφόρηση. Κανένας δεν έχει το χρόνο να τα διαβάσει όλα, εν' τούτοις συχνά οι άνθρωποι παίρνουν κρίσιμες αποφάσεις που βασίζονται σ' αυτό πού είναι σε θέση να αφομοιώσουν.

Συγκεκριμένα, έρευνα που πραγματοποίησε βρετανική εταιρεία σε διευθυντικά στελέχη μεγάλων επιχειρήσεων στη Δυτική Ευρώπη, κατέληξε στο συμπέρασμα ότι οι περισσότεροι απ' αυτούς υποφέρουν από μια νέα ασθένεια: το σύνδρομο της υπερπληροφόρησης. Πρόκειται για το άγχος, την ένταση και την αβεβαιότητα που δημιουργεί ο καθημερινός κατακλυσμός από πληροφορίες και ειδήσεις που καταφθάνουν, μ' όλα αυτά τα υπερσύγχρονα μέσα επικοινωνιών, στα γραφεία των επιχειρήσεων. Η επιτακτική ανάγκη για συνεχή ενημέρωση και αφομοίωση κάθε τι καινούργιου και η δυνατότητα πολλών επιλογών, άρα και δύσκολων αποφάσεων, έχουν σαν αποτέλεσμα υπερβολικό άγχος, εξασθενημένη υγεία, έλλειψη αυτοπεποίθησης, ανόητες αποφάσεις και λανθασμένα συμπεράσματα.

Η τεχνολογία της **αυτόματης περίληψης κειμένου (automatic text summarization)** γίνεται απολύτως αναγκαία για την επίλυση αυτού του προβλήματος. Η αυτόματη περίληψη κειμένου είναι σε γενικές γραμμές η διαδικασία της εξαγωγής των σημαντικότερων πληροφοριών από μια πηγή για να παραχθεί μια συντομευμένη έκδοση. Σημειώνουμε ότι θα εστιάσουμε σε **αυτόματη περίληψη εγγράφων (automatic document summarization)**, δηλαδή σε περίληψη εγγράφων που θεωρούνται ως πηγές πληροφοριών των οποίων το περιεχόμενο αντικατοπτρίζει πράγματα του κόσμου. Η περίληψη κατά αυτή τη στενή θεώρηση αποσκοπεί στην εξαγωγή της πεμπτουσίας της πληροφορίας στα έγγραφα.

Η αυτόματη περίληψη κειμένου είναι μια πολύ διεπιστημονική ερευνητική περιοχή που περιλαμβάνει τομείς όπως η επεξεργασία φυσικής γλώσσας, η ανάκτηση πληροφορίας, η επιστήμη βιβλιοθηκών (library science), η στατιστική, η γνωστική ψυχολογία και η τεχνητή νοημοσύνη.

Στόχος της αυτόματης περίληψης είναι να πάρει ως είσοδο μια πηγή πληροφορίας, να εξάγει περιεχόμενο απ' αυτή και να παρουσιάσει το περισσότερο σημαντικό περιεχόμενο στον χρήστη σε μια συνοπτική μορφή που να είναι συνυφασμένη με τις ανάγκες του ή με τις ανάγκες της εφαρμογής.

Ποιά είναι όμως η ερμηνεία της περίληψης στην καθημερινή πραγματικότητα; Μια από τις ερμηνείες που δίνεται είναι **«η σύντομη απόδοση του περιεχομένου γραπτού ή προφορικού κειμένου»**. «Εν' περιλήψει» σημαίνει στην καθομιλουμένη «με λίγα λόγια», «με συντομία». Λαϊκότερα η περίληψη μπορεί να συναντηθεί ως «ρεζουμέ», ενώ λογιότερα ως «σύνοψη». Η ερμηνεία αυτή, όπως διαπιστώνεται εύκολα δεν απέχει και πολύ από αυτή της αυτόματης περίληψης.

Την αναγκαιότητα για την περιληπτική, περιεκτική και συνοπτική παρουσίαση της πληροφορίας την παρατηρούμε και στα αρχαία χρόνια. Στα κείμενα των αρχαίων συγγραφέων θαυμάζουμε σαφήνεια, πυκνότητα, επιγραμματικότητα και φυσικά ουσία. Είναι οι αρετές του λόγου που περιέχονται στον όρο κλασικό ύφος. Οι Αρχαίοι Έλληνες το είχαν πετύχει με το επίμονο και επίπονο δούλεμα της φράσης. Αποστρέφονταν τη φλυαρία και την ασάφεια και προτιμούσαν τα «λίγα και καλά». Γνωμικά της τότε εποχής με διαχρονική σημασία όπως του Πυθαγόρα «Μη εν πολλοίς ολίγα λέγε, αλλ' εν ολίγοις πολλά», «ουκ εν τω

πολλώ το ευ, αλλ' εν το ευ το πολύ», «το λακωνίζειν εστί φιλοσοφείν», αλλά και του Πλάτωνα «του λόγου μέτρον εστί ουχ ο λέγων, αλλά ο ακούων» αποδεικνύουν περίτρανα τα προλεγόμενα. Ομοίως οι Γάλλοι έλεγαν: Έγραψες δέκα λέξεις; Σβήσε τις πέντε. Έγραψες πέντε λέξεις; Σβήσε τις τρεις.

Πολλά παραδείγματα περίληψης μπορεί να συναντήσει κάποιος στην καθημερινή πραγματικότητα. Τίτλοι εφημερίδων, προκαταρτικές επισκοπήσεις (previews – trailers), περιλήψεις (abstracts) επιστημονικών άρθρων, κριτικές βιβλίων ή ταινιών, πίνακες περιεχομένων βιβλίων ή περιοδικών, προγράμματα τηλεόρασης, δελτία χρηματιστηρίου, δελτία καιρού, βιογραφικά, επικήδειοι, χάρτες, διαφημιστικοί κατάλογοι προϊόντων, ιστοσελίδες που περιέχουν πηγές για μια συγκεκριμένη θεματική περιοχή, αναδρομικοί απολογισμοί γεγονότων, στατιστικές παρουσιάσεις γεγονότων, τα σημαντικότερα σημεία (highlights) ενός γεγονότος, ενός διαλόγου, μιας συνάντησης, μιας κατάθεσης κτλ., περιλήψεις στα πίσω εξώφυλλα βιβλίων, δεδομένα ανάλυσης κίνησης του ιστοχώρου, διαφόρων ειδών λογαριασμοί είναι μόνο μερικά απ' αυτά.

Χρίζει αναφοράς ότι μια περίληψη δεν είναι μόνο σε μορφή κειμένου, αλλά σε πολλές περιπτώσεις είναι μια εικόνα, μια ταινία, ένα ακουστικό κομμάτι. Το ίδιο και η πηγή που συνοψίζεται μπορεί να είναι σ' αυτές τις διαφορετικές πολυμεσικές μορφές. Επιπλέον, σε πολλές περιπτώσεις η θέση όπου υπάρχει η πληροφορία πηγής δεν είναι μοναδική. Για παράδειγμα, ένας επικήδειος πραγματεύεται σε συνοπτική μορφή τη ζωή ενός ανθρώπού με γεγονότα και συμπεράσματα που συλλέγονται από διάφορες πηγές, συμπεριλαμβανομένων των πληροφοριών από τους συγγενείς, τους φίλους, τους συνεργάτες, τα ΜΜΕ κτλ. Μια περίληψη, επίσης, δεν χρειάζεται στην πραγματικότητα να έχει μια πραγματική και ακριβής πηγή πληροφορίας την ώρα της δημιουργίας της,, όπως στην περίπτωσης της περίληψης μιας προσεχής ομιλίας πάνω σ' ένα συγκεκριμένο θέμα.

Ποιά είναι όμως η ειδοποιός διαφορά μεταξύ μιας περίληψης και άλλων αναπαραστάσεων πληροφορίας εγγράφων; Το βασικό χαρακτηριστικό που ξεχωρίζει τις περιλήψεις από άλλου είδους αναπαραστάσεις είναι η έννοια της συμπύκνωσης (condensation) του περιεχόμενου της πληροφορίας των εγγράφων προς όφελος του χρήστη και της εφαρμογής.. Δηλαδή, η ίδια πληροφορία εισόδου μπορεί ανάλογα με τις απαιτήσεις συμπύκνωσης και το βαθμό της λεπτομέρειας του χρήστη (είτε του συγκεκριμένου χρήστη είτε του είδους του χρήστη) και της εφαρμογής να έχει ως αποτέλεσμα πολύ διαφορετικές περιλήψεις. Μια δεδομένη εφαρμογή μπορεί να έχει διάφορες απαιτήσεις για περίληψη, διαφορετικά μεγέθη – μήκη περιλήψεων, διαφορετικές μορφές εξόδου και διαφορετικές οργανώσεις της πληροφορίας σε κάθε περίπτωση. Όσον αφορά στο μέγεθος της περίληψης, αυτό μπορεί σε γενικές γραμμές να εκτείνεται από μόλις μικρότερο από το μέγεθος της εισόδου έως μόλις μεγαλύτερο από το μηδέν. Αυτό σημαίνει ότι το **ποσοστό συμπύκνωσης (condensation rate)** ή αλλιώς **ποσοστό συμπίεσης (compression rate)** (μέγεθος περίληψης προς μέγεθος πηγής) μπορεί να ποικίλλει από μόλις κάτω από 100% έως μόλις πάνω από 0%.Ως συμφωνία, θεωρείται ότι το ποσοστό συμπύκνωσης του 1% είναι μεγαλύτερο συγκρινόμενο με το 99%, διότι το 99% του κειμένου αφαιρείται στην πρώτη περίπτωση και κρατείται στη δεύτερη. Η εναλλακτική θεώρηση είναι ακριβώς το αντίστροφο.

Οι υπολογιστές μπορούν να βοηθήσουν τους ανθρώπους με διάφορους τρόπους να συνοψίσουν. Καταρχήν, η περιοχή της αυτόματης περίληψης μπορεί να χωριστεί σε τρεις υποπεριοχές:

- **Ανθρώπινη περίληψη βοηθημένη από μηχάνημα (Machine Assisted Human Summarization - MAHS):** Ο άνθρωπος εξάγει την περίληψη και το μηχάνημα επικουρεί (π.χ. παρέχει γλωσσάρια τεχνικών όρων, υποψήφια εδάφια)
- **Μηχανική περίληψη βοηθημένη από άνθρωπο (Human Assisted Machine Summarization - HAMS):** Το μηχάνημα εξάγει την περίληψη και ο άνθρωπος την επεξεργάζεται και τη διορθώνει

- **Πλήρως αυτόματη περίληψη (Fully Automatic Summarization - FAS):** Το μηχάνημα δημιουργεί την περίληψη καθαρά από μόνο του

Ο βαθμός στον οποίο οι δραστηριότητες που πραγματοποιούνται παραδοσιακά από τους ανθρώπους που το επάγγελμά τους είναι η εξαγωγή περιλήψεων μπορούν και πρέπει να αυτοματοποιηθούν εξαρτάται εν μέρει από το ρόλο των μηχανών σε σχέση με αυτόν των ανθρώπων, δηλαδή από την επιλογή FAS (Fully Automated Summarization), HAMS (Human Assisted Machine Summarization), ή MAHS (Machine Assisted Human Summarization). Ανεξάρτητα όμως από αυτήν την επιλογή, η μελέτη του πως λειτουργούν οι άνθρωποι που ασχολούνται με την περίληψη παράγει πολλές χρήσιμες ιδέες που επηρεάζουν τον σχεδιασμό των αυτόματων συστημάτων περίληψης.

Λαμβάνοντας υπόψη τα παραπάνω, θα λέγαμε εν συντομία ότι ένα σύστημα αυτόματης περίληψης κειμένου (summarizer) είναι ένα σύστημα του οποίου ο σκοπός είναι να παράγει μια συμπυκνωμένη αναπαράσταση του περιεχομένου της εισόδου του **για περαιτέρω χρήση από τον άνθρωπο**. Αυτός ο ευρύς ορισμός ξεχωρίζει τις περιλήψεις (τις εξόδους που παράγονται από τους summarizers) από τις μη περιλήψεις. Έχουμε πολλά παραδείγματα από γειτονικά επιστημονικά πεδία, τα οποία απάδουν με την περίληψη κάτω από ορισμένες προϋποθέσεις:

- **Συμπίεση Κειμένου (Text Compression):** Στοχεύει στη συμπύκνωση ενός κειμένου εισόδου με τη μεταχείριση της εισόδου ως κώδικα και την εκμετάλλευση του πλεονασμού αυτής. Παρόλα αυτά, η συμπιεσμένη αναπαράσταση προορίζεται για αποδοτική αποθήκευση και μετάδοση μεταξύ των μηχανημάτων, παρά για περαιτέρω χρήση από τον άνθρωπο
- **Ανάκτηση εγγράφων (Document Retrieval):** Η λήψη μιας συλλογής εγγράφων και μιας ανάγκης του χρήστη και η ανάκτηση εγγράφων σε σχέση μ' αυτή την ανάγκη. Δεν υπάρχει η έννοια της συμπύκνωσης του περιεχομένου της συλλογής. Βέβαια ο τρόπος της αναπαράστασης των αποτελεσμάτων μπορεί να περιλαμβάνει την έννοια αυτή σε κάθε άρθρο που ανακτάται είτε με εξαγωγή του τίτλου, είτε με εξαγωγή της πρώτης γραμμής κ.τ.λ.
- **Ευρετηρίαση (Indexing):**Στοχεύει στον προσδιορισμό κατάλληλων όρων από ένα έγγραφο, συνήθως για να διευκολύνει την ανάκτηση πληροφορίας. Η ευρετηρίαση μπορεί να είναι μια περιορισμένη μορφή περίληψης, όταν συγκεκριμένοι περιγραφείς (descriptors) χρησιμοποιούνται για να βοηθήσουν το χρήστη στο χαρακτηρισμό του περιεχομένου της πληροφορίας του εγγράφου και να επιτρέψουν την αποτελεσματική ανάκτηση του. Παρολ' αυτά, όταν οι όροι που «επισυνάπτονται» δεν προορίζονται για χρήση σε περιλήψεις, αλλά χρησιμοποιούνται κυρίως από ένα σύστημα ανάκτησης, η ευρετηρίαση δεν εξυπηρετεί το ρόλο της περίληψης
- **Εξαγωγή Πληροφορίας (Information Extraction):** Στοχεύει στη συμπλήρωση προτύπων (ή πινάκων) από μια είσοδο σε μορφή φυσικής γλώσσας. Εντούτοις, η συμπύκνωση δεν είναι απαραίτητα ο στόχος ενός συστήματος εξαγωγής πληροφορίας. Αυτό μπορεί να γίνει ένας summarizer μόνο όταν η συμπύκνωση γίνει στόχος, επιτρέποντας π.χ. έναν μεγαλύτερο ή μικρότερο κατάλογο ονομασμένων οντοτήτων (named entities), με χρήση μιας συνάρτησης με όρισμα τον επιθυμητό αριθμό αυτών στην είσοδο. Χρίζει αναφοράς ότι το σύστημα εξαγωγής πληροφορίας δεν παράγει μια περίληψη, εάν η είσοδος δεν ταιριάζει με το πρότυπο και έτσι η συμπύκνωση δεν μπορεί να καθοριστεί σ' αυτές τις περιπτώσεις. Τέλος, η εξαγωγή πληροφορίας συχνά χρησιμοποιείται σε ένα από τα συστατικά μέρη ενός συστήματος περίληψης
- **Εξόρυξη Κειμένου (Text Mining):** Στοχεύει στην ανίχνευση νέων ή ανωμάλων ή αλλιώς ενδιαφερουσών πληροφοριών σε μεγάλες αποθήκες κειμένων. Αν και η έξοδος είναι συνήθως μικρότερη από την είσοδο, η εξόρυξη κειμένου δεν εστιάζει στη συμπύκνωση της πληροφορίας στις αποθήκες. Αντιθέτως, επικεντρώνεται στο χαρακτηρισμό ιδιομορφιών στα δεδομένα

- **Απάντηση ερωτήσεων από μια συλλογή εγγράφων (Question answering from a document collection):** Στοχεύει στη λήψη μιας ερώτησης σε φυσική γλώσσα και της εύρεσης μιας απάντησης μέσω ενός συστήματος από μια συλλογή εγγράφων. Ωστόσο, ο στόχος της απάντησης ερωτήσεων δεν είναι να συμπυκνώνει έγγραφα. Και αυτό γιατί παρέχει χονδρικά μόνο την τιμή κάποια μεταβλητής για την οποία τέθηκε η ερώτηση. Εντούτοις, η απάντηση ερωτήσεων μπορεί να ερμηνευτεί ως περίληψη πληροφοριών του εγγράφου που απαντάει σε ερώτηση. Μια περίληψη μπορεί ωστόσο να συνοψίζει πληροφορίες που σχετίζονται με ένα θέμα (topic) σε ένα έγγραφο. Όταν το θέμα είναι στη μορφή ερώτησης φυσικής γλώσσας, τότε η περίληψη μπορεί να περιλαμβάνει την απάντηση ερωτήσεων [3]

2.1.1 Βασικές έννοιες και παράμετροι της περίληψης

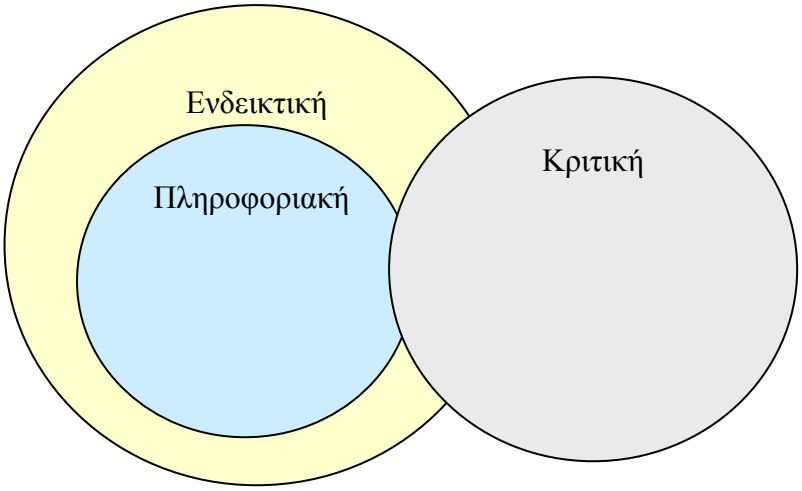
Πολλές βασικές έννοιες της αυτόματης περίληψης επικεντρώνονται στη σχέση μεταξύ της περίληψης και της εισόδου της. Υπάρχει μια θεμελιώδης διαφορά μεταξύ των **εξαγωγών (extracts)** και των **συνόψεων (abstracts).** Μια εξαγωγή είναι μια περίληψη που αποτελείται ολοκληρωτικά από υλικό που αντιγράφεται από την είσοδο. Μπορεί να αποτελείται από προτάσεις, όρους, όπως τεχνικούς όρους, κύρια ονόματα, ονοματικές φράσεις, περικομμένες προτάσεις κτλ. Αντιθέτως, μια σύνοψη είναι μια περίληψη της οποίας τουλάχιστον μερικό από το υλικό της δεν είναι παρόν στην είσοδο, δηλαδή περιλαμβάνει σε κάποιο βαθμό παραφράσεις του περιεχομένου. Τέλος υπάρχει ένα άλλο είδος περίληψης, το οποίο λέγεται **συμπίεση κειμένου (text compaction)** και το οποίο αφορά καταρχήν την εξαγωγή κειμένου από ένα πηγαίο έγγραφο και αφετέρου την περικοπή ή συντόμευση του. Γενικά, οι συνόψεις προσφέρουν τη δυνατότητα για μεγαλύτερο βαθμό συμπύκνωσης*: Μια μικρή σύνοψη μπορεί να προσφέρει περισσότερη πληροφορία απ' ό,τι μια μεγαλύτερη εξαγωγή.*

Οι περιλήψεις μπορούν τώρα παραδοσιακά να χωριστούν σε **ενδεικτικές (indicative)** και **πληροφοριακές (informative)**. Μια ενδεικτική περίληψη (indicative summary) παρέχει μια συνάρτηση αναφοράς για την επιλογή εγγράφων εκ των υστέρων για περισσότερη εμβάθυνση. Μια πληροφοριακή περίληψη (informative summary) καλύπτει όλη την αξιοπρόσεκτη πληροφορία στην πηγή σε κάποιο επίπεδο λεπτομέρειας. Με άλλα λόγια η μεν πρώτη στοχεύει στο να βοηθήσει το χρήστη να αποφασίσει εάν θα διαβάσει το πηγαίο κείμενο ή όχι, ενώ η δεύτερη να κατανοήσει τα βασικά σημεία του πηγαίου κείμενου. Για παράδειγμα στην περίπτωση αναφορών επιστημονικών ερευνών (papers), σύμφωνα με το πρότυπο ANSI (American National Standards Institute), οι ενδεικτικές περιλήψεις θα πρέπει να περιέχουν πληροφορίες για τον σκοπό, το πεδίο δράσης – πλαίσιο (scope) και την προσέγγιση που πραγματοποιείται και όχι για τα αποτελέσματα, τα συμπεράσματα και τις προτάσεις (recommendations). Από την άλλη μεριά, οι πληροφοριακές περιλήψεις θα πρέπει να τα εξετάζουν όλα αυτά.

Ένα άλλο είδος περίληψης που υφίσταται είναι και η κριτική περίληψη. Μια **κριτική περίληψη (critical summary)** αποτιμά το θέμα μελέτης της πηγής, εκφράζοντας τις απόψεις του abstractor σχετικά με την ποιότητα της δουλειάς του συγγραφέα. Οι κριτικές περιλήψεις περιλαμβάνουν κριτικές (reviews), αναδράσεις (feedbacks), προσδιορισμούς αδυναμιών, προτάσεις κτλ. και εν τέλει κάτι εντελώς διαφορετικό από το πηγαίο κείμενο. Είναι εύλογο πως αυτού του είδους οι περιλήψεις είναι πέρα από τον σκοπό και το πεδίο δράσης της αυτόματης περίληψης και είναι περισσότερο κατάλληλες για ανθρώπους abstractors.

Χρίζει αναφοράς ότι τα τρία παραπάνω είδη περίληψης δεν είναι αμοιβαία αποκλειόμενα μεταξύ τους. Αυτό σημαίνει ότι μπορεί να υπάρχουν πληροφοριακές περιλήψεις που να επιτελούν και ενδεικτικές και πληροφοριακές λειτουργίες με αποτέλεσμα οι πληροφοριακές περιλήψεις να μπορούν να ειδωθούν ως ένα γνήσιο υποσύνολο των ενδεικτικών. Παρομοίως, μια κριτική περίληψη μπορεί να τόσο ενδεικτική (π.χ. η αναφορά πως αυτό το κείμενο είναι σπουδαίο παραπέμπει τον αναγνώστη να εντρυφήσει σ' αυτό), όσο και πληροφοριακή. Οι

προαναφερθέντες σχέσεις μεταξύ των παραπάνω ειδών περίληψης απεικονίζονται παρακάτω στο σχήμα 1 μέσω ενός διαγράμματος Venn:



Σχήμα 1 - Σχέσεις μεταξύ ενδεικτικών, πληροφοριακών και κριτικών περιλήψεων

Ένας άλλος τρόπος για να χαρακτηριστεί το περιεχόμενο πληροφοριών μιας περίληψης είναι να μετρηθεί η **σημασιολογική της πληροφοριακότητα (semantic informativeness)**. Η σημασιολογική πληροφοριακότητα βασίζεται στο ότι αντί να μεταχειρίζονται όλες οι προτάσεις ως εξίσου σημαντικές, υπάρχει κάποια αριθμητική συνάρτηση για τον υπολογισμό της σημαντικότητας ή του βάρους μιας πρότασης.

Μια εναλλακτική προσέγγιση είναι αν χαρακτηριστεί η πληροφοριακότητα σύμφωνα με την θεωρητική άποψη ότι μια περίληψη μπορεί να θεωρηθεί πληροφοριακή εάν επιτρέπει την αναδημιουργία του πηγαίου εγγράφου με βάση μόνο αυτή. Συνεπώς, αν ζητούταν από κάποιον να μαντέψει το περιεχόμενο του πηγαίου κειμένου μελετώντας την περίληψη, η καλύτερη περίληψη θα ήταν αυτή που θα επέτρεπε σ' αυτόν να μαντέψει σωστά το πλήρες κείμενο του πηγαίου εγγράφου, λέξη προς λέξη, γράμμα προς γράμμα. Η θεωρητική πληροφοριακότητα ενός κειμένου μπορεί έτσι να χαρακτηριστεί σε σχέση με την προβλεψιμότητα του υπολοίπου του κειμένου, δεδομένου ενός αρχικού μέρους αυτού. Εάν τώρα η περίληψη είναι μια εξαγωγή, θα μπορούσε να καθοριστεί πόσο καλά το υπόλοιπο του πηγαίου κειμένου μπορεί να μαντευτεί γράμμα προς γράμμα (ή λέξη προς λέξη) δεδομένου της εξαγωγής. Η έννοια της προβλεψιμότητας ενός πηγαίου κειμένου εκφράζεται από την εντροπία της πληροφορίας, η οποία είναι το ποσό της πληροφορίας που περιέχεται σε μια τυχαία μεταβλητή, η το μέσο μήκος ενός μηνύματος που χρειάζεται να μεταδοθεί ένα αποτέλεσμα (outcome) αυτής της μεταβλητής. Εν τέλει, θα μπορούσαμε να πούμε ότι η καλύτερη περίληψη θα ήταν αυτή που θα επέτρεπε να μαντευτούν σωστά όλες οι σημαντικές (salient) ιδέες ή προτάσεις στο πλήρες κείμενο του πηγαίου εγγράφου.

Τί εννοούμε όμως λέγοντας σημαντικό (salient) περιεχόμενο και εν τέλει **σημαντικότητα (salience)** του περιεχομένου; *Η* **σημαντικότητα (salience)** *ή αλλιώς* **συνάφεια (relevance)** *είναι το βάρος που συνδέεται με τις πληροφορίες σ' ένα έγγραφο, εκφράζοντας και το περιεχόμενο του εγγράφου αλλά και τη σχέση της πληροφορίας του εγγράφου με την εφαρμογή*. Έχει μεγάλη πρακτική εφαρμογή στην περιοχή της αυτόματης περίληψης, όπως και σε άλλες περιοχές (π.χ. στην ανάκτηση πληροφορίας, στις μηχανές αναζήτησης) και υπάρχουν πολλοί αλγόριθμοι που την εφαρμόζουν.

Μια άλλη κρίσιμη έννοια την περιοχή αυτή είναι αυτή της ***συνοχής (coherence)****, δηλαδή του τρόπου με τον οποίο τα μέρη του κειμένου συλλέγονται μαζί για να σχηματίσουν ένα*

ολοκληρωμένο σύνολο. Ένα ανακόλουθο (incoherent) κείμενο είναι αυτό που είναι ασυνεχές και του οποίου οι προτάσεις δεν συνδέονται ώστε να σχηματίσουν ένα συνεπές λογικά (coherent) σύνολο. Αυτό μπορεί να οφείλεται σε αναφορές που είναι αδιασαφήνιστες στο κείμενο, κενά στον ειρμό, προτάσεις που επαναλαμβάνουν την ίδια ή παρόμοια ιδέα (το οποίο λέγεται **πλεονασμός – redundancy**), έλλειψη καλής οργάνωσης κτλ. Μια ακραία περίπτωση ενός ανακόλουθου κειμένου είναι μια τυχαία συλλογή προτάσεων. Σημειώνεται εδώ ότι το επίπεδο της ανοχής για την ανακολουθία (incoherence) σε μια περίληψη ποικίλει ανάλογα με την εφαρμογή. Μερικές εφαρμογές μπορούν να ικανοποιηθούν με περιλήψεις που είναι αποσπασματικές (π.χ. λίστα λέξεων ή φράσεων συμπεριλαμβανομένων). Επιπλέον, η συνοχή εξαρτάται σε μεγάλο βαθμό από τη μορφή της εξόδου. Μια περίληψη μπορεί να μορφοποιηθεί ως ένας πίνακας, ως ένα έγγραφο με διαφορετικές ενότητες, κεφαλίδες κτλ. Μπορεί να διαταχθεί με διαφορετικούς τρόπους, οριζόντια κατά μήκος της σελίδας, κάθετα, σε στήλες κ.τ.λ.

Ένας άλλος τρόπος διάκρισης μεταξύ των περιλήψεων βασίζεται στον τύπο του χρήστη που η περίληψη προορίζεται. Υπάρχουν λοιπόν **οι περιλήψεις που εστιάζουν στον χρήστη (user-focused)**, οι οποίες διακρίνονται σ' αυτές **που εστιάζουν στο θέμα (topic-focused)** και σ' αυτές **που εστιάζουν στην ερώτηση (query-focused)** και οι **γενικής χρήσης (generic) περιλήψεις**. Οι μεν πρώτες προσαρμόζονται στις απαιτήσεις ενός συγκεκριμένου χρήστη ή μιας ομάδας χρηστών. Με άλλα λόγια, η περίληψη λαμβάνει υπόψη κάποια αναπαράσταση των ενδιαφερόντων των χρηστών, η οποία μπορεί να εκτείνεται από πραγματικά μοντέλα χρηστών μέχρι προφίλ που καταγράφουν όρους θεματικής περιοχής ή ακόμα μια συγκεκριμένη ερώτηση που περιέχει όρους που εκφράζουν την ανάγκη πληροφορίας ενός χρήστη. Γενικότερα, εξετάζει τα ενδιαφέροντα και το υπόβαθρο του χρήστη καθώς επίσης και το περιεχόμενο του εγγράφου. Μια ιδιαίτερα ενδιαφέρουσα μορφή περίληψης που εστιάζει στον χρήστη είναι αυτή που επιστρέφεται σε απάντηση μιας ερώτησης (query-focused) (π.χ. αυτή η περίληψη μπορεί να είναι μια σύντομη απάντηση στην οποία οι πραγματικές πληροφορίες έχουν επιλεγεί από μία ή περισσότερες πηγές). Από την άλλη μεριά, η άλλη μορφή περίληψης που εστιάζει στο χρήστη - αυτή που εστιάζει στο θέμα (topic-focused) - στη γενική μορφή της περιλαμβάνει ένα θέμα που δεν εκφράζεται με τη μορφή μιας ερώτησης, οπότε στη συγκεκριμένη περίπτωση η περίληψη που προκύπτει περιέχει πληροφορίες τις σχετικές με το θέμα.

Από την άλλη μεριά, *οι δε δεύτερες στοχεύουν σε ένα συγκεκριμένο – συνήθως ευρύ – αναγνωστικό κοινό*. Παραδοσιακά γενικής χρήσης περιλήψεις που γράφονται από συγγραφείς η επαγγελματίες abstractors εξυπηρετούσαν ως υποκατάστατα του πλήρους κειμένου. Αυτές μπορούν να είναι ενδεικτικές ή πληροφοριακές.

Αξίζει τέλος να σημειώσουμε ότι τελευταία οι πρώτες έχουν προσλάβει μια αυξανόμενη σπουδαιότητα, λόγω των τεχνικών διευκολύνσεων (π.χ. εξατομικευμένο φιλτράρισμα πληροφοριών) που παρέχουν αφειδώς τα υπολογιστικά περιβάλλοντα.

Οι περιλήψεις μπορούν ακόμα να διαχωριστούν **σε ενός εγγράφου (single document)** και σε **πολλαπλών εγγράφων (multi-document)**. Η διαφορά είναι ότι οι μεν πρώτες προέρχονται από ένα πηγαίο έγγραφο, οι δε δεύτερες από πολλά. Στην περίληψη που προέρχεται από πολλά πηγαία έγγραφα (Multi-Document Summarization – MDS) και στην οποία εμείς θα εστιάσουμε, σε γενικές γραμμές το σύστημα περίληψης (summarizer) προσδιορίζει για παράδειγμα τι είναι κοινό μεταξύ των εγγράφων ή διαφορετικό σε ένα συγκεκριμένο έγγραφο σε σχέση βέβαια με τα άλλα. Παρακάτω θα αναφερθούμε εκτενέστερα στη συγκεκριμένη υποπεριοχή της αυτόματης περίληψης κειμένου και συγκεκριμένα θα εξετάσουμε επισταμένως **την αυτόματη περίληψη από πολλαπλά έγγραφα γεγονότων που εξελίσσονται με το χρόνο (Multi-Document Summarization Of Evolving Events)**.

Επίσης οι περιλήψεις μπορεί να είναι **μονογλωσσικές (monolingual)**, **πολυγλωσσικές (multilingual)** και **ετερογλωσσικές (cross-lingual)**. Οι μεν πρώτες είναι συνυφασμένες με

μια μόνο γλώσσα με την είσοδο και την έξοδο προφανώς γραμμένες στην ίδια γλώσσα. Οι δε δεύτερες σχετίζονται με διάφορες γλώσσες με την είσοδο και την έξοδο όμως γραμμένες στην ίδια γλώσσα. Οι δε τρίτες αφορούν και αυτές διάφορες γλώσσες, αλλά με τη διαφορά τώρα ότι η είσοδος και η έξοδος είναι γραμμένες σε διαφορετικές γλώσσες.

Σημειώνεται ότι οι περιλήψεις μπορούν επίσης να περιορίζονται σε μια συγκεκριμένη **υπογλώσσα (sublanguage)**. Για παράδειγμα ένα τεχνικό εγχειρίδιο μπορεί να χρησιμοποιεί ένα συγκεκριμένο, εξειδικευμένο λεξιλόγιο. Επίσης, οι περιλήψεις που δημιουργούνται για μαθητές ή τουρίστες ή ξένους μπορεί να πρέπει να χρησιμοποιούν περιορισμένα λεξιλόγια και απλούστερες δομές.

Μια άλλη παράμετρος που παίζει ρυθμιστικό ρόλο στην αυτόματη περίληψη είναι το **ύφος (genre)**. Ο όρος ύφος χρησιμοποιείται για να υποδείξει διαφορετικές κατηγορίες κειμένων. Έτσι ένα σύστημα περίληψης μπορεί να χρησιμοποιεί ειδικές στρατηγικές για διαφορετικές κατηγορίες κειμένου, όπως για παράδειγμα επιστημονικές αναφορές, τεχνικές αναφορές, ιστορίες ειδήσεων, μηνύματα ηλεκτρονικού ταχυδρομείου, βιβλία κ.τ.λ. Χρίζει αναφοράς ότι λόγω του ότι δεν υπάρχει μια τυποποιημένη ταξινόμηση αυτών των κατηγοριών κειμένου, το ύφος κατά κάποιο τρόπο παρέχει την ειδοποιό διαφορά.

Οι περιλήψεις επίσης μπορούν να έχουν ως είσοδο, όπως και έξοδο **διάφορους τύπους μέσων (media types)** (π.χ. κείμενο, ήχο, εικόνες, διαγράμματα, πίνακες). Στην **πολυμεσική περίληψη (Multimedia Summarization)** τόσο η είσοδος όσο και η έξοδος μπορούν να αποτελούνται από έναν συνδυασμό από τους διαφορετικούς τύπους μέσων αλλά και η είσοδος και η έξοδος μπορούν να είναι διαφορετικοί τύποι μέσων (π.χ. η είσοδος να είναι κείμενο και η έξοδος να είναι ομιλία (speech) ή το αντίστροφο).

Σε μια δεδομένη εφαρμογή, η σπουδαιότητα αυτών των παραμέτρων ποικίλλει ανάλογα με τις απαιτήσεις και τις καταστάσεις - φάσεις αυτής. Πάντως γεγονός είναι πως είναι απίθανο ένα σύστημα περίληψης να μπορεί να χειριστεί όλες τις προαναφερθέντες παραμέτρους. Συνολικά, οι παράμετροι που προαναφέρθηκαν και επηρεάζουν και διακρίνουν ένα σύστημα περίληψης συνοψίζονται στον παρακάτω πίνακα 1 [3]:

Παράμετρος	Σχόλιο - διάκριση
Ποσοστό συμπίεσης (compression rate)	Μέγεθος περίληψης / Μέγεθος πηγαίου κείμενου
Ακροατήριο (audience)	Με επίκεντρο τον χρήστη (User-focused) – Γενικής χρήσης (Generic)
Σχέση με την πηγή (Relation to source)	Εξαγωγή (Extract) – Σύνοψη (Abstract)
Λειτουργία (Function)	Ενδεικτική (Indicative) – Πληροφοριακή (Informative) – Κριτική (Critical)
Συνοχή (Coherence)	Συνοχή (Coherence) – Ασυνέπεια (Incoherence)
Έκταση (Span)	Ενός εγγράφου (Single-Document) – Πολλαπλών εγγράφων (Multi-Document)
Γλώσσα (Language)	Χρήση μιας γλώσσας (Monolingual) –Χρήση πολλαπλών γλωσσών με την είσοδο και την έξοδο στην ίδια γλώσσα (Multilingual) – Χρήση πολλαπλών γλωσσών με την είσοδο και την έξοδο σε διαφορετική γλώσσα (Cross-lingual)
Ύφος (genre)	Επιστημονικές Αναφορές (Scientific Reports) – Technical Reports (Τεχνικές Αναφορές) – Ιστορίες ειδήσεων (News Stories) – Μηνύματα Ηλεκτρονικού Ταχυδρομείου (Email Messages) – Βιβλία (Books) κ.τ.λ.
Μέσα (Media)	Κείμενο (Text) – Ήχος (Audio) - Εικόνες (Pictures) - Διαγράμματα (Diagrams) - Πίνακες (Tables)

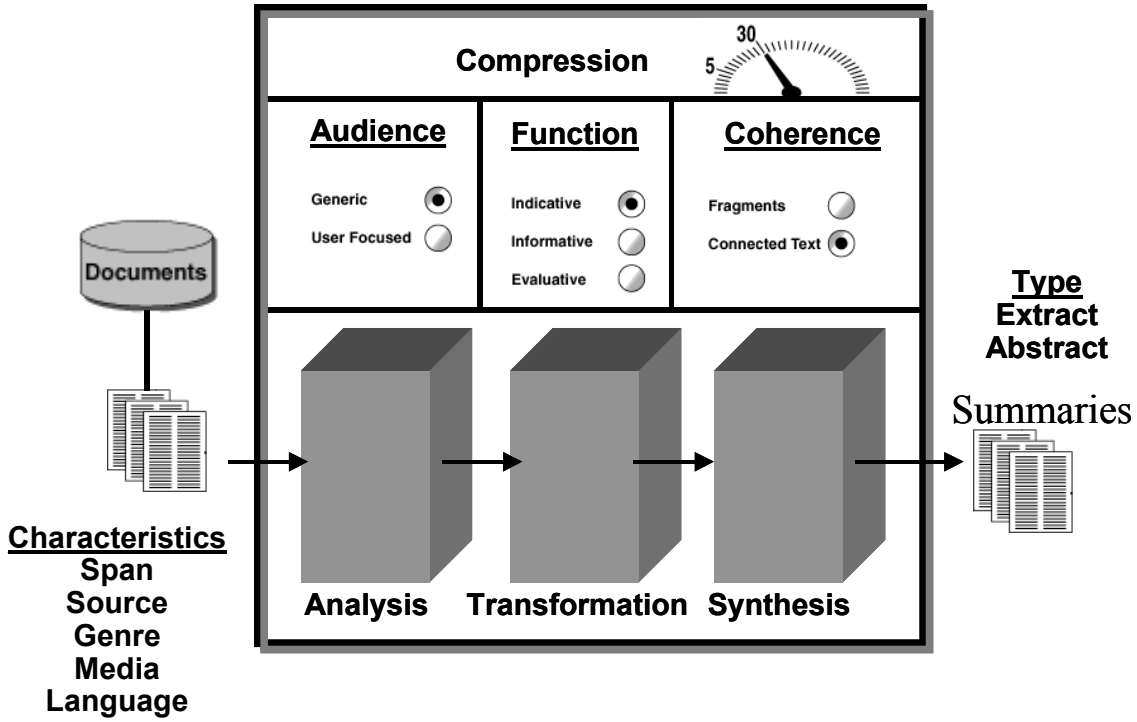
Πίνακας 1- Παράμετροι ενός συστήματος περίληψης

2.1.2 Η Αρχιτεκτονική ενός συστήματος περίληψης

Μια αρχιτεκτονική ενός συστήματος περίληψης και συγκεκριμένα **σύνοψης** μπορεί να περιλαμβάνει κάποιες από τις προαναφερθείσες παραμέτρους (συνήθως οι σημαντικότερες είναι το ποσοστό συμπίεσης, το ακροατήριο, η λειτουργία και η συνοχή) της περίληψης. Το ποσοστό συμπίεσης συνήθως κυμαίνεται μεταξύ 5% και 30%. Η διαδικασία περίληψης θεωρείται ότι αποτελείται βασικά από τρεις φάσεις:

- **Ανάλυση (Analysis):** Αυτή η φάση αναλύει την είσοδο και δημιουργεί μια εσωτερική αναπαράσταση αυτής
- **Μετασχηματισμός (Transformation) ή ραφινάρισμα (Refinement):** Μετατρέπει την εσωτερική αναπαράσταση σε μια αναπαράσταση της περίληψης. Αυτή η φάση είναι εφαρμόσιμη στα συστήματα που παράγουν συνόψεις ή επιτελούν συμπίεση (compaction) ή πολλαπλών εγγράφων περίληψη (multi-document summarization). Τα συστήματα που παράγουν εξαγωγές από απλά έγγραφα (single-document extracts) χωρίς συμπίεση τείνουν να πηγαίνουν άμεσα από τη φάση της ανάλυσης στην έξοδο
- **Σύνθεση (Synthesis):** Η αναπαράσταση της περίληψης μετατρέπεται σε φυσική γλώσσα

Παρακάτω, στο σχήμα 2 απεικονίζεται μια υψηλού επιπέδου αρχιτεκτονική ενός συστήματος περίληψης:



Σχήμα 2 – Μια υψηλού επιπέδου αρχιτεκτονική ενός συστήματος περίληψης

Σε ένα χαμηλότερο τώρα επίπεδο προσέγγισης υπάρχουν τρεις βασικές λειτουργίες συμπύκνωσης (condensation operations), τις οποίες τα συστήματα περίληψης εκτελούν σε καθεμία από τις παραπάνω φάσεις:

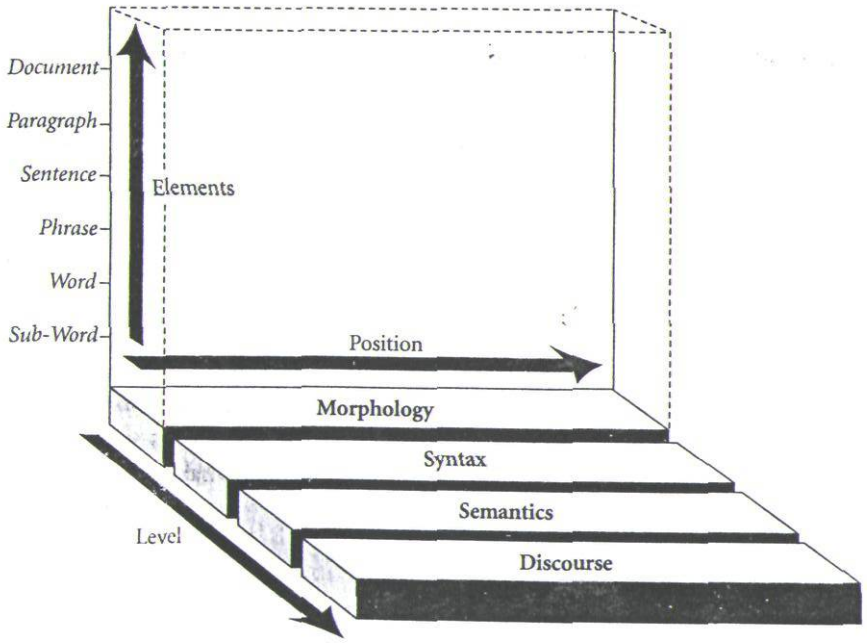
- **Επιλογή (Selection):** Φιλτράρισμα των στοιχείων
- **Συνάθροιση (Aggregation):** Συγχώνευση των στοιχείων
- **Γενίκευση (Generalization):** Αντικατάσταση των στοιχειών με γενικότερα

Άλλες περισσότερο σύνθετες λειτουργίες είναι οι **παραφράσεις** και οι **απλοποιήσεις**. Αυτές οι λειτουργίες εκτελούνται σε διάφορα στοιχεία (elements), όπως λέξη (word), φράση (phrase), πρόταση (clause), πρόταση (sentence), ή ομιλία (discourse). Επίσης η παράγραφος (paragraph) μπορεί να θεωρηθεί στοιχείο, παρόλο που αναφέρεται ειδικά σε μορφοποίηση γραπτού κειμένου. Εν τέλει στα στοιχεία συμπεριλαμβάνεται και το στοιχείο εγγράφου (document element), αν και δεν είναι ένα παραδοσιακό γλωσσολογικό στοιχείο.

Τα στοιχεία μπορούν εν συνεχεία να αναπαρασταθούν σε διαφορετικά **επίπεδα (levels)** της γλωσσολογικής ανάλυσης:

- **Μορφολογικό (Morphological):** Μελετάει τις ελάχιστες σημασιολογικές μονάδες της γλώσσας, τα μορφήματα
- **Συντακτικό (Syntactic):** Εξετάζει τη συντακτική δομή και λειτουργία μιας γλώσσας, τους τρόπους που οι λέξεις συντάσσονται, συνδέονται μεταξύ τους για να σχηματίσουν φράσεις, προτάσεις κ.τ.λ.
- **Σημασιολογικό (Semantic):** Εξετάζει ό,τι έχει σχέση με τη σημασία: Τα διάφορα είδη της σημασίας, τις ποικίλες σημασιολογικές σχέσεις (συνωνυμίας, αντωνυμίας κτλ), την ανάλυση της σημασίας, τη μεταβολή της σημασίας, τη σχέση της σημασίας με τα άλλα επίπεδα της γλώσσας και με τις πραγματικές συνθήκες επικοινωνίας, το λεξιλόγιο μιας γλώσσας κ.τ.λ.
- **Πραγματολογικό (Discourse / Pragmatic):** Εξετάζει τους τρόπους με τους οποίους το περιβάλλον , γλωσσικό ή εξωγλωσσικό, επιδρά στην ερμηνεία μιας πρότασης, όπως αυτή πραγματώνεται ως εκφώνημα μέσα σε συγκεκριμένο χώρο και χρόνο (στις πραγματικές συνθήκες επικοινωνίας) και σε συγκεκριμένα συμφραζόμενα

Οι σχέσεις μεταξύ των στοιχείων, των επιπέδων και των λειτουργιών μπορούν να απεικονιστούν σε έναν πολυδιάστατο χώρο. Το παρακάτω σχήμα 3 απεικονίζει τη δομή αυτού του χώρου. Τα στοιχεία παραθέτονται στον κάθετο άξονα. Η **θέση (position)** εκφράζει τη σειρά των στοιχείων στην είσοδο. Τα επίπεδα απεικονίζονται στην τρίτη διάσταση. Κάθε διαδοχικό επίπεδο υποδηλώνει ένα βαθύτερό του. Η ανάλυση θεωρείται η διαδικασία που πηγαίνει από αβαθή σε βαθύτερα (πιο σημασιολογικά και πραγματικά) επίπεδα. Η σύνθεση ακολουθεί ακριβώς την αντίθετη κατεύθυνση [3].

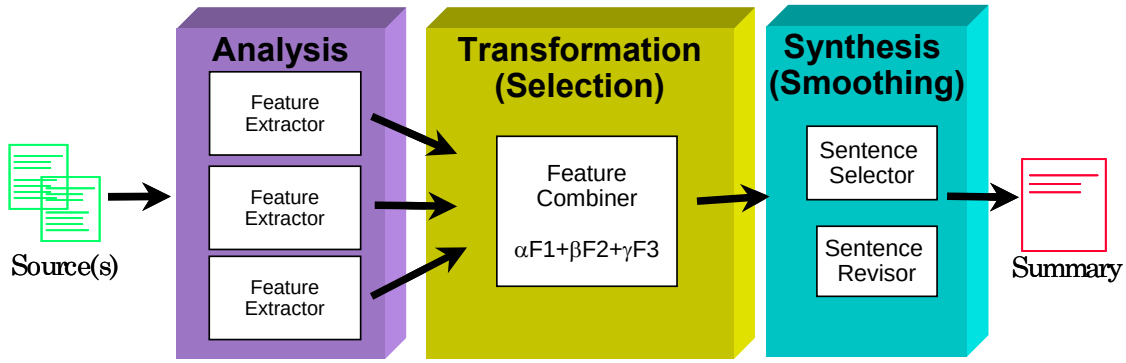


Σχήμα 3 – Ο γλωσσολογικός χώρος

2.1.3 Προσεγγίσεις της περίληψης

Οι βασικές μέθοδοι της περίληψης μπορούν να προσδιοριστούν σε σχέση με το επίπεδο στο γλωσσολογικό χώρο. Υπάρχουν δύο ευρείες προσεγγίσεις:

- **Ρηχές προσεγγίσεις (Shallow Approaches):** Αυτές δεν επιχειρούν αναπαράσταση πέρα από το συντακτικό επίπεδο, αν και διαφορετικά στοιχεία μπορούν να αναπαρασταθούν σε διαφορετικά επίπεδα. Για παράδειγμα, οι λέξεις μπορούν να αναλυθούν μέχρι και το σημασιολογικό επίπεδο, αλλά οι προτάσεις μπορούν να αναλυθούν το πολύ μέχρι το συντακτικό επίπεδο. Αυτές οι προσεγγίσεις συνήθως παράγουν εξαγωγές, συχνά με την εξαγωγή προτάσεων, οι οποίες μάλιστα μπορούν να εξαχθούν αποσπασματικά. Η φάση μετασχηματισμού περιλαμβάνει την επιλογή σημαντικών μονάδων (units). Η φάση σύνθεσης τώρα αναλαμβάνει να επιδιορθώσει τις ανακολουθίες που προκαλούνται από τέτοιες εξαγωγές ή την αναδιοργάνωση του κειμένου κ.τ.λ., κάνοντας το κείμενο περισσότερο συμπαγές (βλέπε παρακάτω το σχήμα 4). Εν ολίγοις, αυτές οι προσεγγίσεις περιορίζονται στην εξαγωγή των σημαντικών μερών του πηγαίου κειμένου και έπειτα στη διευθέτηση αυτών και στην παρουσίασή τους με αποδοτικό τρόπο. Ένα μεγάλο πλεονέκτημα αυτής της προσέγγισης είναι ότι για την εκπαίδευση μπορεί να χρησιμοποιηθεί μια συλλογή κειμένων (corpus)



Σχήμα 4 – Αρχιτεκτονική της ρηχής προσέγγισης

- **Βαθύτερες προσεγγίσεις (Deeper Approaches):** Αυτές επιχειρούν τουλάχιστον ένα προτασιακό σημασιολογικό επίπεδο αναπαράστασης. Παράγουν συνόψεις και η φάση σύνθεσης εδώ συνήθως περιλαμβάνει παραγωγή φυσικής γλώσσας από ένα σημασιολογικό η πραγματολογικό επίπεδο αναπαράστασης. Μπορεί να παράγεται συνεπές λογικά κείμενο με την εφαρμογή διάφορων κανόνων για το πως τα τμήματα του πραγματολογικού επιπέδου μπορούν να συνδεθούν μεταξύ τους. Αυτές οι προσεγγίσεις χρειάζονται κάποια κωδικοποίηση για συγκεκριμένα πεδία. Πολλές απ' αυτές τις προσεγγίσεις έχουν ξεκινήσει με χρήση δομημένων δεδομένων ως πηγή εισόδου και παράγουν για παράδειγμα περιλήψεις εποχιακής απόδοσης από πίνακες στατιστικών καλαθοσφαίρισης. Άλλες προσεγγίσεις χρησιμοποιούν σημασιολογικές προσεγγίσεις μόνο για ορισμένα στοιχεία. Εν' τέλει, οι βαθύτερες προσεγγίσεις προσφέρουν συνήθως περισσότερο πληροφοριακές περιλήψεις

Υπάρχουν επίσης και **υβριδικές προσεγγίσεις (hybrid approaches)** που όλο και περισσότερο χρησιμοποιούνται στην περίληψη. Αυτές περιλαμβάνουν την αφαίρεση στοιχείων κειμένου προκειμένου να το ανάγουν σε περισσότερο συμπαγές, καθώς και αναδιοργάνωση του κειμένου. Εντούτοις, μπορούν να εκτελεστούν χωρίς την απαίτηση χρήσης προτασιακής σημασιολογίας (αν και είναι χρήσιμη η γνώση του τι αναδιοργανώνεται). Αυτού του είδους η προσέγγιση έχει εφαρμοστεί αποτελεσματικότατα στην περίληψη που προέρχεται από πολλά πηγαία έγγραφα, όπου διαφορετικά στοιχεία κειμένου που προέρχονται από διαφορετικές πηγές συγχωνεύονται για να παράγουν συνόψεις [3].

2.1.4 Τρέχουσες εφαρμογές

Υπάρχει μια πληθώρα από εργαλεία και προϊόντα αυτόματης περίληψης που είναι διαθέσιμα στην αγορά, με ίσως ευρύτερα γνωστό το AutoSummarize του Microsoft Office. Υπάρχουν επίσης και πολλές τρέχουσες ερευνητικές εφαρμογές στην περιοχή αυτή. Μερικές από τις εφαρμογές αυτές παρουσιάζονται συνοπτικά παρακάτω:

- **Περιλήψεις ειδήσεων πολυμέσων (Multimedia news summaries):** Αυτή η τεχνολογία επιτρέπει την ανακατασκευή των πολυμεσικών ειδήσεων (π.χ. αυτόματη παρακολούθηση των ειδήσεων απουσία του ενδιαφερομένου και συνοπτική περιγραφή αυτών αφού επιστρέψει)
- **Αρωγή σε γιατρούς (physicians aid):** Παρέχει στους γιατρούς περιλήψεις από την ιατρική βιβλιογραφία που σχετίζονται με τον ιατρικό φάκελο (medical record) ενός ασθενή (π.χ. σύνοψη και σύγκριση των προτεινόμενων θεραπειών γι' αυτόν τον ασθενή)
- **Περίληψη συνεδρίασης (Meeting Summarization):** Επιτρέπει στο χρήστη να επαναπροσδιορίσει το περιεχόμενο των συνεδριάσεων χρησιμοποιώντας τεχνολογίες αυτόματης αναγνώρισης ομιλίας (automatic speech recognition) και αυτόματης

περίληψης. Οι σημαντικές πληροφορίες της συνεδρίασης μπορούν να παρέχονται αφετέρου ως περιλήψεις. Αυτή η ανεξάρτητη από το πεδίο ενδιαφέροντος προσέγγιση καθιστά δυνατή την περίληψη των αποτελεσμάτων τηλεσυνεδριάσεων σε οποιοδήποτε θέμα. (π.χ. συνοπτική παρουσίαση μιας τηλεσυνεδρίασης λόγω απουσίας του ενδιαφερόμενου χρήστη)

- **Επιτυχίες μηχανών αναζήτησης (Search Engine Hits):** Συνοψίζει τις πληροφορίες στις λίστες επιτυχιών που ανακτώνται από τις μηχανές αναζήτησης
- **Συλλογή νοημοσύνης (intelligence gathering):** Παίρνει μια συλλογή από έγγραφα και παράγει έναν φάκελο για ένα πρόσωπο που αναφέρεται σ' αυτά τα έγγραφα για χρήση από ευφυείς αναλυτές (π.χ. δημιουργία της βιογραφίας ενός προσώπου)
- **Φορητές συσκευές (hand-held devices):** Παρέχει μεγέθους μιας οθόνης περιλήψεις (π.χ. βιβλίων) για χρήση από οποιουδήποτε είδους φορητή συσκευή
- **Αρωγή σε άτομα με ειδικές ανάγκες (Aid for the handicapped):** Παρέχει συμπίεση κειμένου για χρήση από άτομα με ειδικές ανάγκες. Για παράδειγμα μπορεί να υπάρχει ένα σύστημα που σαρώνει μια σελίδα ενός βιβλίου και έπειτα να χρησιμοποιείται ένα σύστημα συμπίεσης κειμένου που να τη διαβάζει συμπιεσμένη σε ένα από τα διάφορα πιθανά επίπεδα συμπίεσης [3]

2.2 Αυτόματη Περίληψη πολλαπλών εγγράφων εξελισσόμενων γεγονότων (Multi – Document Summarization of Evolving Events)

Η αυτόματη περίληψη πολλαπλών εγγράφων (Multi-Document Summarization – MDS) είναι εξ ορισμού η επέκταση της αυτόματης περίληψης ενός εγγράφου (single document summarization) σε συλλογές σχετικών έγγραφων. Με άλλα λόγια είναι η διαδικασία της παραγωγής μιας περίληψης από ένα σύνολο σχετικών πηγαίων εγγράφων. Λόγω της αδόμητης πληροφορίας που υπάρχει στον ιστοχώρο, η οποία έχει ως συνέπεια την επανάληψη και την ανακύκλωση της ίδιας πληροφορίας σε διαφορετικές πηγές πληροφορίας, υπάρχει η ανάγκη για εργαλεία που μπορούν να αφαιρούν την περιττή πληροφορία. Έτσι, είναι χρήσιμο να υπάρχει μια περίληψη που αναγνωρίζει τι είναι κοινό σε μια ποικιλία από σχετικά έγγραφα, ή κατά πόσο συγκεκριμένα έγγραφα σε ένα δεδομένο θέμα διαφέρουν μεταξύ τους. Για παράδειγμα, κάποιος ο οποίος αναζητεί για τις θεραπείες μιας συγκεκριμένης ασθένειας σε μια μεγάλη συλλογή ιατρικής βιβλιογραφίας, ίσως να ήθελε να μπορούσε να συγκρίνει και να αντιπαραβάλει τους διαφορετικούς απολογισμούς. Η οποιαδήποτε βοήθεια που θα μπορούσε να προσφέρει η αυτόματη περίληψη σ' αυτή τη διαδικασία θα ήταν πολύ χρήσιμη [2] [3].

Τώρα ο στόχος ενός συστήματος MDS είναι να πάρει μια πηγή πληροφορίας, να εξαγάγει πληροφορία απ' αυτή, ενώ ταυτόχρονα να αφαιρέσει τον πλεονασμό και να λάβει υπόψη τις ομοιότητες και της διαφορές στο πληροφοριακό περιεχόμενο και να παρουσιάσει το σημαντικότερο περιεχόμενο στο χρήστη σε μια συνοπτική μορφή και με έναν τρόπο σύμφωνο με τις ανάγκες του χρήστη ή της εφαρμογής [3].

Σ' αυτή τη σχετικά νέα περιοχή τρία σημαντικά προβλήματα υφίστανται [1]:

1. Η αναγνώριση και η αντιμετώπιση του πλεονασμού
2. Ο προσδιορισμός των σημαντικών ομοιοτήτων και διαφορών μεταξύ των εγγράφων
3. Η εξασφάλιση της συνοχής της περίληψης

Στην MDS, προκειμένου να συνοψιστεί ένα σύνολο από σχετικά έγγραφα θα πρέπει να προσδιοριστούν οι ομοιότητες και οι διαφορές μεταξύ των εγγράφων. Βέβαια, δεν έχει προσδιοριστεί ακόμα το που αυτές οι ομοιότητες και οι διαφορές πρέπει να στοχεύουν. Οι Αφαντενός και άλλοι [5] προτείνουν ότι οι ομοιότητες και οι διαφορές, τουλάχιστον για τα εξελισσόμενα γεγονότα, πρέπει να αντιμετωπισθούν κάτω από δύο προοπτικές: Τον **χρόνο** και την **πηγή**, που σχετίζονται με σχέσεις μεταξύ των εγγράφων (cross-document relations) και συγκεκριμένα με σχέσεις μεταξύ των μηνυμάτων (βλέπε παρακάτω για την έννοια του μηνύματος).

Οι προαναφερθέντες, όπως ίσως γίνεται αντιληπτό, επικεντρώνονται σε μια υποπεριοχή της αυτόματης περίληψης κειμένου που ονομάζεται αυτόματη περίληψη από πολλαπλά έγγραφα εξελισσόμενων γεγονότων (Multi-Document Summarization of Evolving Events). Τονίζεται ότι αυτή πραγματεύεται την έννοια της παραγωγής **περιλήψεων εξελισσόμενων γεγονότων** και όχι των **εξελικτικών περιλήψεων**. Η διαφορά είναι ότι μια εξελικτική περίληψη S_{k+1} είναι η περίληψη μιας είδησης A_{k+1}, όταν οι ειδήσεις από A_1 έως A_k έχουν υποβληθεί ήδη σε επεξεργασία και έχουν παρουσιαστεί σε περιληπτική μορφή στο χρήστη. Η περίληψη S_{k+1} διαφέρει από την προκάτοχό της, S_k, επειδή περιέχει νέες πληροφορίες και παραλείπει πληροφορίες της S_k. Στις περιλήψεις εξελισσόμενων γεγονότων προκύπτει μία περίληψη ύστερα από την επεξεργασία των ειδήσεων A_1 έως A_k.

Οι Αφαντενός και άλλοι [5] διακρίνουν τα διαφορετικά είδη της εξέλιξης από την άποψη του χρόνου που τα περιστατικά ενός γεγονότος συμβαίνουν και από την άποψη του ρυθμού με τον οποίο οι διάφορες πηγές ειδήσεων δημοσιεύουν τις ειδήσεις τους.

Πρέπει να αναφερθεί ότι κάθε **γεγονός** αποτελείται από διάφορα απλούστερα **περιστατικά (incidents)** ή υπο-γεγονότα (subevents). Παραδείγματος χάριν, στην περιοχή του ποδοσφαίρου, τέτοια υπογεγονότα μπορούν να είναι η απόδοση ενός παίκτη ή μιας ομάδας, οι στόχοι που επιτυγχάνονται, οι πιθανοί τραυματισμοί των παικτών, κ.λπ. Σε μια περιοχή που σχετίζεται με ομήρους, τέτοια υπογεγονότα μπορούν να είναι η κατάληψη ενός κτηρίου, οι διαπραγματεύσεις, τα αιτήματα των τρομοκρατών, το γεγονός ότι ελευθέρωσαν έναν όμηρο, κ.λπ.

Σύμφωνα με τους Αφαντενό και άλλους [5] μπορούμε να κατατάξουμε τα είδη των εξελίξεων ενός γεγονότος στο χρόνο σε δύο κατηγορίες: **Γραμμικές** και **μη-γραμμικές**. Η μεταπτυχιακή αυτή επικεντρώνεται στις μη-γραμμικές εξελίξεις. Οι Αφαντενός και άλλοι [4] [5] αντιμετωπίζουν το πρόβλημα αυτό με την εισαγωγή της έννοιας των **μηνυμάτων** και των **διακειμενικών σχέσεων**, τις οποίες χωρίζουν σε **συγχρονικές** και **διαχρονικές**.

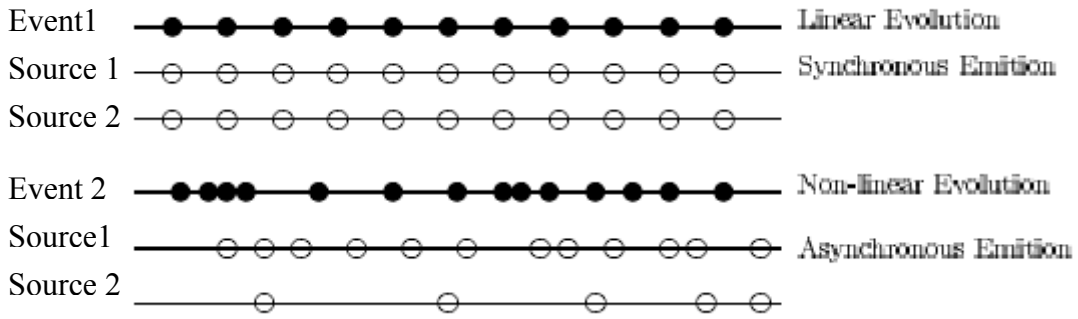
Σχετικά με την εξέλιξη ενός γεγονότος στο χρόνο πραγματοποιείται η διάκριση, όπως προαναφέρθηκε, από τους παραπάνω μεταξύ δύο τύπων εξελίξεων: γραμμική και μη γραμμική εξέλιξη. Στη γραμμική εξέλιξη τα σημαντικότερα περιστατικά ενός γεγονότος συμβαίνουν σε σταθερά και ενδεχομένως προβλέψιμα κβάντα του χρόνου. Αυτό σημαίνει ότι εάν το πρώτο γεγονός q_0 συμβεί τη χρονική στιγμή t_0, κατόπιν κάθε επόμενο γεγονός q_n θα συμβεί στο χρόνο $t_n = t_0 + n*t$, όπου το t είναι το σταθερό χρονικό διάστημα με το οποίο τα γεγονότα συμβαίνουν. Στη μη γραμμική εξέλιξη, αντίθετα, δεν μπορούμε να διακρίνουμε οποιοδήποτε σημαντικό τρόπο που να σχετίζεται με τη σειρά που τα σημαντικότερα περιστατικά ενός γεγονότος συμβαίνουν. Αυτή η διάκριση απεικονίζεται στο παρακάτω σχήμα 5, στο οποίο η εξέλιξη δύο διαφορετικών γεγονότων απεικονίζεται με τους μαύρους κύκλους.

Τα γραμμικά εξελισσόμενα γεγονότα έχουν μια δίκαιη αναλογία στον κόσμο. Συσχετίζονται με τις ανθρώπινες δραστηριότητες που εμφανίζονται σε τακτά χρονικά διαστήματα. Ένα τέτοιο παράδειγμα μπορεί να είναι οι περιγραφές των διάφορων αθλητικών γεγονότων που εμφανίζονται τακτικά. Βέβαια, κάποιος μπορεί να υποστηρίξει ότι τα περισσότερα από τα γεγονότα που βρίσκουμε στις ιστορίες ειδήσεων είναι μη-γραμμικά εξελισσόμενα γεγονότα. Μπορούν να ποικίλουν από πολιτικά γεγονότα, όπως οι εκλογές ή τα διάφορα διεθνή πολιτικά ζητήματα, μέχρι συντριβές αεροπλάνων ή τρομοκρατικά γεγονότα. Εδώ πρέπει να επισημανθεί ότι ιδιαίτερη έμφαση δίνεται από τους Αφαντενό και άλλους [5] στο πεδίο των περιστατικών που αναφέρονται σε ομήρους.

Από την άποψη της αναφοράς ενός γεγονότος από διάφορες πηγές μπορεί να γίνει η διάκριση μεταξύ της **σύγχρονης** και **ασύγχρονης εκπομπής**. Αυτή η διάκριση απεικονίζεται στο παρακάτω σχήμα 5 με άσπρους κύκλους.

Τώρα, θα πρέπει να αναφερθεί ότι καλούνται **συγχρονικές σχέσεις** εκείνες οι σχέσεις που σχετίζονται με τις ομοιότητες και τις διαφορές, μεταξύ των διάφορων πηγών, στον ίδιο χρονικό ορίζοντα και **διαχρονικές σχέσεις** εκείνες οι σχέσεις που σχετίζονται με την εξέλιξη ενός γεγονότος καθώς περιγράφεται από μια πηγή.

Στις περισσότερες από τις περιπτώσεις, όταν έχουμε ένα γεγονός που εξελίσσεται γραμμικά θα έχουμε επίσης μια σύγχρονη εκπομπή των αναφορών των γεγονότων, δεδομένου ότι οι διάφορες πηγές μπορούν εύκολα να προσαρμοστούν στον τρόπο της εξέλιξης ενός γεγονότος. Αυτό δεν μπορεί να ειπωθεί για την περίπτωση της μη γραμμικής εξέλιξης, που έχει ως αποτέλεσμα την ασύγχρονη εκπομπή των αναφορών των γεγονότων από τις διάφορες πηγές.



Σχήμα 5 - Γραμμικές – Μη γραμμικές Εξελίξεις, Σύγχρονη και Ασύγχρονη Εκπομπή

Η μεταπτυχιακή αυτή εργασία εστιάζει και στην έννοια των μηνυμάτων. Τα μηνύματα αναπαριστούν τα περιστατικά των γεγονότων Ένα μήνυμα σύμφωνα με τους Αφαντενό και άλλους [5] αποτελείται από:

- Τον τύπο του
- Ένα σύνολο ορισμάτων, τα οποία παίρνουν τιμές από την οντολογία του υπό εξέταση χώρου
- Την πηγή από την οποία προέρχεται το μήνυμα (m.source)
- Την χρονική στιγμή της έκδοσης του μηνύματος, καθώς και την χρονική στιγμή στην οποία πραγματικά αναφέρεται το μήνυμα (m.time)

Δηλαδή, **message_type(arg_1,…,arg_n), όπου $arg_j \in$ Domain Ontology**

Ο τύπος μηνύματος αναπαριστά τον τύπο του γεγονότος, ενώ τα ορίσματα αναπαριστούν τις κύριες οντότητες που περιλαμβάνονται σ' αυτό το γεγονός. Είναι δυνατό μερικά μηνύματα να συνοδεύονται από μερικούς περιορισμούς στα ορίσματά τους, οι οποίοι αντικατοπτρίζουν τους διάφορους πραγματικούς περιορισμούς.

Όσον αφορά στην πηγή, το m.source κληρονομείται από το έγγραφο που περιέχει το μήνυμα. Αυτό δεν μπορεί να ειπωθεί για το χρόνο επίσης, δεδομένου ότι ο χρόνος που συμβαίνουν τα περιστατικά μπορεί να είναι διαφορετικός από το χρόνο έκδοσης. Αυτό εκφράζεται στο έγγραφο από μια χρονική έκφραση. Κατά συνέπεια, προκειμένου να καθοριστεί ο πραγματικός χρόνος ενός μηνύματος πρέπει να ερμηνεύσουμε εκείνη την έκφραση σε σχέση με το χρόνο της δημοσίευσης του εγγράφου.

Τα παραδείγματα των προδιαγραφών των μηνυμάτων, για μια γραμμική και μη γραμμική εξέλιξη παρουσιάζονται στον παρακάτω πίνακα 2. Τα ορίσματα για κάθε μήνυμα προέρχονται από την οντολογία περιοχής. Κατά συνέπεια, παραδείγματος χάριν, το όρισμα δραστηριότητας (Activity) στο δεύτερο μήνυμα αντιστοιχεί σε ένα σύνολο δραστηριοτήτων που καθορίζονται στην οντολογία της περιοχής. Οι προδιαγραφές για το πρώτο μήνυμα προέρχονται από την περιοχή των αγώνων ποδοσφαίρου [4] και αντιπροσωπεύει την απόδοση ενός παίκτη ή μιας ομάδας για μια συγκεκριμένη περίοδο και μια συγκεκριμένη περιοχή δράσης (π.χ. την άμυνα). Οι προδιαγραφές του δεύτερου μηνύματος προέρχονται από το θέμα που συσχετίζεται με τους ομήρους. Αυτό το μήνυμα αναπαριστά το γεγονός ότι έχουμε μια διαπραγμάτευση μεταξύ δύο οντοτήτων σχετικά με μια συγκεκριμένη δραστηριότητα (π.χ. την απελευθέρωση μερικών ομήρων).

Linear	Non-Linear
performance (entity, in_what, time_span, value)	negotiate ($entity_1$, $entity_2$, about)
entity: Player or Team	$entity_1$: Person
in_what: Action Area	$entity_2$: Person
time_span: Minute or Duration	about : Activity
value: Degree	

Πίνακας 2 – Παραδείγματα Προδιαγραφών Μηνυμάτων

Όπως προαναφέρθηκε, οι σχέσεις μεταξύ των εγγράφων που ικανοποιούνται μεταξύ των μηνυμάτων μπορούν να χωριστούν σε συγχρονικές και διαχρονικές. Οι συγχρονικές σχέσεις προσπαθούν να προσδιορίσουν τις ομοιότητες και τις διαφορές που δύο πηγές έχουν, στον ίδιο χρόνο. Στην περίπτωση της γραμμικής ή σύγχρονης εξέλιξης όλες οι πηγές υποβάλλουν αναφορές στον ίδιο χρόνο. Κατά συνέπεια τις περισσότερες φορές τα περιστατικά που περιγράφονται σε κάθε έγγραφο αναφέρονται στο χρόνο που το άρθρο δημοσιεύθηκε. Ακόμα, σε μερικές περιπτώσεις μπορεί να έχουμε στο κείμενο χρονικές εκφράσεις που τροποποιούν το χρόνο που το άρθρο δημοσιεύτηκε. Σε τέτοιες περιπτώσεις, πριν δημιουργηθεί μια συγχρονική σχέση, πρέπει να τοποθετηθεί αυτό το μήνυμα στον κατάλληλο χρονικό ορίζοντα.

Στην περίπτωση της μη γραμμικής ασύγχρονης εξέλιξης αυτό το φαινόμενο είναι κυρίαρχο. Κάθε πηγή υποβάλλει αναφορά σε άτακτα χρονικά διαστήματα, αναφέροντας ενδεχομένως τα περιστατικά που συνέβησαν πολύ πριν από τη δημοσίευση του άρθρου, και που μια άλλη πηγή μπορεί να έχει ήδη αναφέρει σε ένα άρθρο που δημοσιεύθηκε νωρίτερα. Σε αυτήν την περίπτωση δεν πρέπει να στηριχθούμε πλέον στη δημοσίευση ενός άρθρου, αλλά αντ' αυτού πρέπει να στηριχθούμε στη χρονική ετικέτα που τα μηνύματα έχουν και η οποία έχει τροποποιηθεί κατάλληλα σύμφωνα με τις χρονικές εκφράσεις που βρίσκονται στο κείμενο. Μόλις εκτελεσθεί αυτό, πρέπει έπειτα να δημιουργηθεί ένα χρονικό παράθυρο στο οποίο πρέπει να πλαισιώσουμε τα μηνύματα, και κατά συνέπεια τις σχέσεις ως συγχρονικές. Αυτό το χρονικό παράθυρο, ανάλογα με την περιοχή, μπορεί να ποικίλει από μερικές ώρες μέχρι μια ολόκληρη ημέρα.

Οι διαχρονικές σχέσεις, τώρα, προσπαθούν να συλλάβουν τις ομοιότητες και τις διαφορές, κατά τη διάρκεια του χρόνου, οι οποίες υπάρχουν για ένα γεγονός δεδομένου ότι περιγράφεται από την ίδια πηγή. Από αυτή την άποψη, οι διαχρονικές σχέσεις δεν παρουσιάζουν προβλήματα χρόνου, όπως οι συγχρονικές σχέσεις.

Παραδείγματα συγχρονικών σχέσεων μπορούν να είναι συμφωνία (agreement), διαφωνία (disagreement), γενίκευση (generalization) κ.λπ. Παραδείγματα διαχρονικών σχέσεων μπορούν να είναι θετική ή αρνητική βαθμολόγηση (positive or negative graduation), σταθερότητα (stability), συνέχεια (continuation), επανάληψη (repetition), κ.λπ.

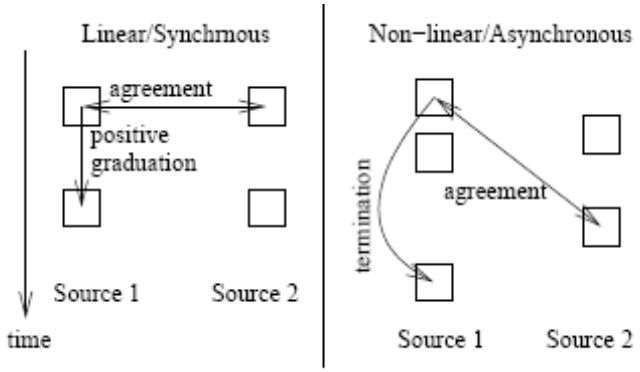
Με πιο επίσημους όρους, εάν αναπαρασταθεί μια σχέση r ως ένα ζευγάρι μηνυμάτων (m1,m2), όπου m1 και m2 είναι δύο μηνύματα, τότε μια σχέση θα είναι συγχρονική εάν και μόνο αν:
m1.time=m2.time και m1.source≠m2.source
και διαχρονική εάν και μόνο αν:
m1.time>m2.time και m1.source=m2.source

Προκειμένου να καθοριστεί μια σχέση σε μια περιοχή πρέπει καταρχάς αυτή να ονομαστεί και να περιγραφούν οι συνθήκες κάτω από τις οποίες θα ικανοποιείται.. Το όνομα της σχέσης είναι στην πραγματικότητα πραγματική πληροφορία, η οποία θα είναι δυνατόν να εκμεταλλευθεί κατάλληλα κατά τη διάρκεια της παραγωγής της περίληψης. Οι συνθήκες

κάτω από τους οποίες μια σχέση μεταξύ δύο μηνυμάτων ικανοποιείται αναπαριστούνται σε σχέση με τις τιμές των ορισμάτων τους, καθώς επίσης και σε σχέση με τον αντίστοιχο χρόνο τους και την πηγή τους.

Ας υποθέσουμε για παράδειγμα ότι έχουμε δύο ίδια μηνύματα. Εάν έχουν την ίδια χρονική ετικέτα, αλλά ανήκουν σε διαφορετικές πηγές, κατόπιν έχουμε π.χ. μια σχέση συμφωνίας (agreement) Εάν, από την άλλη μεριά, έχουν την ίδια πηγή αλλά κάποια χρονολογική απόσταση, τότε έχουμε π.χ. μια σχέση σταθερότητας (stability). Κατά συνέπεια χρίζει αναφοράς ότι εκτός από τα χαρακτηριστικά που τα ορίσματα ενός ζευγαριού μηνυμάτων (m1, m2) πρέπει να παρουσιάζουν, η πηγή και η χρονική απόσταση διαδραματίζουν επίσης έναν σημαντικό ρόλο προκειμένου εκείνο το ζευγάρι να χαρακτηριστεί από μια σχέση.

Στο παρακάτω σχήμα 6, παρουσιάζεται η διαφορά από την άποψη των συγχρονικών σχέσεων μεταξύ ενός πεδίου, το οποίο εξελίσσεται γραμμικά με το χρόνο και έχει μια σύγχρονη εκπομπή ειδησεογραφικών αναφορών από τις διάφορες πηγές και ενός πεδίου που εξελίσσεται μη-γραμμικά και έχει μια ασύγχρονη εκπομπή ειδησεογραφικών αναφορών από τις διάφορες πηγές Στην πρώτη περίπτωση, έχουμε δύο πανομοιότυπα μηνύματα που αφορούν στην απόδοση (βλέπε παραπάνω τον πίνακα 2) από δύο έγγραφα τα οποία έχουν δημοσιευτεί στον ίδιο χρόνο. Σύμφωνα με τις προδιαγραφές των συγχρονικών σχέσεων [4], υφίσταται μεταξύ τους μια σχέση συμφωνίας (agreement relation).Στην δεύτερη περίπτωση έχουμε δυο πανομοιότυπα μηνύματα που αφορούν στην διαπραγμάτευση (βλέπε παραπάνω τον πίνακα 2) από δύο έγγραφα τα οποία έχουν δημοσιευτεί σε διαφορετικό χρόνο. Στην περίπτωση αυτή τροποποιείται η χρονική ετικέτα σε ένα ή και στα δύο μηνύματα σε σχέση με την ημερομηνία δημοσίευσης της είδησης, λαμβάνοντας υπόψη την κανονικοποιημένη μορφή των χρονικών εκφράσεων και σε περίπτωση που αυτή αναφέρεται στην ίδια μέρα τότε έχουμε και πάλι μια συγχρονική σχέση, παρόλο που τα έγγραφα που περιέχουν τα μηνύματα δεν έχουν δημοσιευτεί την ίδια μέρα.



Σχήμα 6 – Παραδείγματα σύγχρονων και ασύγχρονων σχέσεων

Στο ίδιο σχήμα απεικονίζονται δύο διαχρονικές σχέσεις. Στην περίπτωση της γραμμικής εξέλιξης έχουμε δύο μηνύματα του ίδιου τύπου που αφορούν στην απόδοση και τα οποία έχουν τις μορφές performance (entity1, in_what1, time_span1, value1) και performance (entity2, in_what2, time_span2, value2) και ορίσματα τα ίδια, εκτός από το value, όπου έχουμε value1<value2. Σ' αυτή την περίπτωση, σύμφωνα με τις προδιαγραφές των σχέσεων του πεδίου [4] έχουμε μια διαχρονική σχέση θετικής βαθμολόγησης (positive graduation). Στην δεύτερη περίπτωση της μη-γραμμικής εξέλιξης έχουμε δύο διαφορετικά μηνύματα start (entity1, activity1) και end (entity2, activity2), όπου entity1=entity2 και activity1=activity2. Εδώ, σύμφωνα με τις προδιαγραφές έχουμε μια διαχρονική σχέση τερματισμού (termination).

Χρίζει αναφοράς ότι στην πρώτη περίπτωση έχουμε μια διαχρονική σχέση που ικανοποιείται μεταξύ μηνυμάτων του ίδιο τύπου, ενώ στη δεύτερη μεταξύ μηνυμάτων διαφορετικού τύπου.

Επίσης στην πρώτη περίπτωση, τα έγγραφα που περιέχουν τα μηνύματα έχουν απόσταση μονάδα, δηλαδή το ένα ακολουθεί αμέσως το άλλο, ενώ στη δεύτερη έχουν μεγαλύτερη απόσταση.

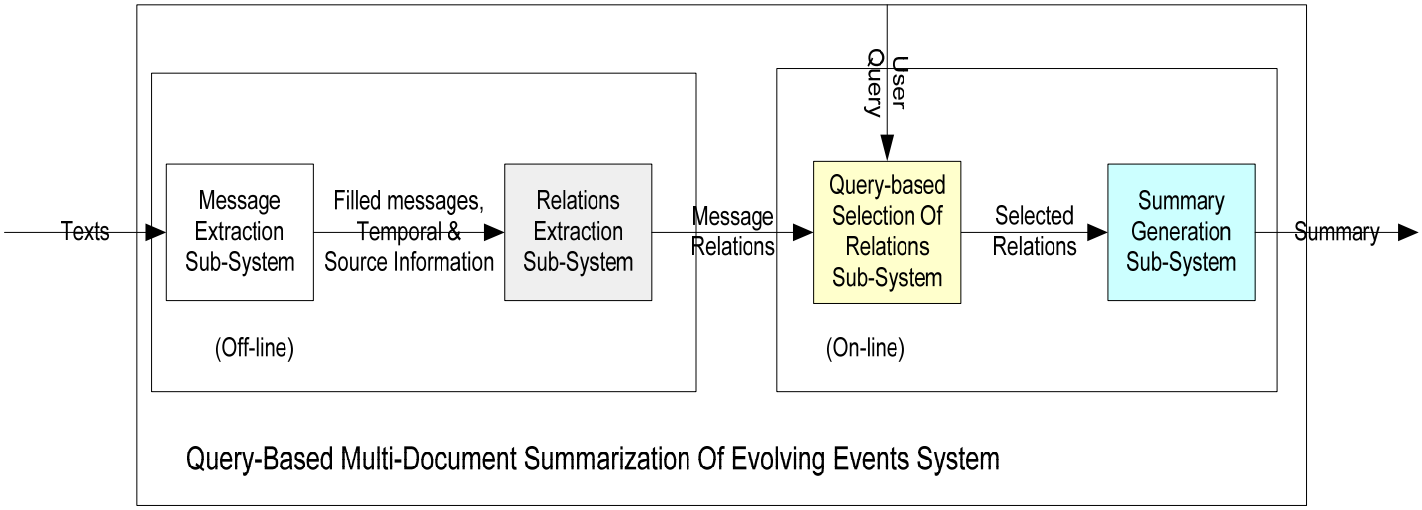
Μπορούν επίσης να υπάρχουν περιπτώσεις που ένα γεγονός περιγράφεται από μία μόνο πηγή. Δεδομένου ότι χρειαζόμαστε τουλάχιστον δύο μηνύματα από διαφορετικές πηγές προκειμένου να έχουμε μια σύγχρονη σχέση δε θα συνδεθεί το μήνυμα αυτό με κανένα άλλο, χάνοντας έτσι ίσως μια σημαντική πληροφορία που η πηγή αναφέρει. Αυτές οι περιπτώσεις μπορούν να αντιμετωπιστούν με την εισαγωγή μιας ελλειπτικής σχέσης (ellipsis relation).

Όπως προαναφέρθηκε, η μεταπτυχιακή αυτή εργασία ασχολείται με την υλοποίηση ενός μέρους ενός ευρύτερου συστήματος, το οποίο θα ωφελήσει στην αυτόματη παραγωγή περιλήψεων γεγονότων τα οποία εξελίσσονται στο χρόνο [4] [5] [6]. Παρακάτω, λοιπόν, προβαίνουμε σε μια συνοπτική περιγραφή αυτού του συστήματος:

Καταρχήν, το σύστημά μας ανήκει στην κατηγορία αυτών που βασίζονται σε μια ερώτηση (query-based), δεδομένου ότι η προκύπτουσα περίληψη είναι αποτέλεσμα μιας ερώτησης σε φυσική γλώσσα που ο χρήστης έχει θέσει. Αποτελείται από τέσσερα υποσυστήματα:

- Υποσύστημα Εξαγωγής Μηνυμάτων (Message Extraction Sub-system)
- Υποσύστημα Εξαγωγής Σχέσεων (Relation Extraction Sub-system)
- Υποσύστημα Επιλογής Σχέσεων Βασισμένο Σε Ερώτηση (Query-Based Selection Of Relation Sub-System)
- Υποσύστημα Παραγωγής Περίληψης (Summary Generation Sub-system)

Χρίζει αναφοράς ότι τα δύο πρώτα υποσυστήματα είναι εκτός σύνδεσης (off- line), ενώ τα δύο τελευταία είναι σε σύνδεση (on-line). Το εν λόγω σύστημα δέχεται ως είσοδο τα κείμενα μιας συλλογής, ενώ η έξοδός του είναι η περίληψη (βλέπε σχήμα 7).



Σχήμα 7 - Σύστημα περίληψης από πολλαπλά έγγραφα εξελισσόμενων γεγονότων που βασίζεται σε ερώτηση

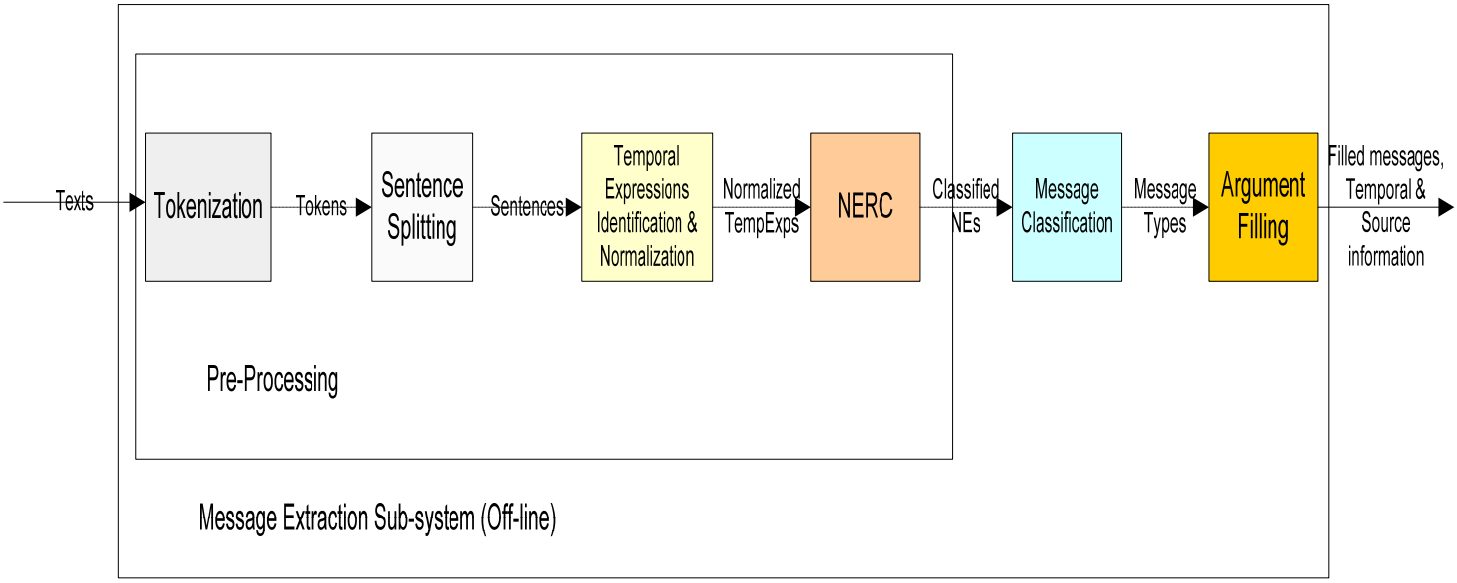
Όσον αφορά το υποσύστημα Εξαγωγής Μηνυμάτων (Message Extraction Sub-system) αυτό αποτελείται από τις εξής μονάδες:

- Διαχωρισμού λεκτικών μονάδων (Tokenization)
- Διαχωρισμού προτάσεων (Sentence Splitting)

- Αναγνώρισης και Κανονικοποίησης Χρονικών Εκφράσεων (Temporal Expressions Identification and Normalization)
- Αναγνώρισης και ταξινόμησης ονομάτων οντοτήτων (Named Entity Recognition and Classification – NERC)
- Ταξινόμησης Τύπων Μηνυμάτων (Message Type Classification)
- Γεμίσματος ορισμάτων (Argument Filling)

Χρίζει αναφοράς ότι οι τέσσερις πρώτες ανήκουν στο στάδιο προεπεξεργασίας (Pre-Processing) και πραγματοποιούνται εκτός σύνδεσης (off-line). Ως είσοδο το υποσύστημα αυτό δέχεται τα κείμενα μιας συλλογής, ενώ η έξοδος του είναι τα μηνύματα με συμπληρωμένα τα ορίσματά τους και συνδεδεμένα τόσο με χρονική πληροφορία στην οποία αναφέρονται, όσο και με πληροφορία σχετικά με την πηγή που ανήκουν. (βλέπε σχήμα 8).

Γίνεται σαφές ότι τα δικά μας αρθρώματα που υλοποιήθηκαν αφορούν στη μονάδα Αναγνώρισης και Κανονικοποίησης Χρονικών Εκφράσεων (Temporal Expressions Identification and Normalization) και στη μονάδα Εξαγωγής Ορισμάτων (Argument Extraction).



Σχήμα 8 - Υποσύστημα εξαγωγής μηνυμάτων

Παρακάτω ακολουθεί μια συνοπτική περιγραφή σχετικά με τί λειτουργίες επιτελούν το κάθε υποσύστημα και οι μονάδες τους:

Σε γενικές γραμμές, λοιπόν, στο υποσύστημα εξαγωγής μηνυμάτων, το οποίο όπως προαναφέρθηκε δέχεται ως είσοδο τα κείμενα, όσον αφορά στο προεπεξεργαστικό του στάδιο πραγματοποιείται καταρχάς διαχωρισμός λεκτικών μονάδων, με την οποία λαμβάνονται λεκτικές μονάδες, ακολούθως λαμβάνει χώρα διαχωρισμός προτάσεων από τον οποίο λαμβάνονται προτάσεις, έπειτα αναγνώριση και κανονικοποίηση χρονικών εκφράσεων από τα οποία λαμβάνονται οι κανονικοποιημένες χρονικές εκφράσεις και τέλος αναγνώριση και ταξινόμηση ονομάτων οντοτήτων σύμφωνα με την υπάρχουσα οντολογία από τα οποία λαμβάνονται ταξινομημένα ονόματα οντοτήτων. Χρίζει αναφοράς ότι η μονάδα αναγνώρισης και κανονικοποίησης χρονικών εκφράσεων την οποία και υλοποιήσαμε θα μπορούσε να εκτελείται πρώτη απ' όλες, δεδομένου ότι όπως θα διαπιστώσουμε παρακάτω είναι αυτόνομη και δεν απαιτείται κάποια προεπεξεργασία των κειμένων μας προκειμένου να λειτουργήσει.

Τα επόμενα δύο στάδια αποτελούν τον πυρήνα του υποσυστήματος εξαγωγής μηνυμάτων. Η μονάδα ταξινόμησης μηνυμάτων προσπαθεί να αναγνωρίσει τον τύπο του εξαγχθέντος μηνύματος. Ως αποτέλεσμα, αυτό που λαμβάνεται απ' εδώ είναι οι τύποι των μηνυμάτων.

Η επόμενη μονάδα στην οποία εμπλεκόμαστε και εν μέρει υλοποιήσαμε αφορά στη συμπλήρωση των τιμών των ορισμάτων των μηνυμάτων και συνεπώς παρέχει στην έξοδό της συμπληρωμένα μηνύματα συνδεδεμένα με χρονική πληροφορία καθώς και με την πηγή από την οποία προέρχονται.

Το υποσύστημα, τώρα, εξαγωγής σχέσεων, το οποίο δέχεται ως είσοδο τα μηνύματα με συμπληρωμένα τα ορίσματά τους, καθώς και πληροφορία που σχετίζεται με τον χρόνο και την πηγή είναι υπεύθυνο για την εξαγωγή των διμερών σχέσεων (συγχρονικών – διαχρονικών) μεταξύ των μηνυμάτων, βάσει της ικανοποίησης κάποιων συνθηκών. Η έξοδός του είναι φυσικά οι διμερείς σχέσεις που συνδέουν τα μηνύματα.

Εν συνεχεία, το υποσύστημα επιλογής σχέσεων βασισμένο σε ερώτηση είναι υπεύθυνο για την επιλογή κάποιων σχέσεων με βάση μια ερώτηση που τίθεται από έναν χρήστη. Στην έξοδό του παρέχονται οι επιλεγμένες σχέσεις.

Τέλος, το υποσύστημα παραγωγής της περίληψης είναι υπεύθυνο για την παραγωγή σε φυσική γλώσσα της περίληψης. Η έξοδός του ,όπως γίνεται αντιληπτό, είναι η περίληψη.

Συνολικά, είναι έκδηλο ότι το εν λόγω σύστημα είναι άρρηκτα συνδεδεμένο με τη γνώση του πεδίου (domain knowledge). Συνεπώς, προκείμενου κάποιος χρήστης να μπορέσει να παράγει κάποια περίληψη σε ένα νέο πεδίο με το εν λόγω σύστημα θα πρέπει να συγκεντρώσει μια συλλογή κειμένων που περιλαμβάνει γεγονότα που συσχετίζονται μεταξύ τους, να δημιουργήσει την κατάλληλη οντολογία του πεδίου (domain ontology), να καθορίσει τις προδιαγραφές των μηνυμάτων (τύπους, ορίσματα) και τέλος τις προδιαγραφές των σχέσεων (είδη, συνθήκες ικανοποίησης).

3 Χρονικές Εκφράσεις

3.1 Εισαγωγή

Ο πρώτος στόχος της μεταπτυχιακής εργασίας είναι η προγραμματιστική - αυτόματη επισημείωση χρονικών εκφράσεων (temporal expressions) που εμφανίζονται στα κείμενα μιας συλλογής. Ο ρόλος της επισημείωσης αυτής είναι η αναγνώριση και η κανονικοποίηση των χρονικών εκφράσεων. Όπως προαναφέρθηκε, αυτό είναι ιδιαίτερα αναγκαίο και χρήσιμο, λόγω του ότι η χρονική πληροφορία με την οποία συνδέεται ένα μήνυμα καθορίζεται από τις χρονικές εκφράσεις που περιλαμβάνονται σ' αυτό σε σχέση με την ημερομηνία της είδησης.

Υπάρχουν δύο προσεγγίσεις τις οποίες θα μπορούσαμε να ακολουθήσουμε. Η μία βασίζεται σε κανόνες (rule-based) και η άλλη σε μηχανική μάθηση (machine learning-based). Η προσέγγιση που βασίζεται σε κανόνες χωρίζεται σε εσωτερική (internal) και εξωτερική (external) καθώς και σε συνδυασμό τους. Η μεν πρώτη λαμβάνει υπόψη μόνο το περιεχόμενο των εκφράσεων, η δε δεύτερη λαμβάνει υπόψη μόνο το συγκείμενο (context), ενώ η τρίτη και τα δύο. Εμείς επιλέξαμε την προσέγγιση που βασίζεται σε κανόνες στην εσωτερική μορφή της, δηλαδή δε λάβαμε υπόψη καθόλου το συγκείμενο. Στην ίδια απόφαση καταλήξαμε και όσον αφορά στην κανονικοποίηση, καθώς οι περιπτώσεις όπου μας χρειαζόταν το συγκείμενο προκειμένου να πραγματοποιηθεί σωστή κανονικοποίηση ήταν πολύ λιγότερες από τις περιπτώσεις στις οποίες αυτό ήταν περιττό.

Δεν χρησιμοποιήθηκε μηχανική μάθηση για το λόγο ότι δεν υπήρχαν αρκετά παραδείγματα εκπαίδευσης για κάθε κατηγορία χρονικών εκφράσεων, ενώ υπήρχαν και πολλές υποπεριπτώσεις σε κάθε κατηγορία χρονικών εκφράσεων (βλέπε παρακάτω τις κατηγορίες).

Η μεθοδολογία που ακολουθήθηκε περιλαμβάνει τα ακόλουθα βήματα:

- Χειρωνακτική επισημείωση των χρονικών εκφράσεων στα έγγραφα της συλλογής μας και οργάνωση των χρονικών εκφράσεων σε κατηγορίες
- Συγγραφή κανόνων ανά κατηγορία με χρήση κανονικών εκφράσεων. Επίσης, για κάθε κατηγορία επινοήθηκαν και κάποιοι τρόποι κανονικοποίησης.
- Αναγνώριση και κανονικοποίηση των χρονικών εκφράσεων με εφαρμογή των κανόνων και των τρόπων κανονικοποίησης αντίστοιχα. Αξιολόγηση των αποτελεσμάτων

Τα βήματα αυτά περιγράφονται στις ενότητες που ακολουθούν.

3.2 Δημιουργία σώματος κειμένων αξιολόγησης

Πραγματοποιήθηκε χειρωνακτική επισημείωση των χρονικών εκφράσεων στα 164 έγγραφα της συλλογής κειμένων αξιολόγησης (βλέπε στο Παράρτημα για τις προδιαγραφές της), λόγω του ότι - όπως προαναφέρθηκε - στο τέλος θα χρειαζόταν να πραγματοποιηθεί σύγκριση μεταξύ της επισημείωσης των χρονικών εκφράσεων και της αντίστοιχης κανονικοποίησής τους που θα πραγματοποιούταν χειρωνακτικά και της επισημείωσης των χρονικών εκφράσεων και της αντίστοιχης κανονικοποίησής τους που θα πραγματοποιούταν υπολογιστικά – προγραμματιστικά. Τα κριτήρια με τα οποία επιλέχθηκαν αφορούν στην επισημείωση αυτών των χρονικών εκφράσεων που η κανονικοποίησή τους να μην καθορίζεται από το συγκείμενο - συμφραζόμενα.

Χρίζει αναφοράς ότι δεν επισημειώθηκαν χειρωνακτικά μόνο οι χρονικές εκφράσεις που περιλαμβάνονταν στα μηνύματα, λόγω του ότι θελήσαμε να κατασκευάσουμε ένα όσο το

δυνατόν γενικότερο εργαλείο. Βέβαια αυτό δημιουργεί κάποιο πρόβλημα στην δημιουργία σαφής εικόνας ορθών αποτελεσμάτων όσον αφορά στην αναγνώριση και στην κανονικοποίηση των χρονικών εκφράσεων που περιλαμβάνονται στα μηνύματα, πράγμα και το οποίο αποτέλεσε και το κίνητρο για τη δημιουργία του προκείμενου αρθρώματος. Βέβαια, όπως θα δούμε παρακάτω τα αποτελέσματα που προκύπτουν είναι αρκετά ικανοποιητικά, πράγμα το οποίο μας επιτρέπει κατά κάποιο τρόπο να βγάλουμε και κάποια συμπεράσματα και για το υποσύνολο των χρονικών εκφράσεων της εν λόγω περίπτωσης.

Επίσης, δεν επισημειώθηκαν χειρωνακτικά όλες οι χρονικές εκφράσεις που περιλαμβάνονται στα μηνύματα και επομένως δεν έχει υλοποιηθεί ούτε η αναγνώρισή τους άλλα ούτε και η κανονικοποίησή τους, (π.χ. χρονικές εκφράσεις όπως «στη συνέχεια», «νωρίτερα» προκειμένου να κανονικοποιηθούν θα έπρεπε να ληφθεί υπόψη το συγκείμενο»).

3.2.1 Κανονικοποίηση χρονικών εκφράσεων

Ο τρόπος κανονικοποίησής τους στηρίχτηκε σε μεγάλο βαθμό στο έργο Mitos του οποίου ανάδοχος φορέας ήταν το Εργαστήριο Τεχνολογίας Γνώσεων & Λογισμικού του Ινστιτούτου Πληροφορικής & Τηλεπικοινωνιών του Ε.Κ.Ε.Φ.Ε. «Δημόκριτος» και αφορούσε εν ολίγοις την υλοποίηση ενός συστήματος Αναζήτησης και Εξόρυξης Πληροφορίας με εφαρμογή σε Χρηματοοικονομικές Ειδήσεις (για περισσότερες πληροφορίες βλέπε http://iit.demokritos.gr/skel/mitos).

Η κανονικοποιημένη αναπαράσταση των χρονικών εκφράσεων μπορεί να λάβει τις εξής μορφές [11]:

- **(ηη/μμ/χχχχ@ωω:λλ, ηη/μμ/χχχχ@ωω:λλ)** που σημαίνει «από ημερομηνία και ώρα μέχρι ημερομηνία και ώρα»
- **(ηη/μμ/χχχχ@ωω:λλ, inf)** που έχει τη σημασία «από ημερομηνία και ώρα μέχρι άγνωστο χρονικό σημείο»
- **(inf, ηη/μμ/χχχχ@ωω:λλ)** με τη σημασία «από άγνωστο χρονικό σημείο μέχρι ημερομηνία και ώρα»

Σε αυτές τις αναπαραστάσεις οι παρενθέσεις χρησιμοποιούνται για το συμβολισμό συγκεκριμένης χρονικής διάρκειας, το σύμβολο "@" για το διαχωρισμό της ημερομηνίας από την ώρα, το κόμμα για το χωρισμό έναρξης και λήξης μιας συγκεκριμένης χρονικής περιόδου.

Γενικά η αναπαράσταση της κανονικοποίησης των χρονικών εκφράσεων (ΧΕ) μπορεί να αποδοθεί με τους παρακάτω δύο κανόνες [11]:

ΧΕ → (ΑΚΡΟ, ΑΚΡΟ) | (inf, ΑΚΡΟ) | (ΑΚΡΟ, inf)
ΑΚΡΟ → ηη/μμ/χχχχ@ωω:λλ

Ο τρόπος κανονικοποίησης σχετίζεται σε ορισμένες περιπτώσεις με την ημερομηνία δημοσίευσης της είδησης, ενώ σε κάποιες άλλες όχι.

Για παράδειγμα, όταν έχουμε αναφορά σε συγκεκριμένη μέρα, π.χ. «την Κυριακή», τότε ως ημέρα τοποθετείται στην κανονικοποίηση η ημέρα της δημοσίευσης της είδησης, αν η ημερομηνία δημοσίευσης της είδησης ταυτίζεται με την ημερομηνία που αντιστοιχεί στη χρονική έκφραση που αναφέρεται στο κείμενό μας (στην περίπτωσή μας «την Κυριακή»), διαφορετικά τοποθετείται το αποτέλεσμα που προκύπτει αν αφαιρέσουμε από την ημέρα που αντιστοιχεί στην ημερομηνία δημοσίευσης της είδησης τόσες μέρες όσες είναι η διαφορά σε ημέρες μεταξύ της ημέρας που αντιστοιχεί στην ημερομηνία δημοσίευσης της είδησης και της αναφερθείσας ημέρας στο κείμενο. Ο μήνας και το έτος λαμβάνονται από την

ημερομηνία δημοσίευσης της είδησης και για την ώρα θεωρούμε ότι μια μέρα ξεκινάει στις 00:00 και τελειώνει στις 23:59.

Έστω ότι η ημερομηνία δημοσίευσης της είδησης είναι η 6/2/2000. Τότε η χρονική έκφραση «την Κυριακή» θα κανονικοποιηθεί ως εξής: **την Κυριακή** (06/02/2000@00:00, 06/02/2000@23:59), όπου η ημερομηνία δημοσίευσης της είδησης ταυτίζεται με την ημερομηνία που αντιστοιχεί στην ημέρα Κυριακή.

Στην περίπτωση που έχουμε αναφορά σε συγκεκριμένο έτος, τότε δεν λαμβάνεται υπόψη η ημερομηνία της είδησης και ως έναρξη και λήξη του έτους θεωρείται η πρώτη και η τελευταία ημέρα του αντίστοιχα. Για παράδειγμα η χρονική έκφραση «το 1999» κανονικοποιείται ως εξής: **το 1999** (01/01/1999@00:00, 31/12/1999@23:59).

Στο Παράρτημα παρακάτω δίνονται κάποια παραδείγματα χρονικών εκφράσεων καθώς και ο τρόπος κανονικοποίησής τους.

3.3 Συγγραφή κανόνων

3.3.1 Κατηγορίες χρονικών εκφράσεων

Η κατηγοριοποίηση των χρονικών εκφράσεων στηρίχτηκε σε μεγάλο βαθμό στο έργο Mitos καθώς και στη μελέτη της συλλογής κειμένων μας.

Οι κατηγορίες χρονικών εκφράσεων που πρέπει να αναγνωρίζονται, να σημειώνεται η έκτασή τους μέσα στο κείμενο και να κανονικοποιούνται είναι οι εξής:

- ✓ Φράσεις που αναφέρονται σε ώρες, π.χ. «στις 12:30», «στις 10 το πρωί» κ.ο.κ.
- ✓ Φράσεις που αναφέρονται σε ημέρες, όπως «την Κυριακή», «από την Τετάρτη», κ.ο.κ.
- ✓ Φράσεις που αναφέρονται σε χρονικό διάστημα ημέρας, π.χ. «το πρωί», «τα ξημερώματα» κ.ο.κ.
- ✓ Φράσεις που αναφέρονται σε εβδομάδες, π.χ. «την περασμένη εβδομάδα», την επόμενη εβδομάδα» κ.ο.κ.
- ✓ Φράσεις που αναφέρονται σε μήνες, π.χ. «τον περασμένο μήνα», «τον επόμενο μήνα» κ.ο.κ.
- ✓ Φράσεις που αναφέρονται σε έτη, π.χ. «τον περασμένο χρόνο», «τον επόμενο χρόνο» κ.ο.κ.
- ✓ Φράσεις που αναφέρονται σε ημερομηνίες, π.χ. «στις 20 Ιουλίου», «6/2/2000» κ.ο.κ.
- ✓ Χρονικά επιρρήματα, π.χ. «σήμερα», «αύριο», «τώρα» κ.ο.κ.
- ✓ Επιθετικοί προσδιορισμοί που δηλώνουν χρόνο όπως «σημερινός», «χθεσινός» κ.ο.κ.
- ✓ Ανάμεικτες χρονικές εκφράσεις, όπως «αυτήν την στιγμή», «για την ώρα» κ.ο.κ.
- ✓ Συνδυασμοί των παραπάνω, όπως π.χ. «Κυριακή 6 Φεβρουαρίου 2000 – 14:30», «τα ξημερώματα της Δευτέρας», «σήμερα το μεσημέρι» κ.ο.κ.

3.3.2 Κανόνες ανά κατηγορία

Όπως προαναφέρθηκε δημιουργήθηκαν κανόνες, οι οποίοι λαμβάνουν υπόψη μόνο το περιεχόμενο των χρονικών εκφράσεων. Οι κανόνες αυτοί αναπαρίστανται με κανονικές εκφράσεις. Αυτές, τώρα, δημιουργήθηκαν με τέτοιο τρόπο, ώστε να αναγνωρίζονται λαθεμένα όσο το δυνατόν λιγότερες χρονικές εκφράσεις - οι οποίες δεν είναι στην πραγματικότητα - και μια μορφή (pattern) κανονικής έκφρασης να αντιστοιχεί σε όσο το δυνατόν περισσότερες χρονικές εκφράσεις. Με άλλα λόγια στο σύνολο των χρονικών

εκφράσεων που λήφθηκαν από τη συλλογή μας, πρώτα πραγματοποιήθηκε ομαδοποίηση αυτών με βάση κάποια κοινά τους στοιχεία και οργάνωση αυτών σε κατηγορίες και αφετέρου κατασκευάστηκαν οι μορφές (patterns) των κανονικών εκφράσεων. Εν τέλει, δημιουργήθηκε μια μορφή (pattern) κανονικών εκφράσεων για το σύνολο των χρονικών εκφράσεων.

Σε γενικές γραμμές, οι κανονικές εκφράσεις που κατασκευάστηκαν συντάσσονται με χρήση κάποιων συμβόλων με ειδικό όνομα που ορίζει η γλώσσα προγραμματισμού Java και τα οποία σύμβολα εντάσσονται σε ένα γενικότερο πλαίσιο αναπαράστασης κανονικών εκφράσεων. Για παράδειγμα προκειμένου να αναγνωριστεί κάθε χρονική έκφραση του τύπου «23:30» που αναπαριστά ώρα δημιουργήθηκε η κανονική έκφραση "\d{2}\.d{2}" που σημαίνει «ακριβώς 2 ψηφία που ακολουθούνται από τελεία που ακολουθείται από ακριβώς 2 ψηφία». Το πρόβλημα που αντιμετωπίζουμε εδώ είναι ότι σε περίπτωση που σε κείμενο της συλλογής μας υπάρχει για παράδειγμα κάποιο ποσό στο οποίο περιλαμβάνεται τελεία ή τελείες, τότε θα αναγνωριστεί λαθεμένα ως χρονική έκφραση. Άλλο ένα παράδειγμα που θα μπορούσαμε να αναφέρουμε αφορά στην αναγνώριση της χρονικής έκφρασης «την Κυριακή». Εδώ δημιουργήθηκε η κανονική έκφραση «[τΤ]p{InGreek}+ Κυριακ.» που σημαίνει «Είτε πεζός είτε κεφαλαίος ελληνικός χαρακτήρας «ταφ» που ακολουθείται διαδοχικά από τον συμβολισμό "\p{InGreek}+" που σημαίνει οποιοσδήποτε ελληνικός χαρακτήρας μία η περισσότερες φορές, το κενό, τους ελληνικούς χαρακτήρες Κ, υ, ρ, ι, α, κ που ακολουθούνται από το σύμβολο "." που σημαίνει οποιοσδήποτε χαρακτήρας». Αυτό επινοήθηκε αρχικά προκείμενου ο ελληνικός τόνος να μη δημιουργεί πρόβλημα αναγνώρισης (στο συγκεκριμένο παράδειγμα τοποθετείται στη θέση του ελληνικού χαρακτήρα «η» όπου τονίζεται, αν και εν τέλει διαπιστώθηκε πως ο ελληνικός τόνος δε δημιουργεί πρόβλημα. Επίσης για να αποφύγουμε προβλήματα σχετικά με το τελικό ν τοποθετήθηκε ο συμβολισμός "\p{InGreek}+" μετά τον χαρακτήρα «ταφ». Στο συγκεκριμένο παράδειγμα θα μπορούσαμε να έχουμε δημιουργήσει την κανονική έκφραση «[τΤ]p{InGreek}+ Κ\p{InGreek}+» που σημαίνει «Είτε πεζός είτε κεφαλαίος ελληνικός χαρακτήρας «ταφ» που ακολουθείται από οποιοδήποτε ελληνικό χαρακτήρα μία ή περισσότερες φορές, το κενό, τον ελληνικό κεφαλαίο χαρακτήρα Κ που ακολουθείται από οποιοδήποτε ελληνικό χαρακτήρα μία ή περισσότερες φορές». Η αναπαράσταση αυτή δεν επιλέχθηκε λόγω του ότι έτσι θα αναγνωριζόταν και κάθε άλλη ελληνική λέξη της οποίας θα προηγούταν για παράδειγμα το άρθρο «την» και θα ξεκινούσε με ελληνικό κεφαλαίο γράμμα Κ. Για την αναγνώριση όμως της χρονικής έκφρασης «την περασμένη Κυριακή» δημιουργήθηκε η κανονική έκφραση «[τΤ]\p{InGreek}+ π\p{InGreek}+ Κυριακ.», διότι στη συγκεκριμένη περίπτωση είναι πολύ δύσκολο να μεσολαβεί διαφορετική λέξη από την «περασμένη» μεταξύ για παράδειγμα των λέξεων «την» και «Κυριακή» ή «Την» και «Κυριακή». Σημειώνουμε εδώ ότι αντί για το ειδικό σύμβολο "." που όπως αναφέρθηκε σημαίνει οποιοσδήποτε χαρακτήρας θα μπορούσε να έχει τοποθετηθεί ο συμβολισμός "\p{InGreek}+", ώστε να αναγνωρίζονται και χρονικές εκφράσεις όπως για παράδειγμα «τις Κυριακές». Χρίζει αναφοράς ότι μεταξύ των κανονικών εκφράσεων που δημιουργήθηκαν τοποθετήθηκε το ειδικό σύμβολο "|" που λειτουργεί στη σύνταξη των κανονικών εκφράσεων όπως το λογικό "OR".

3.4 Εφαρμογή κανόνων

3.4.1 Μεθοδολογία Επισημείωσης και Κανονικοποίησης Χρονικών Εκφράσεων

Η μεθοδολογία της αναγνώρισης χρονικών εκφράσεων στα κείμενα (documents) μιας συλλογής κειμένων (corpus) και της μετέπειτα κανονικοποίησής τους στηρίζεται στο συνδυασμό χρήσης του Java Application Programming Interface (API) [9] της πλατφόρμας επεξεργασίας φυσικής γλώσσας Ellogon [7] [8], της τεχνολογίας Κανονικών Εκφράσεων της γλώσσας προγραμματισμού Java και γενικότερα του API της Java.

Εν συντομία τα βήματα της προσέγγισης που επιλέχθηκε και ονομάζεται SystemAegeanNcsrTempExp είναι τα εξής:

Καταρχάς λαμβάνεται κάθε φορά το κείμενο της συλλογής σε μορφή ακατέργαστων δεδομένων (raw data). Αυτό μετατρέπεται σε συμβολοσειρά (String) και εν συνεχεία σε ακολουθία χαρακτήρων (Charsequence).

Αφετέρου δημιουργούνται οι μορφές (patterns) των κανονικών εκφράσεων για τις χρονικές εκφράσεις σε μορφή συμβολοσειράς.

Έπειτα δημιουργείται ένας Matcher που αντιστοιχίζει γενικά τις κανονικές εκφράσεις με το κείμενο της συλλογής μας. Για όσο τώρα πραγματοποιείται ταίριασμα της μορφής (pattern) των κανονικών εκφράσεων που δημιουργήσαμε με κάποια επόμενη χρονική έκφραση (υποτίθεται ότι έχει προηγηθεί ένα ταίριασμα), δημιουργείται η έκταση (span) της χρονικής έκφρασης που ταιριάστηκε. Εάν τώρα αυτή ταιριάζει με κάποια μορφή (pattern) κανονικής έκφρασης ή κανονικών εκφράσεων (δημιουργήθηκαν μορφές (patterns) κανονικής έκφρασης ή κανονικών εκφράσεων, ούτως ώστε να μην υλοποιηθεί ξεχωριστή κανονικοποίηση για χρονικές εκφράσεις που κανονικοποιούνται με τον ίδιο τρόπο), τότε διαβάζεται (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε στο Παράρτημα την ενότητα «Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία»)) το όνομα του κειμένου της συλλογής και δημιουργείται (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε στο Παράρτημα την ενότητα «Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία»)) τέτοια μορφή (pattern) κανονικών εκφράσεων, ώστε να μπορεί να ταιριαστεί η ημερομηνία και η ώρα δημοσίευσης της είδησης του κειμένου που περιέχονται στο όνομα του αρχείου (\d{4}\d{2}\d{2}|\d{4}\d{2}\d{2}\d{2}\d{2}) (το d σημαίνει ψηφίο (digit), ο αριθμός μέσα στα άγκιστρα δηλώνει το ακριβές πλήθος των ψηφίων, ενώ το σύμβολο "|" εκφράζει το λογικό OR).

Αν ταιριαστούν, τότε (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε στο Παράρτημα την ενότητα «Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία»)) τότε λαμβάνονται όσες πληροφορίες χρειάζονται κάθε φορά (π.χ. μέρα, μήνας, έτος, ώρα) (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε στο Παράρτημα την ενότητα «Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία»)) και δημιουργείται επισημείωση τύπου SystemAegeanNcsrTempExp με ιδιότητα τύπου timeperiod με τιμή την κατάλληλη και αντίστοιχη κανονικοποιημένη τιμή της. Από όλη τη διαδικασία κανονικοποίησης θα πρέπει να τονίσουμε ότι προκειμένου να βρεθεί για παράδειγμα σε ποια ημέρα αντιστοιχεί μια ημερομηνία δημοσίευσης της είδησης χρησιμοποιείται το Γρηγοριανό Ημερολόγιο (Gregorian Calendar), ενώ για να ληφθούν οι κατάλληλες πληροφορίες από τις χρονικές εκφράσεις που αναγνωρίστηκαν χρησιμοποιείται κατάλληλος χειρισμός συμβολοσειρών (Στο Παράρτημα παρουσιάζονται πιο συγκεκριμένα τα βήματα της προκείμενης προσέγγισης).

Τέλος, είναι εξίσου σημαντικό να τονιστεί ότι προκειμένου να εκτελεστεί το συγκεκριμένο άρθρωμα που αναπτύχθηκε για τις χρονικές εκφράσεις, δεν προαπαιτείται η εκτέλεση κάποιου ή κάποιων άλλων αρθρωμάτων (π.χ. κάποιου διαχωριστή λεκτικών μονάδων (Tokenizer)) (με άλλα λόγια δε χρειάζεται η συλλογή μας να είναι εκ των προτέρων επισημειώμενη με κάποιου τύπου επισημειώσεις), δεδομένου ότι σχεδιάστηκε και προγραμματίστηκε, ώστε να λειτουργεί αυτόνομα. Για παράδειγμα, εάν εκτός από την αναγνώριση και την κανονικοποίηση των χρονικών εκφράσεων χρειαζόταν να προσδιορίζονται και τα αναγνωριστικά (ids) των λεκτικών μονάδων (tokens), από τα οποία αποτελείται μια χρονική έκφραση, τότε θα έπρεπε εκ των προτέρων να έχει εκτελεστεί ένας

διαχωριστής λεκτικών μονάδων (tokenizer), ο οποίος θα διασπούσε κάθε κείμενο μιας συλλογής σε λεκτικές μονάδες και θα απέδιδε στην κάθε μία ένα αναγνωριστικό.

3.5 Πειράματα – Αποτελέσματα

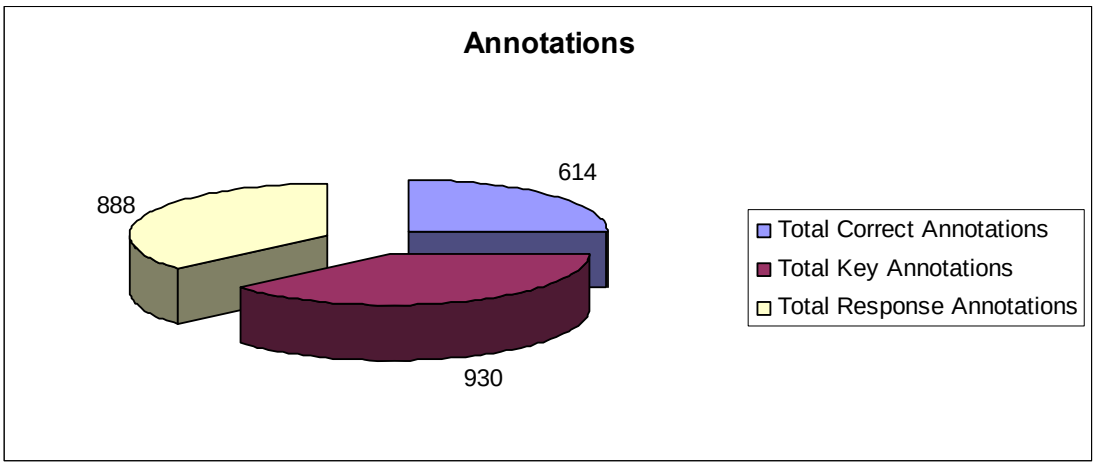
3.5.1 Αξιολόγηση των αποτελεσμάτων του συστήματος αναγνώρισης και κανονικοποίησης χρονικών εκφράσεων

Τα πειράματα πραγματοποιήθηκαν πάνω στη συλλογή gold_corpus_final, η οποία συμπεριλαμβάνεται στο συνοδευτικό cd της μεταπτυχιακής. Συγκρίθηκε η κανονικοποίηση που πραγματοποιήθηκε χειρωνακτικά με την κανονικοποίηση που πραγματοποιήθηκε αυτόματα από το σύστημα. Τα αποτελέσματα τα οποία προέκυψαν απεικονίζονται στον πίνακα 3 (συγκρίνεται η τιμή της ιδιότητας timeperiod μεταξύ της χειρωνακτικής επισημείωσης AegeanNcsrTempExp και της επισημείωσης SystemAegeanNcsrTempExp) του συστήματος (βλέπε παρακάτω στο Παράρτημα για τη διαδικασία της αξιολόγησης στην πλατφόρμα επεξεργασίας φυσικής γλώσσας Ellogon καθώς και για τη σημασία των μετρικών στα αποτελέσματα που προκύπτουν):

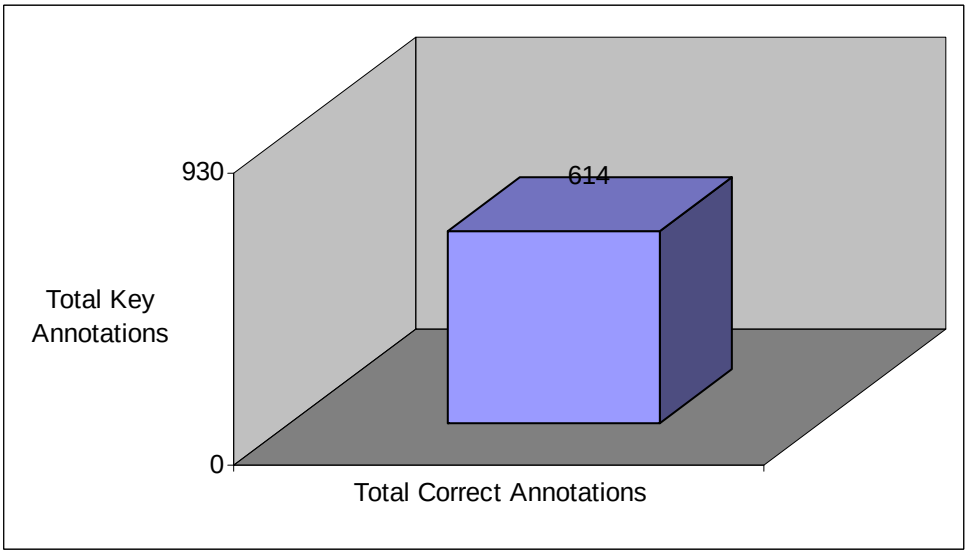
Compare Attribute	Total Correct Annotations	Total Key Annotations	Total Response Annotations	Precision	Recall	F-measure
timeperiod	614	930	888	0,69	0,66	0,67

Πίνακας 3 – Αποτελέσματα κανονικοποίησης χρονικών εκφράσεων

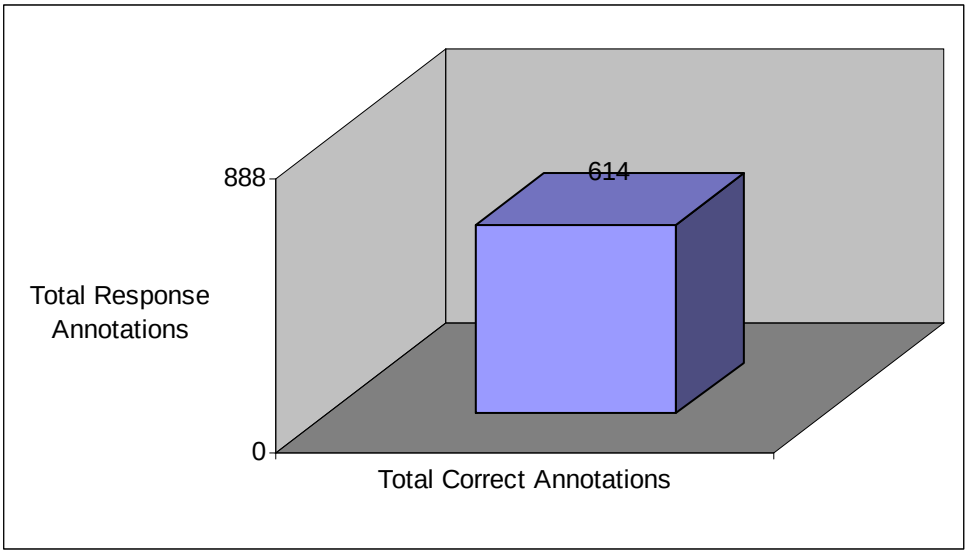
Στα παρακάτω διαγράμματα απεικονίζονται τα παραπάνω πειραματικά αποτελέσματα που ελήφθησαν:



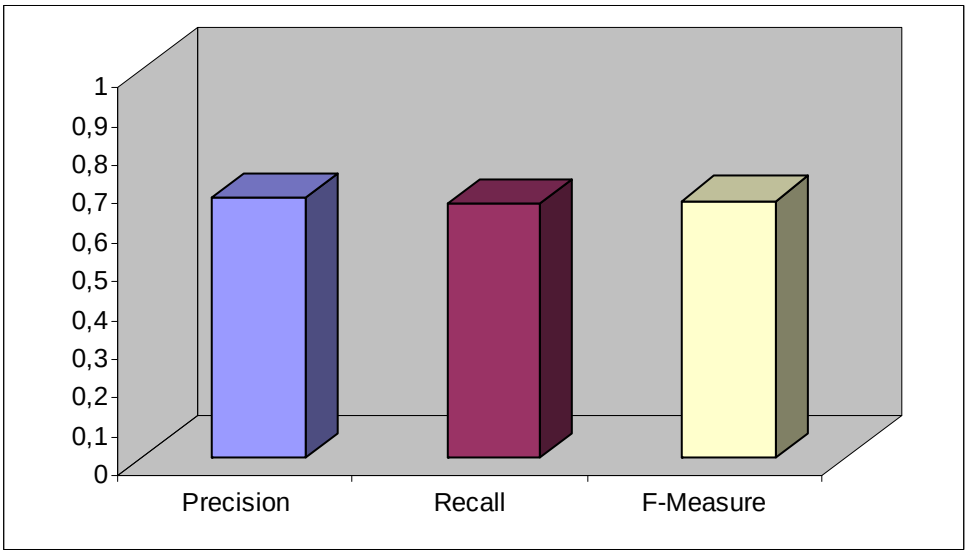
Διάγραμμα 1 – Total Correct Annotations, Total Key Annotations και Total Response Annotations



Διάγραμμα 2 – Total Correct Annotations σε σχέση με τα Total Key Annotations



Διάγραμμα 3 – Total Correct Annotations σε σχέση με τα Total Response Annotations



Διάγραμμα 4 – Precision, Recall και F-Measure

Τα παραπάνω αποτελέσματα είναι ικανοποιητικά και αναμενόμενα αν λάβουμε υπόψη μας τα εξής (για τη σημασία των παραπάνω μετρικών βλέπε στο Παράρτημα την ενότητα «Χρήση του Ellogon»):

- Στην κανονικοποίηση που πραγματοποιείται από το σύστημα δε λαμβάνεται καθόλου υπόψη το συγκείμενο, γι' αυτό και κάποιες χρονικές εκφράσεις κανονικοποιούνται λανθασμένα. Για παράδειγμα σε περίπτωση που η χρονική έκφραση «την Κυριακή» αναφέρεται στην επόμενη Κυριακή, το σύστημα την κανονικοποιεί λαθεμένα με την ημερομηνία δημοσίευσης του άρθρου αν αυτή αντιστοιχεί στην ημέρα Κυριακή. Επιπλέον, όταν έχουμε τη χρονική έκφραση «τώρα» αυτή κανονικοποιείται πάντα όπως η χρονική έκφραση «σήμερα» (βλέπε στο Παράρτημα την ενότητα «Τρόποι και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία»), ενώ χειρωνακτικά μπορεί να έχει κανονικοποιηθεί με την ημερομηνία και την ώρα δημοσίευσης του άρθρου (ακόμα και ο άνθρωπος πολλές φορές δυσκολεύεται να ξεχωρίσει αν το χρονικό επίρρημα «τώρα» αναφέρεται στην στιγμή που μιλάμε ή γενικότερα στο παρόν

- Δε λαμβάνεται υπόψη προγραμματιστικά η αρχή και το τέλος των ημερών, η αρχή και το τέλος των μηνών καθώς και η αρχή και το τέλος των χρόνων (για παράδειγμα αν η ημερομηνία δημοσίευσης του άρθρου είναι η Δευτέρα 02/05/2005 και πρέπει να πάμε για παράδειγμα 5 ημέρες πίσω, γιατί μέσα στο άρθρο αναφέρεται η ημέρα «Πέμπτη», τότε το σύστημα επειδή δεν έχει προβλεφτεί αυτή η περίπτωση θα κανονικοποιήσει λαθεμένα την προκείμενη χρονική έκφραση ως -3/05/2005 και όχι ως 28/04/2004 που είναι και το σωστό)
- Βρίσκονται λαθεμένα επιπλέον αριθμοί (για παράδειγμα μπορεί να βρεθεί το «50.00» το οποίο είναι μέρος ενός ποσού ή ενός αριθμού, λόγω του γεγονότος ότι έχει κατασκευαστεί μια κανονική έκφραση, ώστε να αναγνωρίζονται χρονικές εκφράσεις της μορφής 20.05 που αναφέρονται προφανώς σε ώρα).Αυτό έχει ως συνέπεια και τη λαθεμένη κανονικοποίησή τους
- Δεν έχουν υλοποιηθεί η αναγνώριση και συνεπώς και η κανονικοποίηση όλων των μορφών των χρονικών εκφράσεων που επισημειώθηκαν χειρωνακτικά
- Αξίζει, τέλος, να σημειωθεί ότι κάποιες χρονικές εκφράσεις που επισημειώθηκαν χειρωνακτικά φέρουν κάποια λάθη στην κανονικοποίηση τους, ενώ το σύστημα τις βρίσκει σωστά (λόγω του ότι κατά τη διάρκεια της υλοποίησης του αρθρώματος άλλαξαν οι προδιαγραφές τους), ενώ κάποιες άλλες που βρίσκει το σύστημα παραλείφθηκαν ακουσίως να επισημειωθούν χειρωνακτικά

Ένα τελευταίο πείραμα που πραγματοποιήθηκε ήταν η σύγκριση των χρονικών εκφράσεων που επισημειώθηκαν χειρωνακτικά με αυτές που επισημειώθηκαν αυτόματα από το σύστημα (συγκρίνεται η χειρωνακτική επισημείωση AegeanNcsrTempExp με την επισημείωση SystemAegeanNcsrTempExp του συστήματος. Τα αποτελέσματα τα οποία προέκυψαν απεικονίζονται παρακάτω στον πίνακα 4:

Total Correct Annotations	Total Key Annotations	Total Response Annotations	Precision	Recall	F-measure
680	930	888	0,76	0,73	0,74

Πίνακας 4 – Αποτελέσματα αναγνώρισης χρονικών εκφράσεων

Τα αποτελέσματα κρίνονται αρκετά ικανοποιητικά και θα μπορούν να δικαιολογηθούν ως εξής:

- Κάποιες χρονικές εκφράσεις απ' αυτές που επισημειώθηκαν χειρωνακτικά δεν αναγνωρίζονται καθόλου από το σύστημα (γιατί δεν έχουν κατασκευαστεί κανόνες γι' αυτές) (π.χ. η χρονική έκφραση «το Σαββατοκύριακο»)
- Κάποιες χρονικές εκφράσεις που δεν έχουν επισημειωθεί χειρωνακτικά αναγνωρίζονται επιπλέον από το σύστημα (π.χ. το 80.00 το οποίο το αναγνωρίζει ως χρονική έκφραση που αντιπροσωπεύει ώρα, ενώ στην πραγματικότητα αποτελεί μέρος του ποσού 80.000)

3.5.2 Συγκριτική αξιολόγηση με το σύστημα αναγνώρισης και κανονικοποίησης χρονικών εκφράσεων του έργου (project) Mitos

Κρίθηκε σκόπιμο να πραγματοποιηθεί σύγκριση ανάμεσα στις χρονικές εκφράσεις που επισημειώθηκαν χειρωνακτικά με αυτές που επισημειώθηκαν αυτόματα από το σύστημα αναγνώρισης χρονικών εκφράσεων του έργου Mitos (συγκρίνεται η χειρωνακτική επισημείωση AegeanNcsrTempExp με την επισημείωση tempexp του συστήματος του έργου Mitos). Τα αποτελέσματα τα οποία προέκυψαν απεικονίζονται παρακάτω στον πίνακα 5:

Total Correct Annotations	Total Key Annotations	Total Response Annotations	Precision	Recall	F-measure
290	930	649	0,44	0,31	0,36

Πίνακας 5 – Αποτελέσματα αναγνώρισης χρονικών εκφράσεων του συστήματος του project Mitos

Όπως φαίνεται, τα αποτελέσματα τα οποία προέκυψαν δεν είναι ενθαρρυντικά, αλλά αυτό κατά κάποιο τρόπο ήταν αναμενόμενο, δεδομένου ότι το σύστημα αναγνώρισης και κανονικοποίησης χρονικών εκφράσεων υλοποιήθηκε για συγκεκριμένη συλλογή κειμένων (corpus), όπως άλλωστε και το δικό μας σύστημα κατά κάποιο τρόπο. Έτσι όταν εφαρμόστηκε στη δική μας συλλογή απέτυχε να αναγνωρίσει ένα ικανοποιητικό πλήθος χρονικών εκφράσεων.

Όσον αφορά στην κανονικοποίηση των χρονικών εκφράσεων τώρα, δεν ήταν εφικτό να πραγματοποιηθεί σύγκριση λόγω του ότι το σύστημα του Mitos προσθέτει στην τιμή της ιδιότητας και άλλη μια πληροφορία, η οποία αφορά στα αναγνωριστικά (ids) των λεκτικών μονάδων από τα οποία αποτελείται μια χρονική έκφραση. Εφόσον, λοιπόν, η τιμή της κοινής ιδιότητας timeperiod της χειρωνακτικής επισημείωσης AegeanNcsrTempExp και της επισημείωσης του συστήματος Mitos tempexp διαφέρει εκ των προτέρων, κατέστη αδύνατο να πραγματοποιηθεί σύγκριση, δεδομένου ότι δε θα είχαμε κανένα σωστό αποτέλεσμα. Άλλωστε και σ' αυτή την περίπτωση - για τον προαναφερθέντα λόγο – εικάζουμε πως τα αποτελέσματα δεν θα ήταν ικανοποιητικά.

Ως αποτέλεσμα, αυτό που γίνεται φανερό από τα προλεγόμενα είναι ότι το σύστημά μας για τη συγκεκριμένη συλλογή κειμένων (corpus) υπερτερεί.

4 Μηνύματα

4.1 Εισαγωγή

Ο δεύτερος στόχος της μεταπτυχιακής διπλωματικής εργασίας είναι η συμπλήρωση των ορισμάτων των μηνυμάτων που ορίζονται για τα κείμενα μιας συλλογής, με τις τιμές τους. Υπενθυμίζουμε πως αυτό είναι καθοριστικής σημασίας, γιατί οι τιμές των ορισμάτων των μηνυμάτων αποτελούν έναν από τους παράγοντες που καθορίζουν το είδος της σχέσης που θα δημιουργηθεί μεταξύ των μηνυμάτων (π.χ. αν θα είναι Agreement, Near Agreement ή Disagreement προκείμενου για μία συγχρονική σχέση, εξαρτάται τα αποτελέσματα της σύγκρισης των τιμών των ορισμάτων των μηνυμάτων).

Λόγω του γεγονότος ότι η προσέγγιση όσον αφορά στην συμπλήρωση των ορισμάτων των μηνυμάτων είναι ενδεικτική επιλέχτηκε ο τύπος μηνύματος free που έχει και τη μεγαλύτερη συχνότητα εμφάνισης (για τις προδιαγραφές του μηνύματος free καθώς και των υπολοίπων μηνυμάτων βλέπε στο Παράρτημα την ενότητα «Προδιαγραφές των μηνυμάτων») Οι τύποι των ορισμάτων βασίζονται στην οντολογία του πεδίου (domain ontology) (βλέπε στο Παράρτημα την ενότητα «Οντολογία των μηνυμάτων»).

Η ενδεικτική αυτή προσέγγισή μας έχει τα εξής χαρακτηριστικά:

- Αναλύει μόνο προτάσεις που αντιστοιχούν σε μηνύματα τύπου free συμπληρώνοντας τα ορίσματά τους
- Χρησιμοποιεί ευριστικά (heuristics) που δε λαμβάνουν υπόψη τους το συγκείμενο (context), αλλά μόνο τους περιορισμούς (constraints) των ορισμάτων των μηνυμάτων, αν βέβαια υπάρχουν (βλέπε στο Παράρτημα τους περιορισμούς των ορισμάτων των μηνυμάτων)
- Χρησιμοποιεί κανόνες με τη μορφή κανονικών εκφράσεων που βασίζονται μόνο στο περιεχόμενο των εκφράσεων, προκειμένου να αναγνωριστεί πληροφορία στα κείμενά μας που δεν έχει επισημειωθεί (π.χ. πληροφορίας που σχετίζεται με κάποια ποσότητα)

Η μεθοδολογία που ακολουθήθηκε περιλαμβάνει τα ακόλουθα βήματα:

- Χειρωνακτική επισημείωση των μηνυμάτων με καθορισμό του τύπου τους και των τιμών των ορισμάτων τους καθώς και των οντοτήτων και των τύπων τους στα κείμενα της συλλογής μας.
- Χρήση ευριστικών που δε λαμβάνουν υπόψη τους το συγκείμενο (context), αλλά μόνο τους περιορισμούς (constraints) των ορισμάτων των μηνυμάτων. Χρήση κανονικών εκφράσεων για ορισμένους τύπους ορισμάτων
- Γέμισμα των ορισμάτων των μηνυμάτων Αξιολόγηση των αποτελεσμάτων

Τα βήματα αυτά περιγράφονται στις ενότητες που ακολουθούν.

4.2 Δημιουργία σώματος κειμένων αξιολόγησης

Πραγματοποιήθηκε στα πλαίσια της μεταπτυχιακής εργασίας των κ.κ. Μαρία Σαλαπάτα και Κωνσταντίνα Λιοντού χειρωνακτική επισημείωση των μηνυμάτων (συνήθως προτάσεις του κειμένου) στα 164 έγγραφα της συλλογής κειμένων αξιολόγησης και καθορίστηκαν οι τύποι τους (δημιουργήθηκαν, δηλαδή, χειρωνακτικά επισημειώσεις τύπου original_message με τιμές στις ιδιότητές τους τούς τύπους των μηνυμάτων), καθώς και πληροφορίες όσον αφορά στους ανθρώπους και τους τύπους τους (δημιουργήθηκαν δηλαδή χειρωνακτικά επισημειώσεις τύπου Persons με τιμές στις ιδιότητες τους Offender, Rescue Team, Hostage

κ.ο.κ.), στις τοποθεσίες και τους τύπους τους (δημιουργήθηκαν, δηλαδή, χειρωνακτικά επισημειώσεις τύπου Place με τιμές στις ιδιότητες τους City, Country κ.ο.κ.) κ.ο.κ. (βλέπε στο Παράρτημα την οντολογία των μηνυμάτων). Ακολούθησε χειρωνακτική επισημείωση των μηνυμάτων και των τιμών των ιδιοτήτων (ορισμάτων) τους, λόγω του ότι στο τέλος θα χρειαζόταν να πραγματοποιηθεί σύγκριση μεταξύ της επισημείωσης των μηνυμάτων και των τιμών των ιδιοτήτων τους που συμπληρώθηκαν χειρωνακτικά και της επισημείωσης των μηνυμάτων και των τιμών των ιδιοτήτων τους που θα συμπληρώνονταν υπολογιστικά – προγραμματιστικά.

4.3 Μεθοδολογία συμπλήρωσης των ορισμάτων των μηνυμάτων

Η μεθοδολογία της συμπλήρωσης των ορισμάτων των μηνυμάτων στα κείμενα (documents) μιας συλλογής κειμένων (corpus) στηρίζεται στο συνδυασμό χρήσης του Java Application Programming Interface (API) [9] της πλατφόρμας επεξεργασίας φυσικής γλώσσας Ellogon [7] [8], της τεχνολογίας Κανονικών Εκφράσεων της γλώσσας προγραμματισμού Java και γενικότερα του API της Java.

Εν συντομία τα βήματα της προσέγγισης για τη συμπλήρωση των ορισμάτων των μηνυμάτων που επιλέχθηκε και ονομάζεται SystemAegeanNcsrMessageArgumentFilling είναι τα εξής:

Καταρχάς λαμβάνεται κάθε φορά το κείμενο της συλλογής σε μορφή ακατέργαστων δεδομένων (raw data). Αυτό εν συνεχεία μετατρέπεται σε συμβολοσειρά (String).

Έπειτα επιλέγονται όλες οι επισημειώσεις τύπου original_message και εφαρμόζεται φιλτράρισμα λαμβάνοντας τελικά μόνο αυτές τις ιδιότητες που έχουν τιμή free. Για κάθε τώρα επισημείωση τύπου original_message με ιδιότητες που έχουν τιμή free λαμβάνεται το κείμενο της επισημείωσης σε μορφή ακολουθίας από bytes (ByteSequence), μετατρέπεται σε συμβολοσειρά και λαμβάνονται τα όρια της έκτασης της επισημείωσης.

Αφετέρου επιλέγονται όλες τις επισημειώσεις τύπου Persons και εφαρμόζεται φιλτράρισμα λαμβάνοντας τελικά μόνο αυτές μόνο αυτές τις ιδιότητες που έχουν τιμή Offender ή Rescue_Team. Για κάθε τώρα επισημείωση τύπου Persons με ιδιότητες που έχουν τιμή Offender ή Rescue_Team λαμβάνονται τα όρια της έκτασης της επισημείωσης και αν αυτά είναι μεταξύ των ορίων της έκτασης της επισημείωσης τύπου original_message, τότε αποθηκεύεται σε μορφή συμβολοσειράς το id της επισημείωσης και τελειώνει εδώ η επανάληψη, διαφορετικά τίθεται η τιμή "null" σε μορφή συμβολοσειράς.

Ομοίως επιλέγονται όλες τις επισημειώσεις τύπου Persons και εφαρμόζεται φιλτράρισμα λαμβάνοντας τελικά μόνο αυτές μόνο αυτές τις ιδιότητες που έχουν τιμή Hostage. Για κάθε τώρα επισημείωση τύπου Persons με ιδιότητες που έχουν τιμή Hostage λαμβάνονται τα όρια της έκτασης της επισημείωσης και αν αυτά είναι μεταξύ των ορίων της έκτασης της επισημείωσης τύπου original_message, τότε αποθηκεύεται σε μορφή συμβολοσειράς το id της επισημείωσης και τελειώνει εδώ η επανάληψη, διαφορετικά τίθεται η τιμή "null" σε μορφή συμβολοσειράς.

Έπειτα επιλέγονται όλες τις επισημειώσεις τύπου Place. Για κάθε τώρα επισημείωση τύπου Place λαμβάνονται τα όρια της έκτασης της επισημείωσης και αν αυτά είναι μεταξύ των ορίων της έκτασης της επισημείωσης τύπου original_message, τότε αποθηκεύεται σε μορφή συμβολοσειράς το id της επισημείωσης και τελειώνει εδώ η επανάληψη, διαφορετικά τίθεται η τιμή "null" σε μορφή συμβολοσειράς.

Στη συνέχεια δημιουργούνται οι μορφές (patterns) των κανονικών εκφράσεων για την ποσότητα (quantity) σε μορφή συμβολοσειράς και μετατρέπεται το κείμενο της επισημείωσης original_message σε ακολουθία χαρακτήρων (Charsequence). Έπειτα δημιουργείται ένας Matcher που αντιστοιχίζει γενικά τις κανονικές εκφράσεις με την ακολουθία χαρακτήρων της επισημείωσης original_message. Για όσο τώρα πραγματοποιείται ταίριασμα της μορφής (pattern) των κανονικών εκφράσεων που δημιουργήσαμε με κάποια επόμενη ποσότητα (υποτίθεται ότι έχει προηγηθεί ένα ταίριασμα) και εάν αυτή ταιριάζει με κάποιο pattern κανονικής έκφρασης που αναπαριστά ποσότητα, τότε αποθηκεύεται η ποσότητα που ταιριάστηκε σε μορφή συμβολοσειράς και τελειώνει η επανάληψη, διαφορετικά τίθεται η τιμή “null” σε μορφή συμβολοσειράς.

Εν τέλει, δημιουργείται επισημείωση τύπου SystemAegeanNcsrMessageArgumentFilling με ιδιότητες τύπου free, entity1, entity2, from_place και quantity.

Η ίδια διαδικασία μπορεί να ακολουθηθεί και για τους υπόλοιπους τύπους μηνυμάτων (Στο Παράρτημα παρουσιάζονται πιο συγκεκριμένα τα βήματα της προκείμενης προσέγγισης).

4.4 Πειράματα – Αποτελέσματα

Τα πειράματα πραγματοποιήθηκαν πάνω στη συλλογή gold_corpus_final, η οποία συμπεριλαμβάνεται στο συνοδευτικό cd της μεταπτυχιακής. Συγκρίθηκε η συμπλήρωση των ορισμάτων του μηνύματος free που πραγματοποιήθηκε χειρωνακτικά με τη συμπλήρωση των ορισμάτων του μηνύματος τύπου free που πραγματοποιήθηκε αυτόματα από το σύστημα. Συγκεκριμένα, πραγματοποιήθηκαν πειράματα διακριτά όσον αφορά στα ορίσματα του μηνύματος τύπου free entity1, entity2, from_place, quantity, συνδυασμό των entity1 & entity2 καθώς και συνδυασμός των entity1, entity2, from_place & quantity (Στο Ellogon τα ορίσματα αυτά ονομάζονται ιδιότητες και αυτό που συγκρίνεται είναι οι τιμές τους). Τα αποτελέσματα παρουσιάζονται στον πίνακα 6 παρακάτω (βλέπε παρακάτω στο Παράρτημα για τη διαδικασία της αξιολόγησης στην πλατφόρμα επεξεργασίας φυσικής γλώσσας Ellogon καθώς και για τη σημασία των μετρικών στα αποτελέσματα που προκύπτουν):

Compare Attribute	Total Correct Annotations	Total Key Annotations	Total Response Annotations	Precision	Recall	F-measure
entity1	86	127	133	0,64	0,67	0,66
entity2	88	127	133	0,66	0,69	0,67
from_place	64	127	133	0,48	0,50	0,49
Quantity	36	127	133	0,27	0,28	0,27
entity1, entity2	66	127	133	0,49	0,51	0,50
entity1, entity2, from_place, quantity	8	127	133	0,06	0,06	0,06

Πίνακας 6 - Αποτελέσματα σύγκρισης της τιμής των ιδιοτήτων entity1, entity2, from_place & quantity, entity1 & entity2 από κοινού & entity1, entity2, from_place & quantity από κοινού για μηνύματα τύπου free

Τα παραπάνω αποτελέσματα θα μπορούσαμε να τα σχολιάσουμε ως εξής:

- Όσον αφορά στις συγκρίσεις της τιμής των ιδιοτήτων entity1 και entity2 τόσο μεμονωμένα όσο και από κοινού θα λέγαμε ότι τα αποτελέσματα είναι αρκετά ικανοποιητικά δεδομένου ότι δε λαμβάνεται υπόψη κανενός είδους γλωσσολογική πληροφορία (για παράδειγμα μπορεί να υπάρχει στο κείμενο του μηνύματος μια αναφορά σε μια οντότητα (π.χ. «αυτοί») και το σύστημα να τοποθετεί λαθεμένα την τιμή "null"). Επίσης σε περίπτωση που μέσα σ' ένα μήνυμα ενός κειμένου βρίσκονται περισσότερες από μία τιμές του ίδιου τύπου που μπορεί να πάρει η ίδια η ιδιότητα (π.χ. όταν έχουμε δύο ομήρους), τότε επιλέγεται κάθε φορά μόνο αυτή που βρίσκει πρώτη το σύστημα με συνέπεια οι υπόλοιπες τιμές να αγνοούνται και να προκύπτουν λάθος αποτελέσματα. Εν τέλει, λάθη μπορεί να οφείλονται στο γεγονός ότι μπορεί να μην έχει επισημειωθεί καθόλου μια οντότητα Person στο κείμενό μας με αποτέλεσμα το σύστημά μας να μην μπορεί να τη βρει και να τοποθετεί λαθεμένα την τιμή null, είτε να έχει επισημειωθεί χειρωνακτικά λαθεμένα παρόλο που το σύστημα βρίσκει τη σωστή τιμή
- Όσον αφορά στη σύγκριση της τιμής της ιδιότητας from_place τα αποτελέσματα δεν είναι και τόσο ενθαρρυντικά και αυτό γιατί είτε δεν έχει επισημειωθεί καθόλου μια οντότητα Place στο κείμενό μας και επομένως το σύστημά μας δεν μπορεί να τη βρει και τοποθετεί λαθεμένα την τιμή null, είτε υπάρχει με τη μορφή αναφοράς (π.χ. «εκεί») που οδηγεί αναπόφευκτα στο ίδιο αποτέλεσμα, είτε έχει επισημειωθεί χειρωνακτικά λαθεμένα, παρόλο που το σύστημα βρίσκει τη σωστή τιμή .Άλλος ένας λόγος θα μπορούσε να είναι το γεγονός ότι σ' ένα μήνυμα ενός κειμένου μπορεί να υπάρχουν περισσότερες από δύο πιθανές τιμές του ίδιου τύπου της παραπάνω ιδιότητας (π.χ. δύο χώρες), όποτε επιλέγεται κάθε φορά μόνο αυτή που βρίσκει πρώτη το σύστημα με συνέπεια οι υπόλοιπες τιμές να αγνοούνται και να προκύπτουν λάθος αποτελέσματα
- Όσον αφορά στη σύγκριση της τιμής της ιδιότητας quantity, θα λέγαμε ότι και εδώ τα αποτελέσματα δεν είναι καθόλου ενθαρρυντικά, πράγμα όμως το οποίο δικαιολογείται, δεδομένου ότι το σύστημά μας μπορεί και αναγνωρίζει μόνο αριθμητικές ποσότητες (π.χ. το «τρεις» δεν αναγνωρίζεται)
- Όσον αφορά στις συγκρίσεις της τιμής των ιδιοτήτων entity1, entity2, from_place και quantity από κοινού τα αποτελέσματα θα λέγαμε ότι είναι απογοητευτικά (μόνο 8 σωστές επισημειώσεις), πράγμα αναμενόμενο από τα προλεγόμενα

5 Συμπεράσματα και προτάσεις για μελλοντικές κατευθύνσεις

Υλοποιήθηκε μία αρχική έκδοση του συστήματος αναγνώρισης και κανονικοποίησης χρονικών εκφράσεων, η οποία για την αναγνώρισή τους λαμβάνει υπόψη μόνο το περιέχομενό τους και για την κανονικοποίησή τους συνήθως την ημερομηνία δημοσίευσης της είδησης, ενώ δε λαμβάνει υπόψη καθόλου το συγκείμενο.

Αυτή θα μπορούσε να βελτιωθεί σημαντικά και αυτό είναι δυνατό να επιτευχθεί κατά κύριο λόγο με την αξιοποίηση ορισμένων παραμέτρων, οι οποίοι στην προκείμενη μεταπτυχιακή εργασία δε λήφθηκαν υπόψη.

Καθοριστικής σημασίας παράμετρος θεωρείται η χρήση γλωσσολογικής πληροφορίας, όπως είναι το συγκείμενο. Με τη χρήση αυτού θα μπορούσε για παράδειγμα να αποφεύγεται η αναγνώριση χρονικών εκφράσεων που δεν είναι στην πραγματικότητα (π.χ η αναγνώριση του «80.00» ως χρονική έκφραση που παριστάνει ώρα, ενώ στην πραγματικότητα είναι μέρος ενός ποσού).

Το ίδιο ισχύει και για την κανονικοποίηση. Ο συνυπολογισμός του συγκείμενου μαζί με την ημερομηνία δημοσίευσης της είδησης θα μπορούσε να έχει ως αποτέλεσμα την αποφυγή λαθών κανονικοποίησης χρονικών εκφράσεων που δεν προσδιορίζονται άμεσα (π.χ. «την Κυριακή»).

Από την στιγμή που θα λαμβάναμε υπόψη μας το συγκείμενο, θα μπορούσε να προστεθεί άλλη μια κατηγορία χρονικών εκφράσεων η οποία θα περιλάμβανε χρονικές εκφράσεις που η σωστή κανονικοποίησή τους συνδέεται άρρηκτα μ’ αυτό, όπως π.χ οι χρονικές εκφράσεις «στη συνέχεια», «πριν από», «μετά από» κ.ο.κ. Θα έπρεπε όμως τότε να επισημειώσουμε ανάλογα χειρωνακτικά τη συλλογή μας.

Επιπλέον, άλλη μια παράμετρος θα μπορούσε να ήταν ο συνδυασμός του συστήματος που υλοποιήθηκε με το σύστημα αναγνώρισης και κανονικοποίησης χρονικών εκφράσεων που χρησιμοποιήθηκε στο έργο Mitos.

Τέλος, εφόσον γίνουν όλες οι απαραίτητες διορθώσεις στην κανονικοποίηση των χρονικών εκφράσεων που πραγματοποιείται λανθασμένα και ληφθούν υπόψη όλοι οι παράμετροι, καλό θα ήταν να επισημανθούν όλες οι ελληνικές χρονικές εκφράσεις και να προστεθούν στο σύστημα, ώστε αυτό να θεωρείται πλήρης και να μπορεί να εφαρμοστεί και σε άλλες συλλογές κειμένων.

Πρέπει να σημειώσουμε, ωστόσο, ότι λόγω του ότι αυτό θέλουμε να αξιοποιήσουμε ακριβώς είναι η χρονική πληροφορία, η οποία συνδέεται με τα μηνύματα μέσω μιας χρονικής έκφρασης, είναι καθοριστικής σημασίας να πραγματοποιηθεί χειρωνακτική επισημείωση μόνο των χρονικών εκφράσεων που περιλαμβάνονται σ’ αυτά.. Επίσης θα πρέπει να προσαρμοστεί ανάλογα και το σύστημά μας, ώστε να λαμβάνει ως είσοδο μόνο το κείμενο που αντιστοιχεί στα μηνύματα. Μόνο έτσι θα μπορούσε να σύστημά μας να ενσωματωθεί στο γενικότερο σύστημα αυτόματης περίληψης από πολλαπλά έγγραφα εξελισσόμενων γεγονότων που βασίζεται σε μια ερώτηση.

Τέλος, πρέπει να αναφερθεί ότι στο σύστημά μας θα μπορούσε να προστεθεί και μία λειτουργία, η οποία κατά την αναγνώριση μιας χρονικής έκφρασης θα καθόριζε και τα αναγνωριστικά (ids) των λεκτικών μονάδων (tokens) από τις οποίες αυτή αποτελείται.

Όσον αφορά στα μηνύματα υλοποιήθηκε μία αρχική έκδοση του συστήματος συμπλήρωσης των ορισμάτων αυτών που δε λαμβάνουν υπόψη τους το συγκείμενο (context), αλλά μόνο

τους περιορισμούς (constraints) των ορισμάτων τους. Συγκεκριμένα συμπληρώνονται αυτόματα οι τιμές των ορισμάτων του μηνύματος τύπου free που είχε και τη μεγαλύτερη συχνότητα εμφάνισης στα κείμενα της συλλογής μας.

Το σύστημά μας διαθέτει προοπτικές επέκτασης και βελτίωσης. Ένα πρώτο βήμα θα ήταν η προσθήκη και των υπολοίπων τύπων μηνυμάτων.

Κατόπιν, η συμπλήρωση των ορισμάτων των μηνυμάτων με τις τιμές τους θα μπορούσε να αντιμετωπιστεί ως ένα πρόβλημα ικανοποίησης περιορισμών (constraint satisfaction problem), όπου μεταβλητές είναι οι τύποι των ορισμάτων των μηνυμάτων (οι τύποι των ιδιοτήτων), πεδίο τιμών για την κάθε μεταβλητή οι τιμές της ιδιότητας οι οποίες ανήκουν στο πεδίο της οντολογίας (domain ontology) και περιορισμοί αυτοί που προκύπτουν από τις προδιαγραφές των μηνυμάτων (τόσο μοναδιαίοι (unary), όσο και δυαδικοί (binary) (βλέπε στο Παράρτημα τους περιορισμούς για τον κάθε τύπο μηνύματος που καθορίζονται στην ενότητα «Προδιαγραφές των μηνυμάτων»).

Επιπλέον, καθοριστικής σημασίας θα ήταν η εξέταση και πάλι του συγκείμενου, έτσι ώστε να μην αγνοούνται τιμές στην περίπτωση που σ’ ένα μήνυμα υπάρχουν περισσότερες από δύο πιθανές τιμές του ίδιου τύπου μιας ιδιότητας (όπως π.χ. η entity1), ενώ όταν υπάρχουν αναφορές (π.χ. «αυτοί»). σ’ αυτό, αυτές να αποσαφηνίζονται.

Τέλος, ένα γενικό συμπέρασμα στο οποίο καταλήξαμε είναι ότι η αυτόματη περίληψη κειμένου και γενικότερα η επεξεργασία φυσικής γλώσσας είναι άρρηκτα συνδεδεμένη με το πεδίο (domain) στο οποίο εφαρμόζεται. Αλλαγή αυτού του πεδίου οδηγεί αναπόφευκτα σε μη ενθαρρυντικά αποτελέσματα. Έτσι λοιπόν περιοριζόμαστε επί του παρόντος σε πολύ συγκεκριμένη χρήση της αυτόματης περίληψης κειμένου, ενώ η γενική χρήση είναι ένας πολύ μακροπρόθεσμος στόχος.

6 Βιβλιογραφία – Αναφορές

[1] S. D. Afantenos, V. Karkaletsis, and P. Stamatopoulos. Summarization from medical documents: A survey. Journal of Artificial Intelligence in Medicine, 33(2):157–177, February 2005.
[2] D. R. Radev, E. H. Hovy, K. McKeown. Introduction to the Special Issue on Summarization. Computational Linguistics 28(4): 399-408 (2002).
[3] I. Mani. Automatic Summarization, volume 3 of Natural Language Processing. John Benjamins Publishing Company, Amsterdam / Philadelphia, 2001.
[4] S. D. Afantenos, I. Doura, E. Kapellou, and V. Karkaletsis. Exploiting cross-document relations for multi-document evolving summarization. In G. A. Vouros and T. Panayiotopoulos, editors, Methods and Applications of Artificial Intelligence: Third Hellenic Conference on AI, SETN 2004, volume 3025 of Lecture Notes in Computer Science, pages 410–419, Samos, Greece, May 2004. Springer-Verlag Heidelberg.
[5] S. D. Afantenos, K. Liontou, M. Salapata, and V. Karkaletsis. An introduction to the summarization of evolving events: Linear and non-linear evolution. In Natural Language Understanding and Cognitive Science NLUCS - 2005, pages 91–99, Maiami, USA, May 2005.
[6] Afantenos, S. D., and V. Karkaletsis. 2004, December. "Linear Evolving Summarization: The First Results." Technical Report 2004/6, Institute of Informatics & Telecommunications, N.C.S.R. "Demokritos", Athens, Greece.
[7] G. Petasis, V. Karkaletsis, G. Paliouras, and C. D. Spyropoulos, "Using the Ellogon Natural Language Engineering Infrastructure". In Proceedings of the Workshop on Balkan Language Resurces and Tools, 1st Balkan Conference in Informatics (BCI 2003), Thessaloniki, Greece, November 21, 2003.
[8] S. D. Afantenos, G. Petasis and V. Karkaletsis. Ellogon User Guide, Software & Knowledge Engineering Laboratory, Institute of Informatics & Telecommunications, National Centre for Scientific Research (NCSR) “Demokritos”, June 2002.
[9] S. D. Afantenos, G. Petasis and V. Karkaletsis. Developers’ Guide to Ellogon, Software & Knowledge Engineering Laboratory, Institute of Informatics & Telecommunications, National Centre for Scientific Research (NCSR) “Demokritos”.
[10] S. D. Afantenos, G. Petasis and V. Karkaletsis. Ellogon Components’ Specifications, Software & Knowledge Engineering Laboratory, Institute of Informatics & Telecommunications, National Centre for Scientific Research (NCSR) “Demokritos”, June 2002.
[11] Περιγραφή Απαιτήσεων του Συστήματος Κανονικοποιήσεως Χρονικών Εκφράσεων, 3 Δεκεμβρίου 1999, Τεκμηρίωση εργαλείων ΜΙΤΟΣ, Εργαστήριο Τεχνολογίας Γνώσεων και Λογισμικού, Ινστιτούτο Πληροφορικής και Τηλεπικοινωνιών, Ε.Κ.Ε.Φ.Ε. «Δημόκριτος».

7 Παράρτημα

7.1 Περιγραφή της συλλογής κειμένων που χρησιμοποιήθηκε

Η συλλογή κειμένων (corpus) που χρησιμοποιήθηκε αποτελείτο από 164 κείμενα από διάφορες ηλεκτρονικές ειδησεογραφικές πηγές στο διαδίκτυο και αφορούν γεγονότα ομηριών ανθρώπων από το 1999 έως το 2004. Συγκεκριμένα τα κείμενα περιέχουν περιγραφές σχετικών γεγονότων, δεδομένου ότι αυτά εξελίσσονται με το χρόνο, από διάφορες πηγές. Οι πηγές αυτές ήταν ο διαδικτυακός τόπος in.gr, ο διαδικτυακός τόπος bbc.com (British Broadcasting Corporation), ο διαδικτυακός τόπος nea.gr (εφημερίδα «τα Νέα»), ο διαδικτυακός τόπος mpa.gr (Μακεδονικό Πρακτορείο Ειδήσεων), ο διαδικτυακός τόπος enet.gr (εφημερίδα «Ελευθεροτυπία»), ο διαδικτυακός τόπος ert.gr (Ελληνική Ραδιοφωνία Τηλεόραση) και ο διαδικτυακός τόπος ant1.gr (ANTENNA GROUP).Τα ονόματα των κειμένων είναι της μορφής nameOfSource_yyyymmddhhmm ή nameOfSource_yyyymmdd, όπου nameofSource το όνομα της πηγής του κειμένου, yyyy το έτος, mm ο μήνας, dd η μέρα, hh η ώρα και mm τα λεπτά. Σημειώνουμε ότι η προκείμενη περιγραφή της ημερομηνίας αφορά στην ημερομηνία δημοσίευσης του κειμένου. Για παράδειγμα το όνομα ant1_200409080940 ενός αρχείου σημαίνει ότι η πηγή της είδησης είναι ο διαδικτυακός τόπος ant1.gr και ότι η είδηση του άρθρου δημοσιεύτηκε στις 8/9/2004 και ώρα 9:40 π.μ. Τέλος, χρίζει αναφοράς ότι τα κείμενα της συλλογής είναι αρχεία txt.

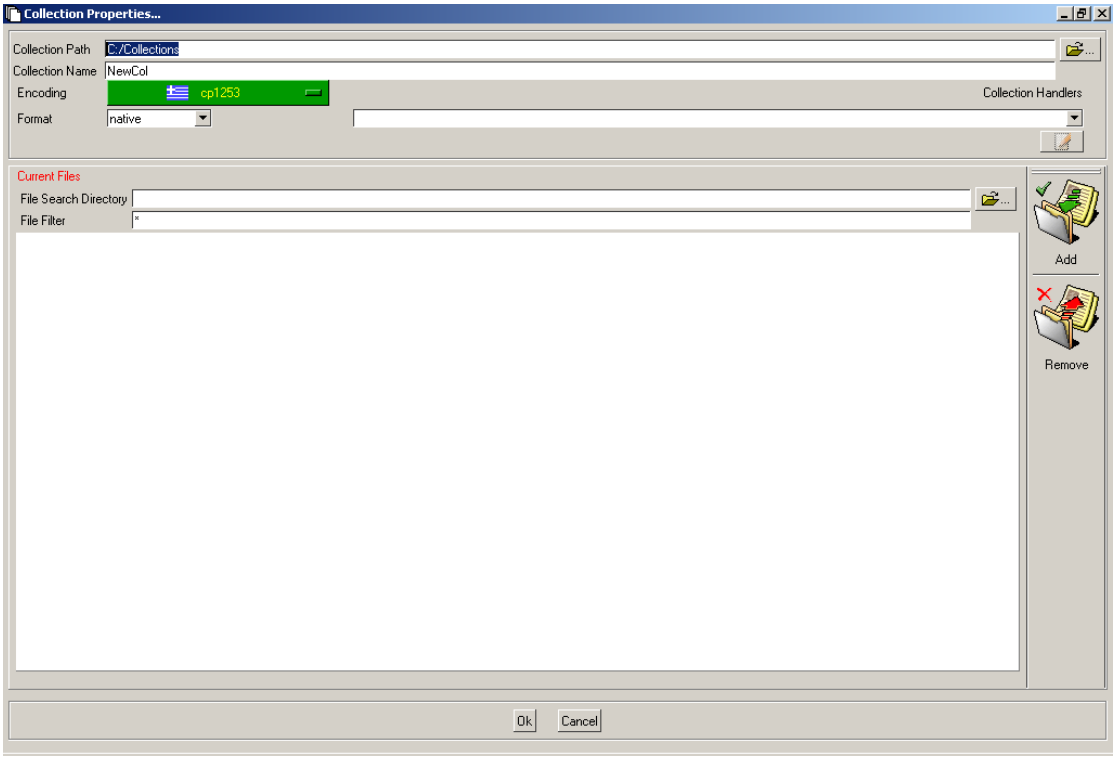
7.2 Χρήση του Ellogon

7.2.1 Χρονικές Εκφράσεις

7.2.1.1 Εκτέλεση του αρθρώματος TemporalExpressions

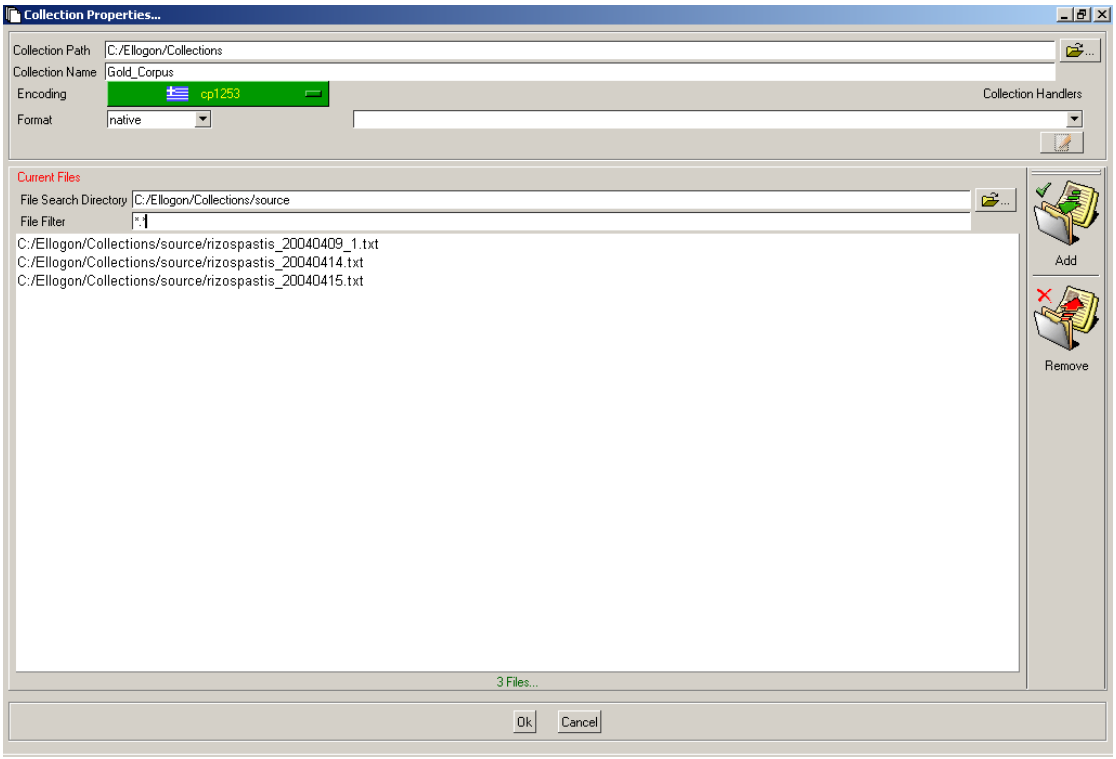
Προκειμένου να τρέξουμε το άρθρωμα TemporalExpressions (αλλά και γενικότερα κάποιο άρθρωμα) στην πλατφόρμα Ellogon ακολουθούμε την εξής διαδικασία, υποθέτοντας ότι έχουμε εγκαταστήσει το δυαδικό (binary) αρχείο της windows έκδοσης 1.7.0 του Ellogon μαζί με την TCL / TK 8.4 ή νεότερη έκδοση και το JAVA JDK / JRE 1.4.1 ή νεότερο) (για περισσότερες πληροφορίες ανατρέξτε στον οδηγό χρήστη (User Guide) της πλατφόρμας επεξεργασίας φυσικής γλώσσας Ellogon) [8]:

Πρώτα απ' όλα δημιουργούμε μια συλλογή επιλέγοντας από το κύριο παράθυρο του Ellogon "Create Collection" από το μενού "Collection". Μετά από την ενέργεια αυτή θα προκύψει το παρακάτω παράθυρο (βλέπε εικόνα 1):



Εικόνα 1 - Δημιουργία Συλλογής

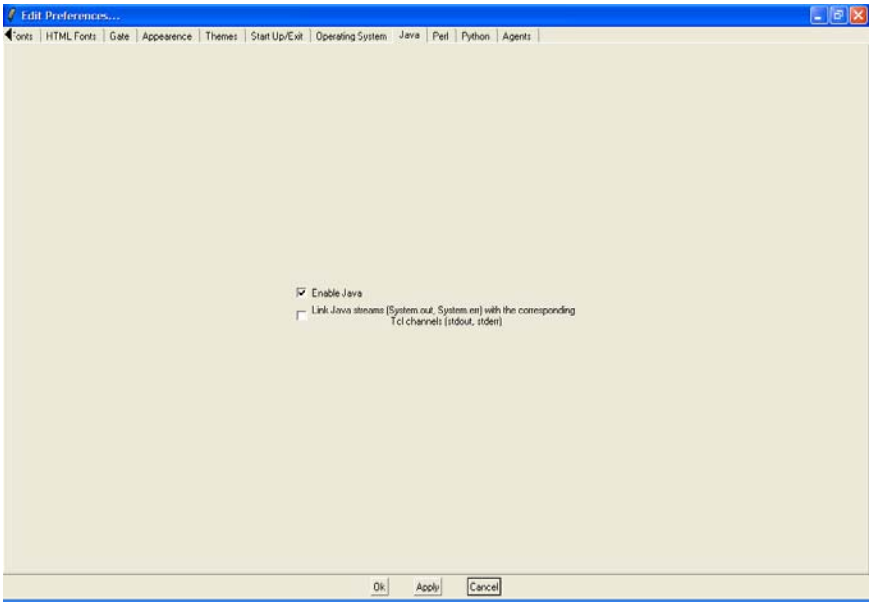
Δίνουμε στο πεδίο κειμένου "Collection Path" το μονοπάτι όπου θα δημιουργηθεί η συλλογή (για παράδειγμα τοποθετούμε C:/Ellogon/Collections) και στο πεδίο κειμένου "Collection Name" το όνομα της συλλογής (στην περίπτωσή μας Gold_Corpus_Final) (αυτός θα είναι ο κατάλογος κάτω από τον οποίο θα δημιουργηθεί η συλλογή). Στο πεδίο κείμενου "File Search Directory" δίνουμε το μονοπάτι, όπου βρίσκονται τα έγγραφα της συλλογής μας, τα οποία πρόκειται να επισημειωθούν (για παράδειγμα C:/Ellogon/Collections/source) και στο πεδίο κειμένου "File Filter" κάποιους χαρακτήρες μπαλαντέρ (wildcard characters) μαζί με τις επιθυμητές προεκτάσεις ονομάτων αρχείων, προκειμένου να επιλέξουμε τα έγγραφα που θέλουμε να επισημειώσουμε (προκειμένου να τα επιλέξουμε όλα βάζουμε *.*). Τα άλλα πεδία τα αφήνουμε όπως έχουν. (για περισσότερες πληροφορίες ανατρέξτε στον οδηγό χρήστη του Ellogon). Αν τα έχουμε πράξει όλα αυτά, τότε θα μας παρουσιαστεί ένα παράθυρο παρόμοιο με την παρακάτω εικόνα 2:



Εικόνα 2 - Προσθήκη εγγράφων προς επισημείωση

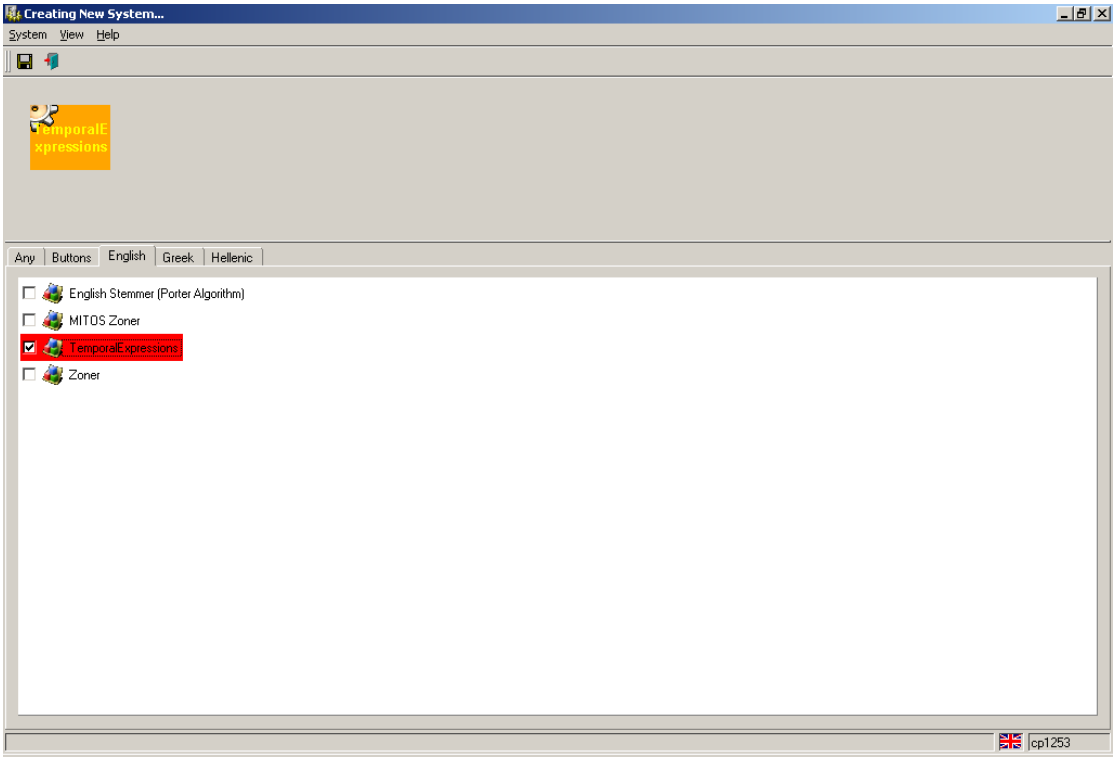
Πατάμε το κουμπί “add”, ώστε να τα προσθέσουμε στη συλλογή μας. Τέλος, πατώντας το κουμπί “Ok” η συλλογή δημιουργείται. Δεν ξεχνάμε να προσθέσουμε το μονοπάτι, όπου αποθηκεύσαμε τη συλλογή μας, επιλέγοντας “Modify Search Path” από το μενού “Collection” του βασικού παραθύρου του Ellogon. Αξίζει να σημειωθεί ότι όταν η συλλογή μας είναι ήδη επισημειώμενη, το μόνο που έχουμε να κάνουμε είναι να την τοποθετήσουμε για παράδειγμα κάτω από τον υποκατάλογο C:/Ellogon/Collections.

Κατόπιν θα πρέπει να δημιουργήσουμε το σύστημα TemporalExpressions. Προτού το επιδιώξουμε αυτό, δεν θα πρέπει να ξεχάσουμε να ενεργοποιήσουμε στο Ellogon την υποστήριξη της Java. Για να επιτευχθεί αυτό, επιλέγουμε “options” από το μενού “File” του βασικού παραθύρου και στην ετικέτα Java επιλέγουμε το πλαίσιο ελέγχου (check box) “Enable Java” (βλέπε εικόνα 3). Βγαίνουμε από το Ellogon και ξαναεισερχόμαστε, προκειμένου να ισχύσουν οι αλλαγές. Ελέγχουμε αν επιλέγοντας “Component Programming Languages” από το μενού “Modules” του βασικού παραθύρου του Ellogon είναι επιλεγμένη η γλώσσα Java.



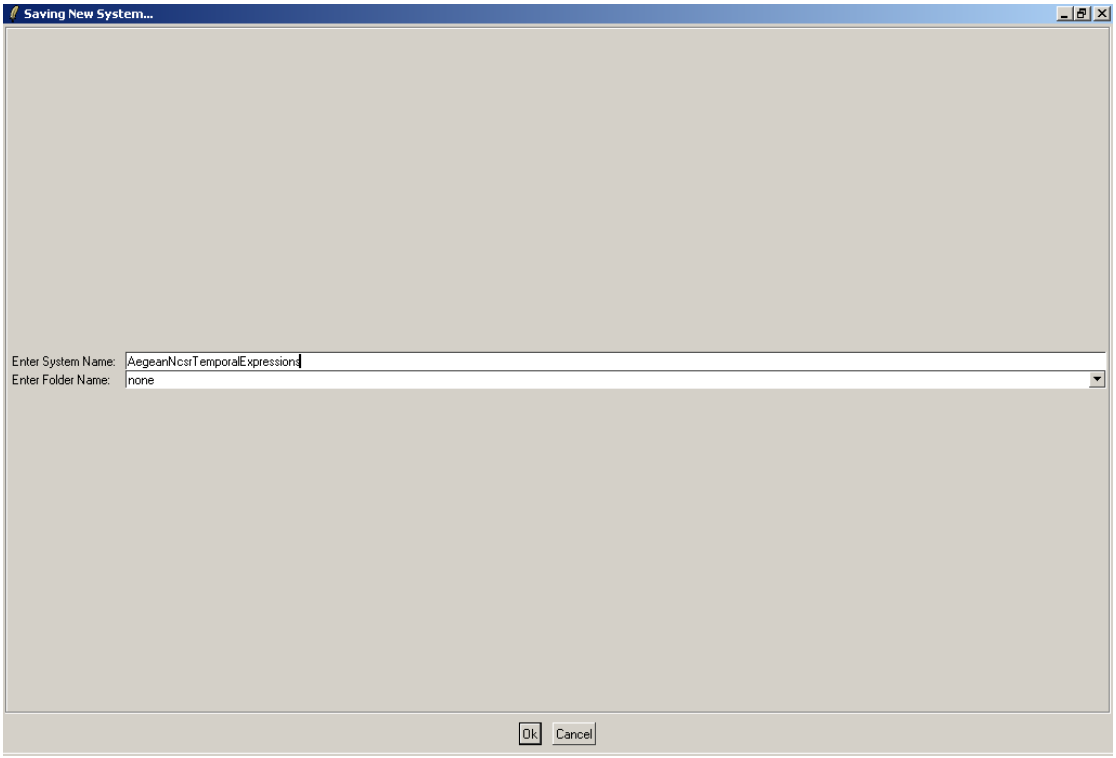
Εικόνα 3 - Ενεργοποίηση της Java

Τώρα είμαστε έτοιμοι να δημιουργήσουμε το σύστημα TemporalExpressions. Καταρχήν έχουμε φροντίσει να τοποθετήσουμε τον υποκατάλογο TemporalExpressions κάτω από το μονοπάτι C:\Ellogon\modules. Καλό θα ήταν να αντιγραφούν όλα τα αρθρώματα (modules) του Ellogon κάτω από το μονοπάτι αυτό (προσοχή τα ονόματα των καταλόγων ή υποκαταλόγων που αποθηκεύονται τα αρθρώματα να μην έχουν κενούς χαρακτήρες, διότι το άρθρωμα δεν θα είναι διαθέσιμο στο Ellogon). Για να δημιουργήσουμε το σύστημα, τώρα, επιλέγουμε “create new system” από το μενού του βασικού παραθύρου του Ellogon και επιλέγουμε το πλαίσιο ελέγχου TemporalExpressions στην ετικέτα English (βλέπε εικόνα 4).



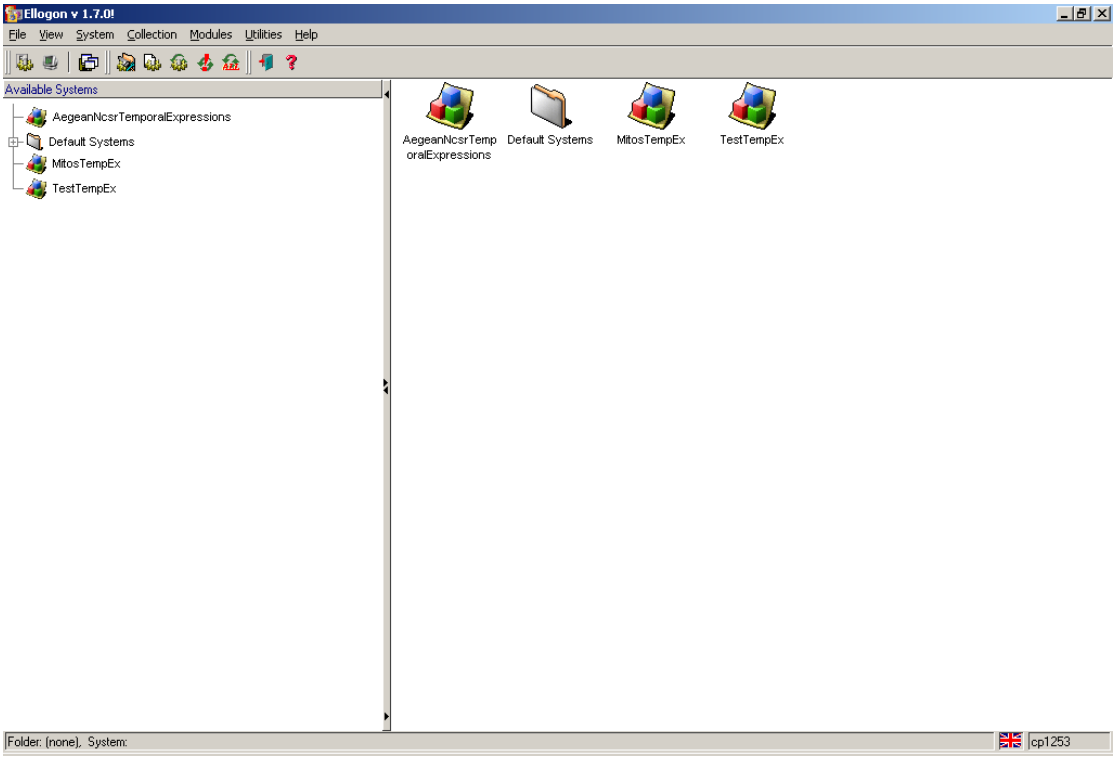
Εικόνα 4 - Δημιουργία Συστήματος

Πατάμε το κουμπί “Save” στη γραμμή εργαλείων του παραθύρου της δημιουργίας καινούργιου συστήματος, δίνοντας το όνομα AegeanNcsrTemporalExpressions στο πεδίο κειμένου “Enter System Name” και τον υποκατάλογο που θα αποθηκευτεί στο πτυσσόμενο πλαίσιο συνδυασμού (Drop-Down Combination Box) “Enter Folder Name” (βλέπε εικόνα 5).



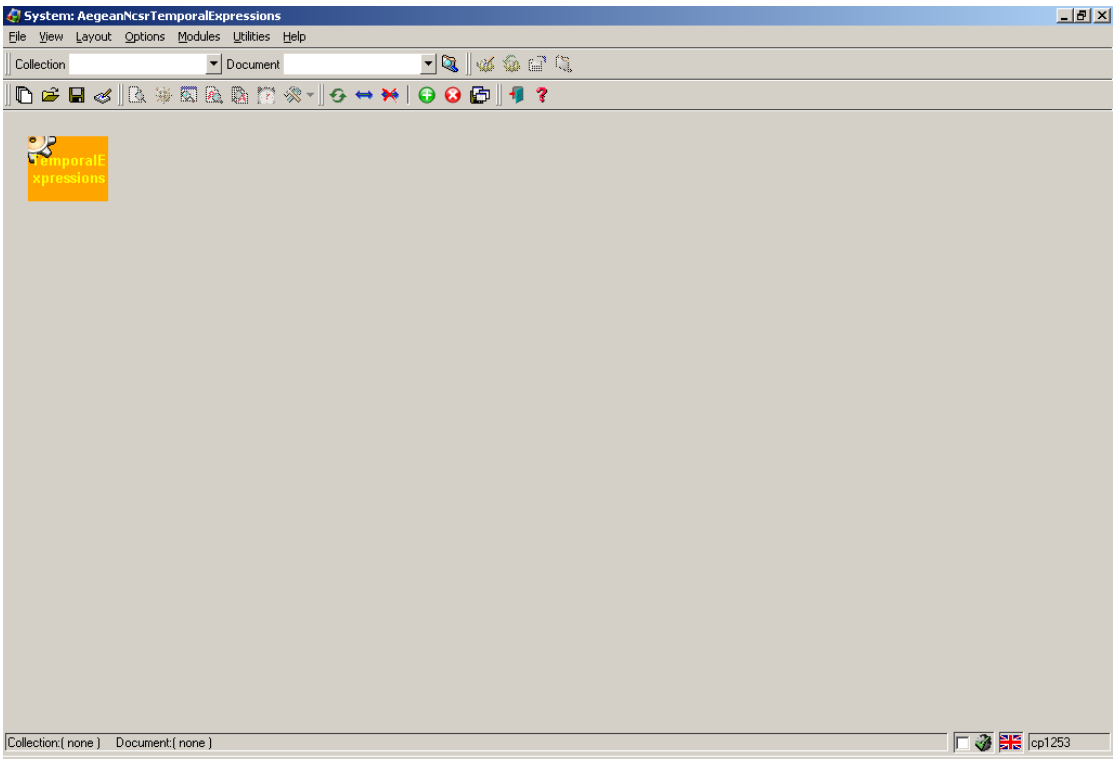
Εικόνα 5 - Αποθήκευση νέου συστήματος

Όπως φαίνεται στην παρακάτω εικόνα 6 το σύστημα μας προστέθηκε επιτυχώς στο σύστημα:



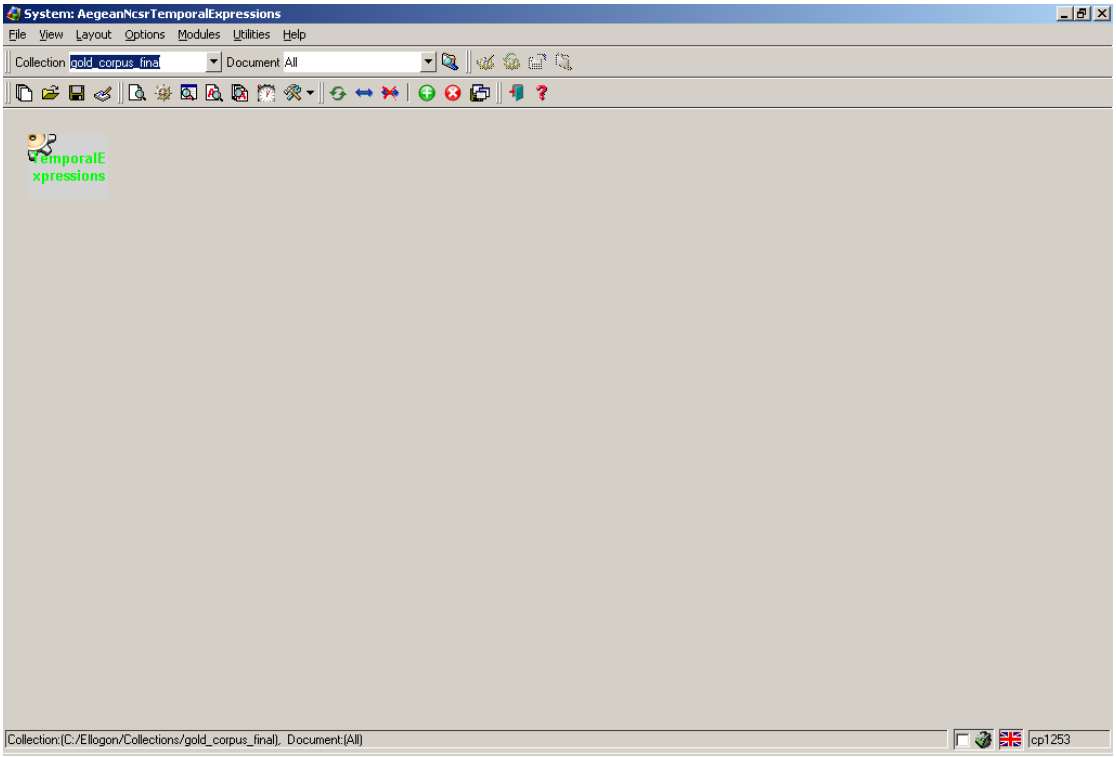
Εικόνα 6 – Επιτυχής προσθήκη νέου συστήματος

Αν κάνουμε διπλό κλικ στο σύστημα AegeanNcsrTemporalExpressions τότε θα εμφανιστεί η παρακάτω εικόνα 7:



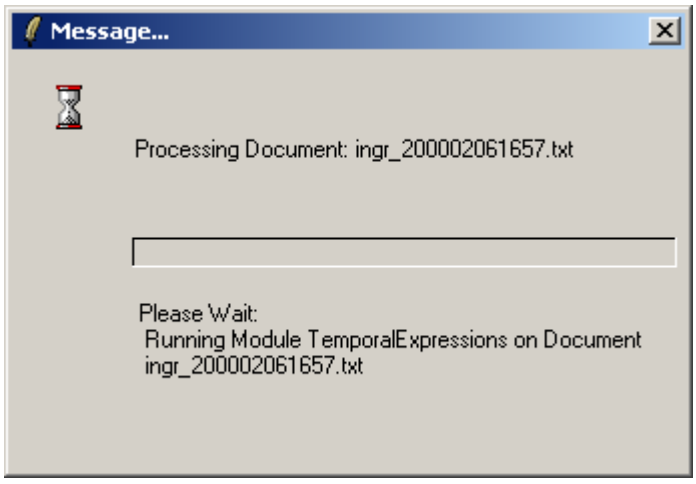
Εικόνα 7 – Το παράθυρο του συστήματος AegeanNcsrAegeanTempExp

Στο παράθυρο του συστήματος που εμφανίζεται επιλέγουμε τη συλλογή που επιθυμούμε (στην περίπτωσή μας τη συλλογή Gold_Corpus_Final) στο πτυσσόμενο πλαίσιο λίστας (Drop-Down List Box) "Collection" και το έγγραφο που θέλουμε να επισημειώσουμε στο πτυσσόμενο πλαίσιο λίστας "Documents" (αν τα θέλουμε όλα αφήνουμε την προκαθορισμένη επιλογή "All") και πατάμε το ακριβώς διπλανό κουμπί, ώστε να ανοίξουμε το έγγραφο ή όλη τη συλλογή. Τότε το άρθρωμά μας θα είναι έτοιμο να εκτελεστεί (βλέπε εικόνα 8).



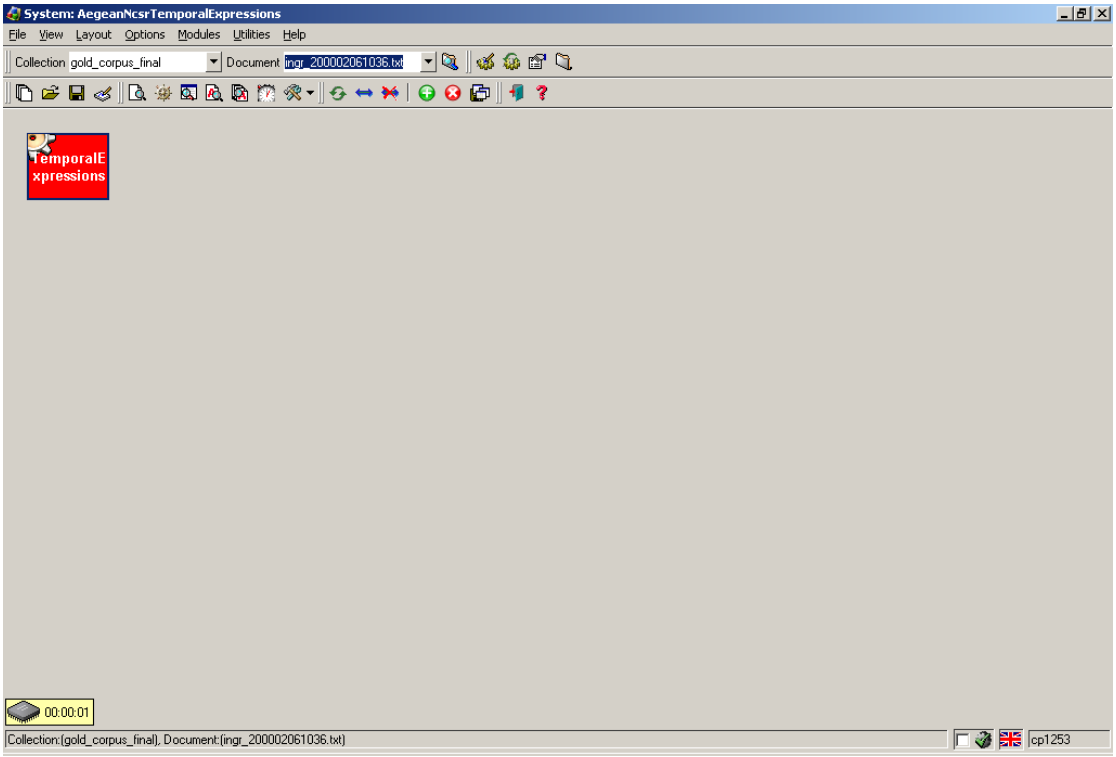
Εικόνα 8 - Το άρθρωμα TemporalExpressions είναι έτοιμο προς εκτέλεση

Αφετέρου κάνουμε διπλό κλικ πάνω στο άρθρωμα TemporalExpressions. Τότε αυτό θα αρχίσει να εκτελείται, όπως φαίνεται και στην παρακάτω εικόνα 9:



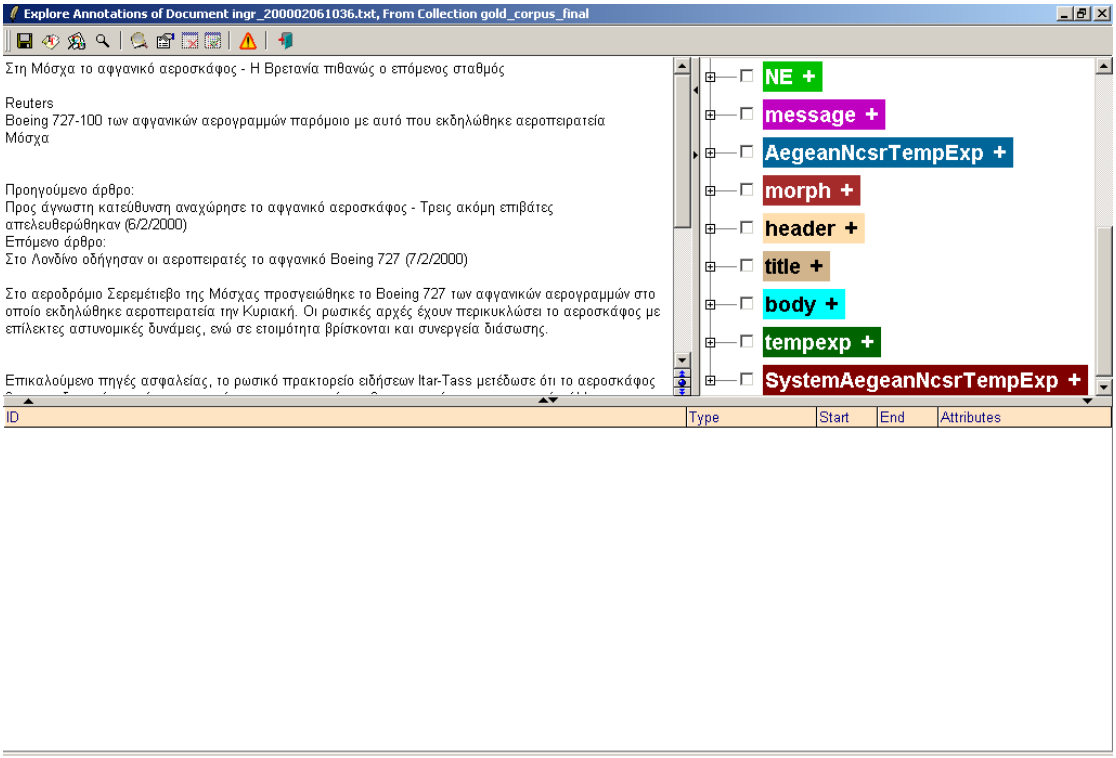
Εικόνα 9 - Το άρθρωμα TemporalExpressions υπό εκτέλεση

Αν ήταν επιτυχής η εκτέλεσή του, αυτό θα μετατραπεί σε κόκκινο χρώμα. Σημειώνεται ότι στην κάτω αριστερή γωνία εμφανίζεται ο χρόνος (elapsed time) που χρειάστηκε το άρθρωμα προκειμένου να εκτελεστεί (στην περίπτωσή μας χρειάστηκε μόλις ένα δευτερόλεπτο) (βλέπε εικόνα 10). Όπως αναφέρθηκε παραπάνω, αυτό αναγνωρίζει τις χρονικές εκφράσεις σ' ένα έγγραφο και τις κανονικοποιεί προσθέτοντας τις ανάλογες επισημειώσεις σ' αυτό το έγγραφο. Εν τέλει, δεν παραλείπουμε να αποθηκεύσουμε τη συλλογή μας πατώντας το κουμπί με τη δισκέτα στη γραμμή εργαλείων.



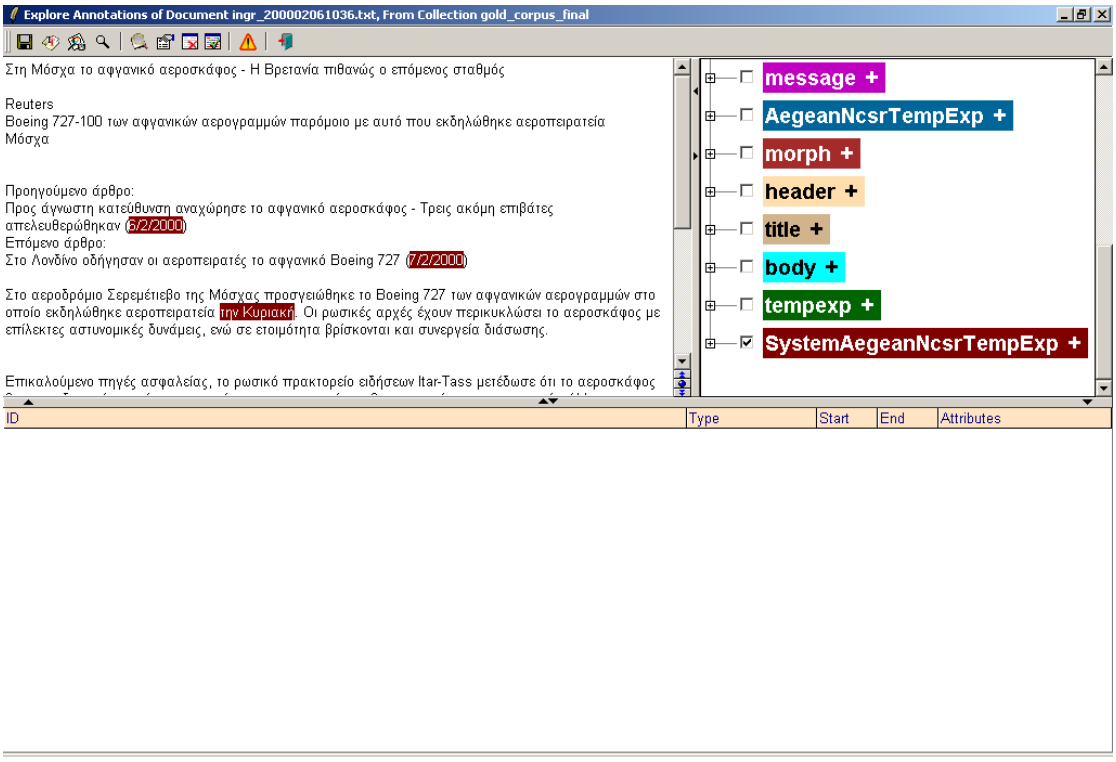
Εικόνα 10 - Επιτυχής εκτέλεση του αρθρώματος TemporalExpressions

Αν τώρα κάνουμε κλικ πάνω στο εκτελεσμένο άρθρωμα, τότε θα εμφανιστεί ένα παράθυρο, παρακινώντας μας να επιλέξουμε ένα απ' όλα τα έγγραφα, αν έχουμε εκτελέσει όλη τη συλλογή μας. Από την στιγμή που το επιλέγουμε εμφανίζεται ένα αναδυόμενο μενού (pop-up menu) που μας ρωτάει να επιλέξουμε το πρόγραμμα παρουσίασης (viewer) με το οποίο επιθυμούμε να δούμε γραφικά τις επισημειώσεις (αυτό μπορεί να είναι και μοναδικό) (προφανώς αν έχουμε εκτελέσει το άρθρωμά μας πάνω σ' ένα μόνο έγγραφο, δε θα χρειαστεί να επιλέξουμε το έγγραφο για το οποίο θα εμφανιστεί το πρόγραμμα παρουσίασης, αλλά κατευθείαν θα εμφανιστεί το αναδυόμενο μενού με τη λίστα από προγράμματα παρουσίασης). Αν επιλέξουμε το πρόγραμμα παρουσίασης "Explore Annotations", τότε θα εμφανιστεί το παρακάτω παράθυρο (βλέπε εικόνα 11) (εναλλακτικά θα μπορούσαμε να επιλέξουμε κατευθείαν το κουμπί με το πηδάλιο από τη γραμμή εργαλείων του συστήματός μας, AegeanNcsrTemporalExpressions):



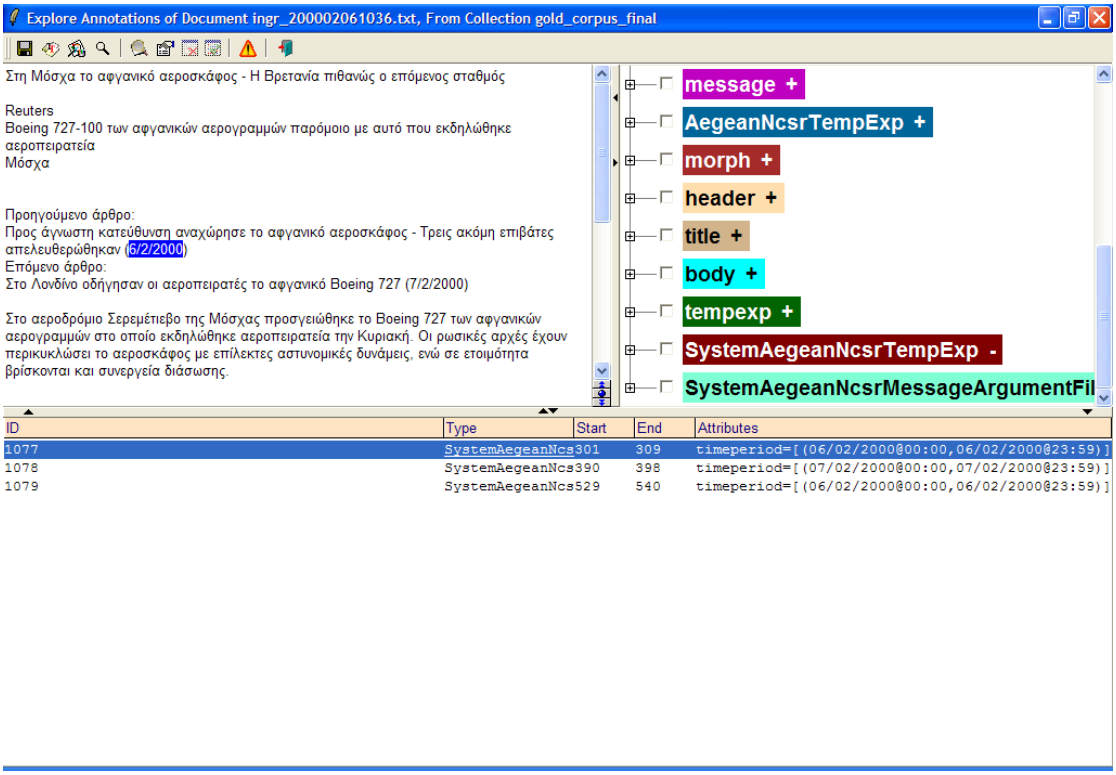
Εικόνα 11 - Το πρόγραμμα παρουσίασης "Exprore Annotations"

Αν τώρα επιλέξουμε το πλαίσιο ελέγχου της χρωματισμένης επισημείωσης SystemAegeanNcsrTempExp θα παρατηρήσουμε τις χρονικές εκφράσεις στο κείμενο να χρωματίζονται με χρώμα ανάλογο αυτού της επισημείωσης, όπως φαίνεται στην παρακάτω εικόνα 12:



Εικόνα 12 – Οι χρονικές εκφράσεις

Αν τώρα πατήσουμε και το σύμβολο + που βρίσκεται δίπλα στην επισημείωση με όνομα SystemAegeanNcsrTempExp, τότε θα μπορούμε να δούμε και όλες τις επισημειώσεις τύπου SystemAegeanNcsrTempExp στο κείμενό μας, όπως φαίνεται στην παρακάτω εικόνα 13 (στην περίπτωσή μας, κάθε φορά που επιλέγουμε μια επισημείωση η αντίστοιχη χρονική έκφραση στο κείμενό μας χρωματίζεται με μπλε χρώμα):

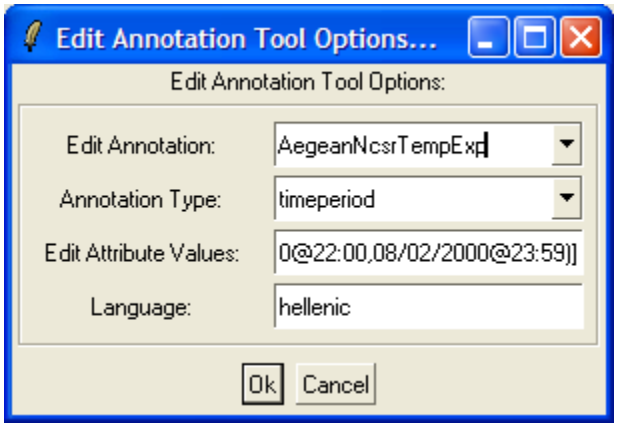


Εικόνα 13 – Εμφάνιση επισημειώσεων

Όπως φαίνεται στο πρόγραμμα παρουσίασης παραπάνω, η επισημείωση που έχουμε επιλέξει έχει αναγνωριστικό (identification – id) 1077, είναι τύπου SystemAegeanNcsrTempExp, τα όρια της έκτασής (span) της είναι 301και 309 και έχει μια ιδιότητα τύπου timeperiod με τιμή (value) [(06/02/2000@00:00,06/02/2000@23:59)]. Η επισημείωση αυτή πραγματοποιήθηκε για τη χρονική έκφραση «6/2/2000» στο κείμενό μας.

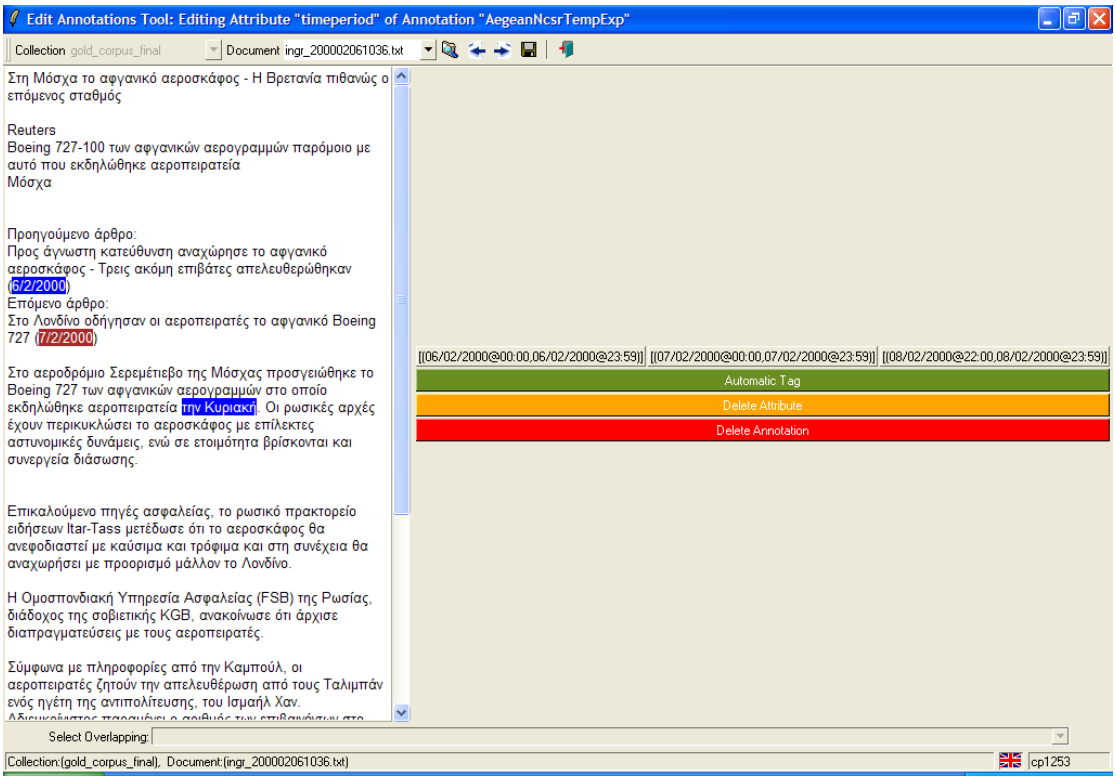
7.2.1.2 Διαδικασία χειρωνακτικής επισημείωσης

Η διαδικασία της χειρωνακτικής επισημείωσης που ακολουθήθηκε περιγράφεται παρακάτω:
Μέσα από ένα παράθυρο ενός συστήματος που έχουμε δημιουργήσει (στην περίπτωση μας από το παράθυρο του συστήματος AegeanNcsrTemporalExpressions) και αφού έχουμε φορτώσει τη συλλογή μας, από το μενού "utilities" επιλέγουμε την επιλογή "Annotation Tools" και από εκεί την επιλογή "Tool for editing Annotations". Στο παράθυρο "Edit Annotation Tool Options" στο πτυσσόμενο πλαίσιο συνδυασμού (Drop-Down Combination Box) "Edit Annotation:" τοποθετούμε το τύπο της επισημείωσης (στην περίπτωση μας τοποθετούμε AegeanNcsrTempExp), στο πτυσσόμενο πλαίσιο συνδυασμού "Annotation Type:" τον τύπο της ιδιότητας (στην περίπτωσή μας τοποθετούμε timeperiod), στο πεδίο κειμένου "Edit Attribute Values" τοποθετούμε την τιμή της ιδιότητας (στην περίπτωσή μας π.χ. [(08/02/2000@22:00,08/02/2000@23:59)] και στο πεδίο "Language" τοποθετούμε τη γλώσσα του κειμένου μας (στην περίπτωση μας τοποθετούμε hellenic) (βλέπε εικόνα 14).



Εικόνα 14 –Χειρωνακτική εισαγωγή επισημειώσεων

Αφού πατήσουμε "Ok" θα εμφανιστεί το παρακάτω παράθυρο (βλέπε εικόνα 15):

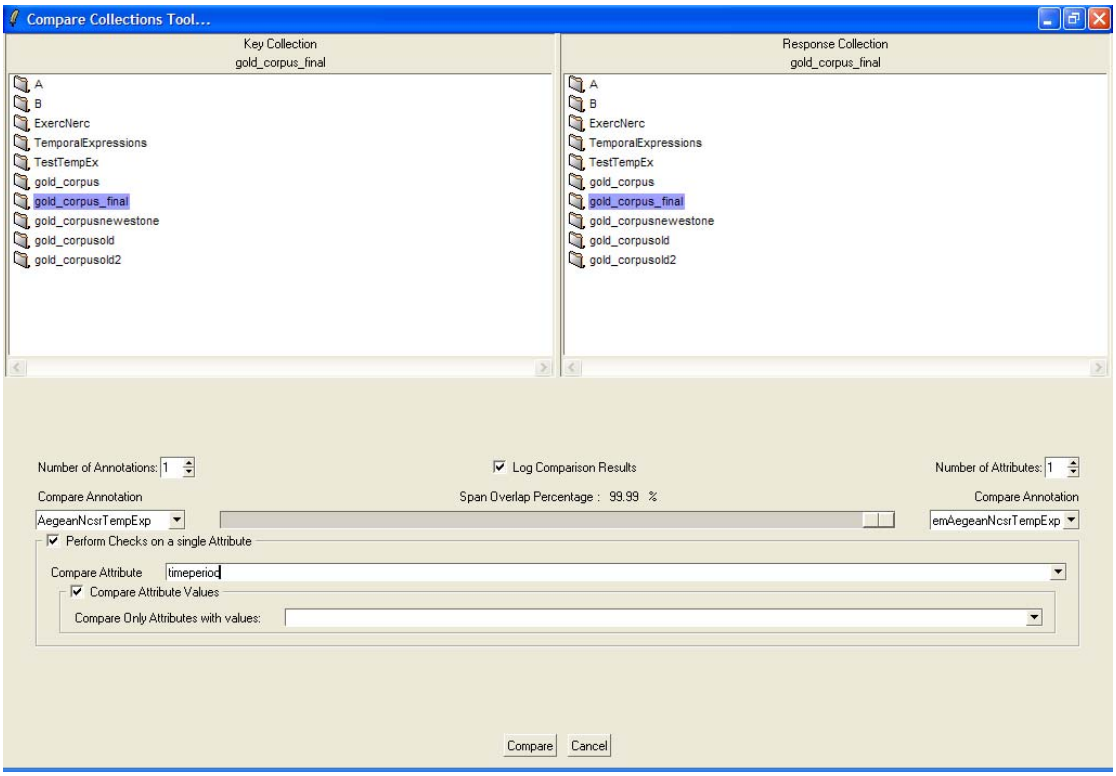


Εικόνα 15 – Χειρωνακτική επισημείωση χρονικών εκφράσεων με το πάτημα ενός κουμπιού

Στο προκείμενο παράθυρο βλέπουμε ήδη κάποιες επισημειωμένες χειρωνακτικά χρονικές εκφράσεις. Είναι αυτές που είναι χρωματισμένες. Σε περίπτωση που είναι χρωματισμένες με το ίδιο χρώμα, αυτό υποδεικνύει έμμεσα στη συγκεκριμένη περίπτωση ότι έχουν κανονικοποιηθεί με τον ίδιο τρόπο, σε διαφορετική περίπτωση ανόμοια. Αυτό που έχουμε να κάνουμε τώρα είναι αφότου επιλέξουμε τη χρονική έκφραση που θέλουμε να επισημειώσουμε να πατήσουμε το ανάλογο κουμπί, ώστε να επισημειωθεί σωστά (στην περίπτωση μας θα μπορούσαμε να επιλέξουμε το κουμπί [(06/02/2000@22:00,06/02/2000@23:59)] ώστε να επισημειώσουμε τη χρονική έκφραση «Την Κυριακή» - εδώ βέβαια είναι ήδη επισημειώμενη). Αφότου την επισημειώσουμε, κάνουμε "Save" (επιλέγουμε το σύμβολο της δισκέτας από τη γραμμή εργαλείων).

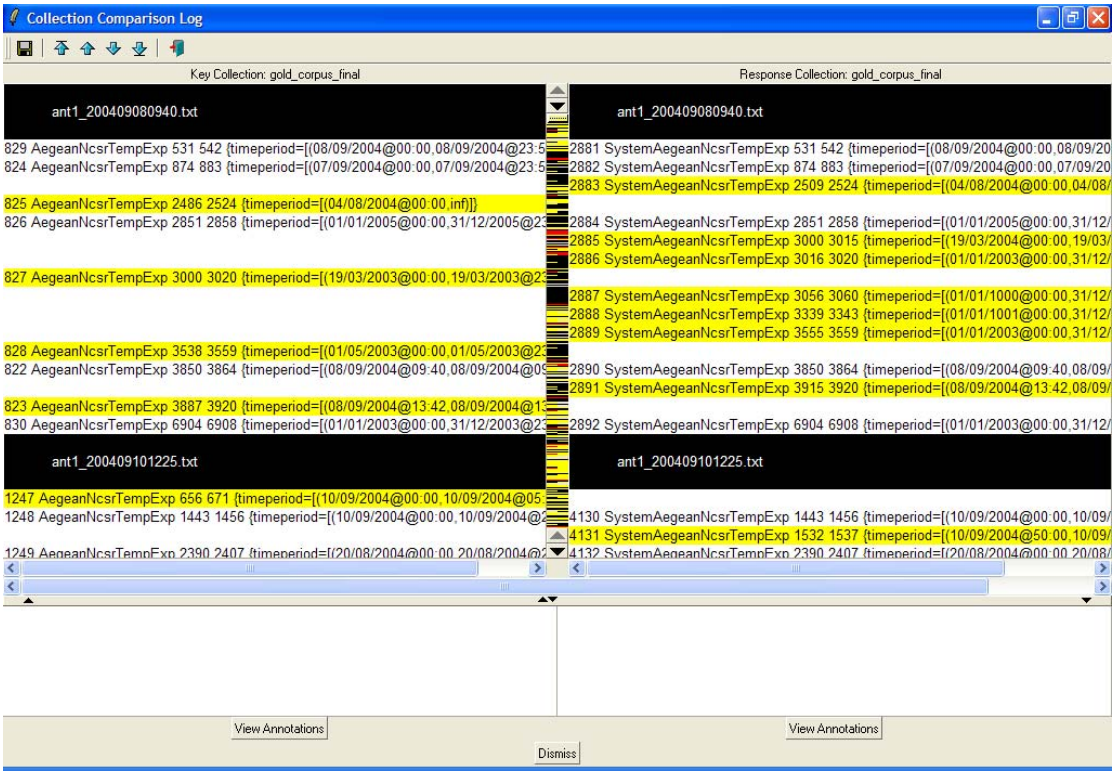
7.2.1.3 Μεθοδολογία αξιολόγησης

Μέσα από το βασικό παράθυρο της πλατφόρμας Επεξεργασίας Φυσικής Γλώσσας Ellogon επιλέγουμε από το μενού "Utilities" την επιλογή "Comparisons" και από εκεί την επιλογή "Collection Comparison Tool". Στο παράθυρο που εμφανίζεται επιλέγουμε στο αριστερό μέρος τη συλλογή που έχει επισημειωθεί χειρωνακτικά (συλλογή αναφοράς) και στο δεξιό μέρος τη συλλογή που έχει επισημειωθεί αυτόματα από το σύστημα (συλλογή αποτίμησης). Στο αριστερό στο πτυσσόμενο πλαίσιο συνδυασμού (Drop-Down Combination Box) "Compare Annotation" τοποθετούμε τον τύπο της χειρωνακτικής επισημείωσης (στην περίπτωσή μας AegeanNcsrTempExp) και στο δεξιό πτυσσόμενο πλαίσιο συνδυασμού "Compare Annotation" τοποθετούμε τον τύπο της επισημείωσης του συστήματος (στην περίπτωσή μας SystemAegeanNcsrTempExp). Έπειτα τσεκάρουμε το πλαίσιο ελέγχου "Performs Checks on a single attribute" και στο πεδίο κειμένου "Compare Attributes" τοποθετούμε τον κοινό τύπο των ιδιοτήτων που θέλουμε να συγκρίνουμε (στην περίπτωση μας τοποθετούμε timeperiod που είναι κοινή ιδιότητα της χειρωνακτικής επισημείωσης AegeanNcsrTempExp και της επισημείωσης του συστήματος SystemAegeanNcsrTempExp). Τέλος τσεκάρουμε το πλαίσιο ελέγχου "Compare Attribute Values" και πατάμε το κουμπί "Compare" (βλέπε την παρακάτω εικόνα 16). Για την περίπτωση όπου θέλουμε να συγκρίνουμε μόνο τις επισημειώσεις, απλώς δεν τσεκάρουμε το πλαίσιο ελέγχου "Performs Checks on a single attribute".

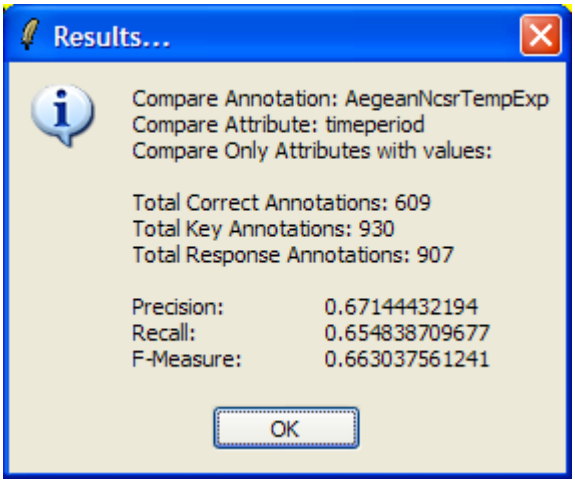


Εικόνα 16 – Σύγκριση τιμής ιδιότητας χειρωνακτικών επισημειώσεων με τιμή ιδιότητας επισημειώσεων που παράγονται αυτόματα από το σύστημα για το άρθρωμα των χρονικών εκφράσεων

Το αποτέλεσμα απεικονίζεται στο παράθυρο της Εικόνας 17. Στο μικρότερο παράθυρο (βλέπε εικόνα 18) απεικονίζονται οι παράμετροι της σύγκρισης, και τα αποτελέσματα της αξιολόγησης σύμφωνα με τα μέτρα Precision, Recall, F-measure.



Εικόνα 17 – Αποτελέσματα Σύγκρισης



Εικόνα 18 – Παράμετροι της σύγκρισης και αποτελέσματα της αξιολόγησης

Τα αποτελέσματα τα οποία θα μας επιστράφηκαν αφορούν στα:

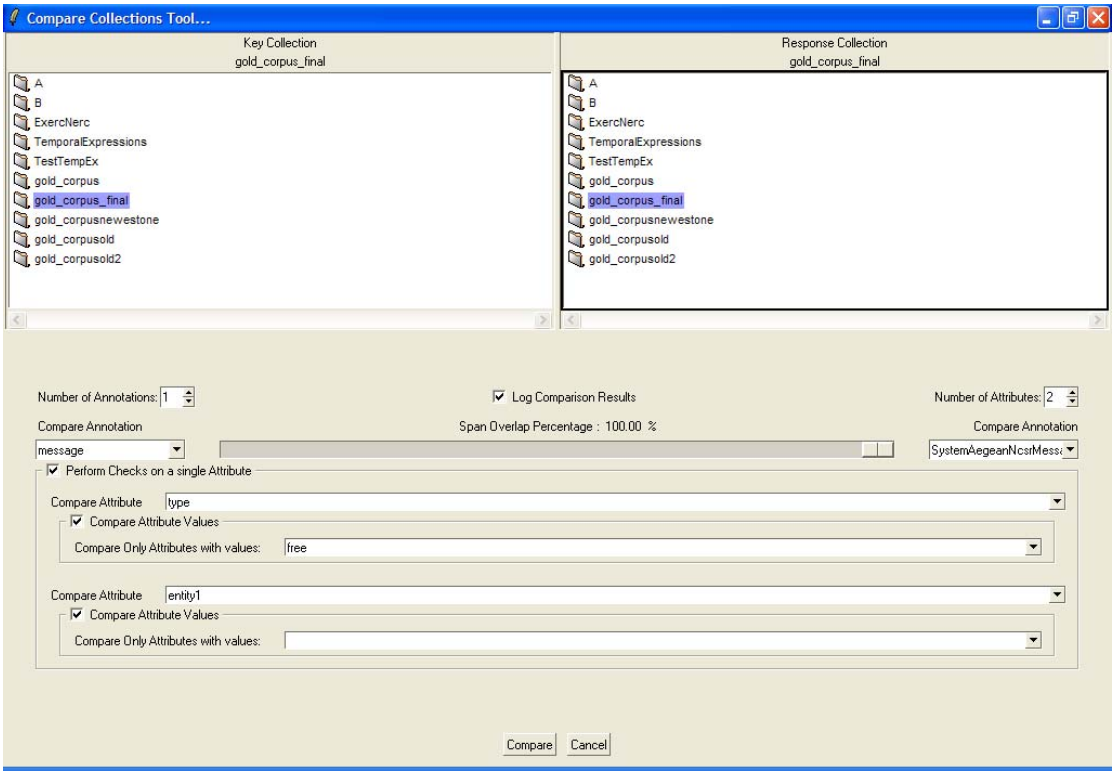
- **Total Correct Annotations**, δηλαδή πόσες επισημειώσεις απ' αυτές που βρήκε το σύστημα ήταν σωστές σε σχέση με αυτές που επισημειώθηκαν
- **Total Key Annotations**, δηλαδή το σύνολο των επισημειώσεων οι οποίες επισημειώθηκαν χειρωνακτικά
- **Total Response Annotations**, δηλαδή το σύνολο των επισημειώσεων οι οποίες βρέθηκαν αυτόματα από το σύστημα
- **Precision (Ακρίβεια)**, δηλαδή το λόγο των Total Correct Annotations προς τα Total Response Annotations (Precision = Total Correct Annotations / Total Response Annotations) και με άλλα λόγια πόσα σωστά βρέθηκαν ως προς αυτά που βρήκε το σύστημα

- **Recall (Ανάκληση)**, δηλαδή το λόγο των Total Correct Annotations προς τα Total Key Annotations (Recall = Total Correct Annotations / Total Key Annotations) και με άλλα λόγια πόσα σωστά βρέθηκαν ως προς αυτά που έπρεπε να βρεθούν (σύμφωνα με τις χειρωνακτικές επισημειώσεις)
- **F-measure = 2 * Precision * Recall / Precision + Recall**

7.2.2 Μηνύματα

Την ίδια περίπου διαδικασία ακολουθούμε και για τα μηνύματα, στη φάση όμως της αξιολόγησης πραγματοποιούμε συγκεκριμένα τα παρακάτω:

Καταρχήν αφήνουμε το αριστερό spin box “Number of Annotations” όπως έχει, δηλαδή με τον αριθμό 1, ενώ στο δεξιό spin box “Number of Attributes” τοποθετούμε τον αριθμό 2 (ή παραπάνω ανάλογα με τον αριθμό των ιδιοτήτων που θέλουμε να συγκρίνουμε). Στο αριστερό πτυσσόμενο πλαίσιο συνδυασμού (Drop-Down Combination Box) “Compare Annotation” τοποθετούμε τον τύπο της χειρωνακτικής επισημείωσης (στην περίπτωσή μας message) και στο δεξιό πτυσσόμενο πλαίσιο συνδυασμού “Compare Annotation” τοποθετούμε τον τύπο της επισημείωσης του συστήματος (στην περίπτωσή μας SystemAegeanNcsrMessageArgumentFilling). Έπειτα τσεκάρουμε το πλαίσιο ελέγχου “Performs Checks on a single attribute” και στο πτυσσόμενο πλαίσιο λίστας (Drop-Down List Box) “Compare Attributes” τοποθετούμε τον κοινό τύπο των ιδιοτήτων που θέλουμε να συγκρίνουμε (στην περίπτωση μας τοποθετούμε type). Μετά τσεκάρουμε το πλαίσιο ελέγχου “Compare Attribute Values” και στο πτυσσόμενο πλαίσιο λίστας “Compare Only Attributes with values” τοποθετούμε τον τύπο του μηνύματος, free (δηλαδή την τιμή της ιδιότητας type της επισημείωσης SystemAegeanNcsrMessageArgumentFilling). Αφετέρου στο δεύτερο πτυσσόμενο πλαίσιο λίστας “Compare Attributes” τοποθετούμε πάλι τον κοινό τύπο των ιδιοτήτων που θέλουμε να συγκρίνουμε (αυτή την φορά τοποθετούμε entity1) και δεν παραλείπουμε να τσεκάρουμε το δεύτερο πλαίσιο ελέγχου “Compare Attribute Values”. Αν θέλουμε να τσεκάρουμε και άλλη ιδιότητα πράττουμε ομοίως αυξάνοντας όμως πρώτα τον αριθμό του spin box “Number of Attributes” σε τρία. Με άλλα λόγια ρυθμίζουμε τον αριθμό του spin box ανάλογα με το πόσες ιδιότητες θέλουμε να συγκρίνουμε (βλέπε εικόνα 19). Εν τέλει πατάμε το κουμπί “Compare”.



Εικόνα 19 - Σύγκριση τιμής ιδιότητας χειρωνακτικών επισημειώσεων με τιμή ιδιότητας επισημειώσεων που παράγονται αυτόματα από το σύστημα για το άρθρωμα των μηνυμάτων

7.3 Χρονικές Εκφράσεις

7.3.1 Τρόποι και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία

Ακολουθούν παραδείγματα με τις κατηγορίες των χρονικών εκφράσεων και τον απαιτούμενο τρόπο κανονικοποιήσεώς τους καθώς και τρόποι επίλυσης κάποιων προβλημάτων που εμφανίζονται σε ορισμένες κατηγορίες. Χρίζει αναφοράς και το γεγονός ότι έχουν προβλεφτεί προβλήματα όσον αφορά στην ευαισθησία σε μικρό – κεφαλαίο (case sensitive) του πρώτου γράμματος μιας χρονικής έκφρασης, πράγμα το οποίο σημαίνει εν ολίγοις ότι αναγνωρίζεται τόσο η χρονική εκφράση «την Κυριακή» όσο και η χρονική έκφραση «Την Κυριακή» και προφανώς κανονικοποιούνται και με τον ίδιο τρόπο.

7.3.1.1 Χρονικές εκφράσεις που αναφέρονται σε ώρες

Οι χρονικές εκφράσεις που αναφέρονται σε ώρες πρέπει να μετατρέπονται σε εικοσιτετράωρη μορφή ωω:λλ (π.χ. η 1 αναπαριστάται ως 13:00). Αν το κείμενο δεν προσδιορίζει την ημερομηνία στην οποία αναφέρεται ένας προσδιορισμός της ώρας, ως ημερομηνία λαμβάνεται η ημερομηνία της είδησης. Έστω ότι η ημερομηνία δημοσίευσης της είδησης είναι η 6/2/2000. Ακολουθούν παραδείγματα χρονικών εκφράσεων που αναφέρονται σε ώρες και ο τρόπος κανονικοποίησής τους:

στις 22:45 (06/02/2000@22:45, 06/02/2000@22:45), ως ημερομηνία μπαίνει η ημερομηνία δημοσίευσης της είδησης

στις 22.45 (06/02/2000@22:45, 06/02/2000@22:45), ως ημερομηνία μπαίνει η ημερομηνία δημοσίευσης της είδησης

7.3.1.2 Χρονικές εκφράσεις που αναφέρονται σε ημέρες

Όταν έχουμε αναφορά σε συγκεκριμένη μέρα, π.χ. «την Κυριακή», τότε ως ημέρα τοποθετείται στην κανονικοποίηση η ημέρα της δημοσίευσης της είδησης, αν η ημερομηνία δημοσίευσης της είδησης ταυτίζεται με την ημερομηνία που αντιστοιχεί στη χρονική έκφραση που αναφέρεται στο κείμενό μας (στην περίπτωσή μας «την Κυριακή»), διαφορετικά τοποθετείται το αποτέλεσμα που προκύπτει αν αφαιρέσουμε από την ημέρα που αντιστοιχεί στην ημερομηνία δημοσίευσης της είδησης τόσες μέρες όσες είναι η διαφορά σε ημέρες μεταξύ της ημέρας που αντιστοιχεί στην ημερομηνία δημοσίευσης της είδησης και της αναφερθείσας ημέρας στο κείμενο. Ο μήνας και το έτος λαμβάνονται από την ημερομηνία δημοσίευσης της είδησης και για την ώρα θεωρούμε ότι μια μέρα ξεκινάει στις 12:00 π.μ. και τελειώνει στις 11:59 μ.μ..

Έστω ότι η ημερομηνία δημοσίευσης της είδησης είναι η 6/2/2000. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αναφέρονται σε συγκεκριμένη ημέρα και ο τρόπος κανονικοποίησής τους:

την Κυριακή (06/02/2000@00:00, 06/02/2000@23:59), η ημερομηνία δημοσίευσης της είδησης ταυτίζεται με την ημερομηνία που αντιστοιχεί στην ημέρα Κυριακή

εντός της Κυριακής (06/02/2000@00:00, 06/02/2000@23:59)

από την Κυριακή (06/02/2000@00:00, inf)

το Σάββατο (05/02/2000@00:00, 05/02/2000@23:59), δεδομένου ότι η ημερομηνία δημοσίευσης της είδησης αντιστοιχεί στην ημέρα «Κυριακή» και στο κείμενο μας αναφέρεται η χρονική έκφραση «το Σάββατο», αφαιρείται μία ημέρα

Στην περίπτωση που έχουμε αναφορά σε περασμένες ή επόμενες μέρες τότε αφαιρούνται ή προστίθενται στην ημερομηνία της είδησης τόσες μέρες ανάλογα με τη χρονική έκφραση που έχουμε στο κείμενό μας.

Έστω ότι η ημερομηνία δημοσίευσης της είδησης είναι η προαναφερόμενη. Παρακάτω, παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αναφέρονται ομοίως σε συγκεκριμένη ημέρα και ο τρόπος κανονικοποίησής τους:

το περασμένο Σάββατο (05/02/2000@00:00, 05/02/2000@23:59)

το επόμενο Σάββατο (12/02/2000@00:00, 12/02/2000@23:59)

Αν η αναφορά στην ημέρα δεν είναι άμεση, αλλά έμμεση τότε και πάλι η κανονικοποίηση πραγματοποιείται σε σχέση με την ημερομηνία δημοσίευσης της είδησης. Έστω ότι αυτή είναι πάλι η προαναφερθείσα. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων ανάλογων με αυτή την περίπτωση και ο τρόπος κανονικοποίησής τους:

εντός των επόμενων ημερών (07/02/2000@00:00, inf)

εντός των προσεχών ημερών (07/02/2000@00:00, inf)

εντός της ημέρας (06/02/2000@00:00, 06/02/2000@23:59)

7.3.1.3 Χρονικές εκφράσεις που αναφέρονται σε χρονικό διάστημα ημέρας

Για τα χρονικά διαστήματα ημέρας θεωρούμε τις παρακάτω παραδοχές:

- ✓ Τα ξημερώματα ξεκινούν στις 3:00 π.μ. και τελειώνουν στις 5:59 π.μ.
- ✓ Το πρωί ξεκινάει στις 6:00 π.μ. και τελειώνει στις 11:59 π.μ.
- ✓ Το μεσημέρι ξεκινάει στις 12:00 μ.μ. και τελειώνει στις 15:59 μ.μ.
- ✓ Το απόγευμα ξεκινάει στις 16:00 μ.μ. και τελειώνει στις 19:59 μ.μ.
- ✓ Το βράδυ και η νύχτα ξεκινούν στις 20:00 μ.μ. και τελειώνουν στις 23:59 μ.μ.
- ✓ Το νωρίς τα ξημερώματα θεωρείται ότι είναι από τις 3:00 π.μ. μέχρι και τις 3:59 π.μ.
- ✓ Τα αργά τα ξημερώματα θεωρείται ότι είναι από τις 5:00 π.μ. μέχρι και τις 5:59 π.μ.
- ✓ Το νωρίς το πρωί θεωρείται ότι είναι από τις 6:00 π.μ. μέχρι και τις 7:59 π.μ.
- ✓ Το αργά το πρωί θεωρείται ότι είναι από τις 10:00 π.μ. μέχρι και τις 11:59 π.μ.
- ✓ Το νωρίς το μεσημέρι θεωρείται ότι είναι από τις 12:00 μ.μ. μέχρι και τις 13:59 μ.μ.
- ✓ Το αργά το μεσημέρι θεωρείται ότι είναι από τις 14:00 μ.μ. μέχρι και τις 15:59 μ.μ.
- ✓ Το νωρίς το απόγευμα θεωρείται ότι είναι από τις 16:00 μ.μ. μέχρι και τις 17:59 μ.μ.
- ✓ Το αργά το απόγευμα θεωρείται ότι είναι από τις 18:00 μ.μ. μέχρι και τις 19:59 μ.μ.
- ✓ Το νωρίς το βράδυ και το νωρίς τη νύχτα θεωρούνται ότι είναι από τις 20:00 μ.μ. μέχρι και τις 21:59 μ.μ.
- ✓ Το αργά το βράδυ και το αργά τη νύχτα θεωρούνται ότι είναι από τις 22:00 μ.μ. μέχρι και τις 23:59 μ.μ.

Ως ημερομηνία τοποθετείται φυσικά η ημερομηνία δημοσίευσης της είδησης. Έστω ότι αυτή είναι η 6/2/2000. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αναφέρονται σε χρονικό διάστημα ημέρας και ο τρόπος κανονικοποίησής τους:

τα ξημερώματα (06/02/2000@03:00, 06/02/2000@05:59)

το πρωί (06/02/2000@06:00, 06/02/2000@11:59)

το μεσημέρι (06/02/2000@12:00, 06/02/2000@15:59)

το απόγευμα (06/02/2000@16:00, 06/02/2000@19:59)

το βράδυ (06/02/2000@20:00, 06/02/2000@23:59)

τη νύχτα (06/02/2000@20:00, 06/02/2000@23:59)

νωρίς το πρωί (06/02/2000@06:00, 06/02/2000@07:59)

αργά το πρωί (06/02/2000@10:00, 06/02/2000@11:59)

νωρίς το μεσημέρι (06/02/2000@12:00, 06/02/2000@13:59)

αργά το μεσημέρι (06/02/2000@14:00, 06/02/2000@15:59)

νωρίς το απόγευμα (06/02/2000@16:00, 06/02/2000@17:59)

αργά το απόγευμα (06/02/2000@18:00, 06/02/2000@19:59)

νωρίς το βράδυ (06/02/2000@20:00, 06/02/2000@21:59)

αργά το βράδυ (06/02/2000@22:00, 06/02/2000@23:59)

7.3.1.4 Χρονικές εκφράσεις που αναφέρονται σε εβδομάδες

Αν έχουμε χρονική έκφραση που αναφέρεται σε εβδομάδα, τότε αντιστοιχίζουμε το ένα άκρο της στην ημερομηνία που αντιστοιχεί στην ημέρα Δευτέρα και το άλλο άκρο της στην ημερομηνία που αντιστοιχεί στην ημέρα Κυριακή (θεωρούμε ότι η εβδομάδα ξεκινάει τη Δευτέρα και τελειώνει την Κυριακή). Αν πρόκειται για περασμένη εβδομάδα, τότε πηγαίνουμε ανάλογα πίσω στο χρόνο, ενώ αν είναι επόμενη πηγαίνουμε ανάλογα μπροστά. Ο μήνας και το έτος λαμβάνονται από την ημερομηνία δημοσίευσης της είδησης. Έστω ότι αυτή είναι η 17/6/2005. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αναφέρονται σε συγκεκριμένη εβδομάδα και ο τρόπος κανονικοποίησής τους:

την περασμένη εβδομάδα (06/06/2005@22:00, 12/06/2005@23:59)

την προηγούμενη εβδομάδα (06/06/2005@22:00, 12/06/2005@23:59)

την επόμενη εβδομάδα (20/06/2005@22:00, 26/06/2005@23:59)

την προσεχή εβδομάδα (20/06/2005@22:00, 26/06/2005@23:59)

7.3.1.5 Χρονικές εκφράσεις που αναφέρονται σε μήνες

Στην περίπτωση που έχουμε αναφορά σε συγκεκριμένο μήνα τοποθετείται ως αρχική ημέρα η πρώτη μέρα του μήνα και ως τελική η τελευταία μέρα αυτού. Ως έτος θεωρείται το έτος που αναγράφεται στην ημερομηνία της είδησης. Αν έχουμε αναφορά σε περασμένο ή επόμενο μήνα τότε αυτός κανονικοποιείται πάντα σε σχέση με το μήνα της ημερομηνίας δημοσίευσης (αφαιρείται ή προστίθεται ανάλογος αριθμός μηνών). Στην περίπτωση που δεν προσδιορίζεται ακριβώς ο μήνας, αλλά εκφράζεται έμμεσα, τότε αφαιρούμε ή προσθέτουμε στον μήνα που λαμβάνουμε από την ημερομηνία της είδησης τόσους μήνες ανάλογα με τη χρονική έκφραση που έχουμε στο κείμενό μας. Έστω ότι η ημερομηνία της είδησης είναι η 30/8/1999. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αναφέρονται σε συγκεκριμένο μήνα και ο τρόπος κανονικοποίησής τους:

για τον Σεπτέμβριο (01/09/1999@00:00, 30/09/1999@23:59)

τον Σεπτέμβριο (01/09/1999@00:00, 30/09/1999@23:59)

τον περασμένο Σεπτέμβριο (01/09/1998@00:00, 30/09/1998@23:59)

τον προηγούμενο Σεπτέμβριο (01/09/1998@00:00, 30/09/1998@23:59)

τον επόμενο Σεπτέμβριο (01/09/1999@00:00, 30/09/1999@23:59)

τον προσεχή Σεπτέμβριο (01/09/1999@00:00, 30/09/1999@23:59)

τον προηγούμενο μήνα (01/07/1999@00:00, 31/07/1999@23:59)

τον επόμενο μήνα (01/09/1999@00:00, 30/09/1999@23:59)

7.3.1.6 Χρονικές εκφράσεις που αναφέρονται σε έτη

7.3.1.6.1 Αναφορά σε έτη

Στην περίπτωση που έχουμε αναφορά σε συγκεκριμένο έτος, τότε ως έναρξη και λήξη αυτού θεωρείται η πρώτη και η τελευταία ημέρα του αντίστοιχα. Στην περίπτωση που δεν προσδιορίζεται ακριβώς το έτος, αλλά εκφράζεται έμμεσα, τότε αφαιρούμε ή προσθέτουμε στο έτος που λαμβάνουμε από την ημερομηνία της είδησης τόσα χρόνια ανάλογα με τη χρονική έκφραση που έχουμε στο κείμενό μας.

Έστω ότι η ημερομηνία της είδησης είναι η 25/5/2004. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αναφέρονται σε συγκεκριμένο έτος και ο τρόπος κανονικοποίησής τους:

έως το 2004 (inf, 01/01/2004@00:00)

μέχρι το 2004 (inf, 01/01/2004@00:00)

το 1999 (01/01/1999@00:00, 31/12/1999@23:59)

μέσα στο 2000 (01/01/2000@00:00, 31/12/2000@23:59)

από το 2004 έως το 2005 (01/01/2004@00:00, 31/12/2005@23:59)

Μεταξύ 1998 και 2000 (01/01/1998@00:00, 31/12/2000@23:59)

Πρόπερσι (01/01/2002@00:00, 31/12/2002@23:59)

Πέρσι (01/01/2003@00:00, 31/12/2003@23:59)

Τον περασμένο χρόνο (01/01/2003@00:00, 31/12/2003@23:59)

Τον προηγούμενο χρόνο (01/01/2003@00:00, 31/12/2003@23:59)

Τον επόμενο χρόνο (01/01/2005@00:00, 31/12/2005@23:59)

Τον προσεχή χρόνο (01/01/2005@00:00, 31/12/2005@23:59)

7.3.1.7 Χρονικές εκφράσεις με τη μορφή ημερομηνιών

Σ' αυτή την περίπτωση ως ημερομηνία τοποθετείται αυτή που λαμβάνεται από τη χρονική έκφραση του κειμένου μας και ως ώρα τα όρια του χρονικού διαστήματος που εκφράζουν την ημέρα. Σε περίπτωση που δεν υπάρχει το έτος στη χρονική έκφραση ως έτος θεωρείται αυτό που αναγράφεται στην ημερομηνία της είδησης. Έστω ότι αυτή είναι η 30/8/1999. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που έχουν τη μορφή ημερομηνιών και ο τρόπος κανονικοποίησής τους:

για τις 15 Απριλίου (15/03/1999@00:00, 15/03/1999@23:59)

στις 10 Φεβρουαρίου (10/02/1999@00:00, 10/02/1999@23:59)

6/2/2000 (06/02/2000@00:00, 06/02/2000@23:59)

7.3.1.8 Χρονικές εκφράσεις με τη μορφή χρονικών επιρρημάτων

Η κανονικοποίηση χρονικών εκφράσεων που έχουν τη μορφή χρονικών επιρρημάτων πραγματοποιείται σε σχέση με την ημερομηνία που αναγράφεται στην είδηση. Έστω ότι αυτή είναι η 20/12/1998. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που έχουν τη μορφή χρονικών επιρρημάτων και ο τρόπος κανονικοποίησής τους:

σήμερα (20/12/1998@00:00, 20/12/1998@23:59)

τώρα (20/12/1998@00:00, 20/12/1998@23:59)

χθες (19/12/1998@00:00, 19/12/1998@23:59)

προχθές (18/12/1998@00:00, 18/12/1998@23:59)

αντιπροχθές (17/12/1998@00:00, 17/12/1998@23:59)

παραπροχθές (17/12/1998@00:00, 17/12/1998@23:59)

αύριο (21/12/1998@00:00, 21/12/1998@23:59)

μεθαύριο (22/12/1998@00:00, 22/12/1998@23:59)

από σήμερα (20/12/1998@00:00, inf)

έως σήμερα (inf, 20/12/1998@00:00)

μέχρι σήμερα (inf, 20/12/1998@00:00)

7.3.1.9 Χρονικές εκφράσεις με τη μορφή επιθετικών προσδιορισμών συγκεκριμένης χρονικής διάρκειας

Είναι απαραίτητη η κανονικοποίηση χρονικών επιθέτων που αναφέρονται σε καθορισμένο χρονικό διάστημα σε κάθε είδηση. Αυτά είναι επίθετα όπως χθεσινός, αυριανός, κτλ. Ακολουθούν παραδείγματα χρονικών εκφράσεων που έχουν τη μορφή επιθετικών προσδιορισμών συγκεκριμένης χρονικής διάρκειας και ο προτεινόμενος τρόπος κανονικοποίησής τους:

χθεσινή(30/12/1998@00:00, 30/12/1998@23:59)

σημερινή (02/02/1999@00:00, 02/02/1999@23:59)

7.3.1.10 Ανάμεικτες (miscellaneous) χρονικές εκφράσεις

Αν η είδηση δημοσιεύτηκε στις 31/12/1998 στις 16:30 ώρα Ελλάδος τότε οι παρακάτω χρονικές εκφράσεις κανονικοποιούνται ως εξής:

για την ώρα (31/12/1998@16:30, 31/12/1998@16:30)

αυτή την ώρα (31/12/1998@16:30, 31/12/1998@16:30)

επί του παρόντος (31/12/1998@16:30, 31/12/1998@16:30)

προς το παρόν (31/12/1998@16:30, 31/12/1998@16:30)

αυτή την ώρα (31/12/1998@16:30, 31/12/1998@16:30)

αυτή την στιγμή (31/12/1998@16:30, 31/12/1998@16:30)

μέχρι στιγμής (inf, 31/12/1998@16:30

προς στιγμήν (inf, 31/12/1998@16:30)

Οι παραπάνω χρονικές εκφράσεις λαμβάνουν υπόψη την ώρα δημοσίευσης της είδησης. Αντιθέτως οι παρακάτω δεν τη λαμβάνουν:

την ίδια ώρα (31/12/1998@00:00, 31/12/1998@23:59)

την ίδια στιγμή (31/12/1998@00:00, 31/12/1998@23:59)

7.3.1.11 Συνδυασμός των παραπάνω χρονικών εκφράσεων

Οι χρονικές εκφράσεις που αποτελούν κάποιο συνδυασμό άλλων χρονικών εκφράσεων κανονικοποιούνται σύμφωνα με τις προαναφερθείσες προδιαγραφές. Έστω ότι η ημερομηνία δημοσίευσης της είδησης είναι η 6/2/2000. Παρακάτω παρουσιάζονται παραδείγματα χρονικών εκφράσεων που αποτελούν κάποιο συνδυασμό άλλων χρονικών εκφράσεων και ο τρόπος κανονικοποίησής τους:

25/11 | 16:45 (25/11/1998@16:45, 25/11/1998@16:45), ως έτος λαμβάνεται το έτος δημοσίευσης της είδησης

Κυριακή 6 Φεβρουαρίου 2000 – 14:30 (06/02/2000@14:30, 06/02/2000@14:30)

στις 6 το απόγευμα (06/02/2000@18:00, 06/02/2000@18:00), ως ημερομηνία μπαίνει η ημερομηνία δημοσίευσης της είδησης

από το πρωί της Κυριακής (06/02/2000@06:00, inf)

έως το βράδυ της Κυριακής (inf, 06/02/2000@23:59)

τα ξημερώματα της Κυριακής (06/02/2000@03:00, 06/02/2000@05:59)

νωρίς το πρωί της Κυριακής (06/02/2000@06:00, 06/02/2000@07:59)

αργά το βράδυ της Κυριακής (06/02/2000@22:00, 06/02/2000@23:59)

7.3.2 Τρόπος υλοποίησης του αρθρώματος των χρονικών εκφράσεων

Εν συντομία τα βήματα του αλγορίθμου σε πολύ αφαιρετικό επίπεδο που επινοήθηκε και ονομάζεται SystemAegeanTempExp είναι τα εξής:

1. Πάρε το κείμενο από το κάθε κείμενο της συλλογής σε μορφή ακατέργαστων δεδομένων (raw data)
2. Μετέτρεψέ το σε συμβολοσειρά (String)
3. Μετέτρεψε τη συμβολοσειρά σε ακολουθία χαρακτήρων (Charsequence)
4. Δημιούργησε τις μορφές (patterns) των κανονικών εκφράσεων για τις χρονικές εκφράσεις σε μορφή συμβολοσειράς
5. Μεταγλώττισε τις κανονικές εκφράσεις
6. Δημιούργησε Matcher που αντιστοιχεί τη μορφή (pattern) των κανονικών εκφράσεων με το κείμενο της συλλογής μας
7. Για όσο πραγματοποιείται ταίριασμα της μορφής (pattern) των κανονικών εκφράσεων που δημιουργήσαμε με κάποια επόμενη χρονική έκφραση (υποτίθεται ότι έχει προηγηθεί ένα ταίριασμα)
   7.1. Δημιούργησε τα όρια της χρονικής έκφρασης που ταιριάστηκε
   7.2. Αν η χρονική έκφραση που ταιριάστηκε με χρήση του Matcher ταιριάζει με κάποια μορφή (pattern) κανονικής έκφρασης ή κάποιων κανονικών εκφράσεων τότε:
      7.2.1. Διάβασε (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία)) το όνομα του κειμένου της συλλογής
      7.2.2. Δημιούργησε (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία)) τέτοια κανονική έκφραση ώστε να μπορέσεις να ταιριάξεις την ημερομηνία και την ώρα δημοσίευσης της είδησης του κειμένου που περιέχονται στο όνομα του αρχείου
      7.2.3. Αν μπορέσεις και τα ταιριάξεις (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία)) τότε:
         7.2.3.1. Αφού πάρεις όσες πληροφορίες χρειάζεσαι κάθε φορά (π.χ. μέρα, μήνας, έτος, ώρα) (αν χρειάζεται – εξαρτάται από τη χρονική έκφραση που ταιριάστηκε (βλέπε Τρόπους και παραδείγματα κανονικοποίησεως των χρονικών εκφράσεων κατά κατηγορία)) δημιούργησε επισημείωση τύπου SystemAegeanNcsrTempExp με ιδιότητα τύπου timeperiod με τιμή την κατάλληλη και αντίστοιχη κανονικοποιημένη τιμή της χρονικής έκφρασης

7.4 Μηνύματα

7.4.1 Οντολογία των μηνυμάτων

Οι τύποι των ορισμάτων των μηνυμάτων στηρίζονται στην παρακάτω οντολογία:

- Persons
  - Offender
  - Hostage
  - Rescue Team
  - Governmental Executives
  - Relatives
  - Professional
  - Demonstrators

- Place
  - Location of Conduct
  - Country
  - City

- Temporal Concept

- Year
- Date
- Hour
- Part of Day
  - Dawn
  - Evening

- Activities
  - Arrest
  - Ask
  - Attack
  - Deny
  - End
  - Enter
  - Free_entity
  - Give
  - Hijack
  - Hurt
  - Keep hostages
  - Kidnap
  - Kill
  - Negotiate
  - Not_ recall_army
  - Recall_army
  - Return
  - Satisfy_demands

- Armament
  - Explosive
  - Gas
  - Gun
  - Tank

- Vehicle
  - Bus
  - Plane
  - Car
  - Lorry

- Physical Condition
  - Hunger
  - Thirst
  - Exhaustion
  - Frightened
  - Sick
  - Alife
  - Good
  - Mediocre
  - Bad

- Media
  - Newspaper/Press
  - Radio
  - Internet

- T.V.

- Quantity
  - Cardinal
  - Quantitative
    - Some
    - Many
    - Few

- Amount
  - (integer)
  - value (A) {Euro, Dollar, drachmas}

- Public Institution

7.4.2 Τύποι των μηνυμάτων

Έχουμε τους εξής τύπους μηνυμάτων:

free, hold, kill, ask_for, announce, kidnap, deny, arrive, physical_condition, transport, negotiate, enter, threaten, work_for, meet, speak_on_the_phone, return, hospitalized, escape_from, help, head_towards, encircle, arrest, armed, leave, take_control_of, give_deadline, located, inform, end, organize, aim_at, hijack, trade, prevent_from, lead, take_on_responsibility, put, stay_parked, pay_ransom, accept, block_the_way, assure, be_afraid, interrogate, start, explode, give_asylum

7.4.3 Προδιαγραφές των μηνυμάτων

Οι προδιαγραφές των μηνυμάτων περιγράφονται παρακάτω:

- accept (entity 1, entity 2): sb accepts sb else or sth
  - entity 1: persons
  - entity 2: persons or activity
  - Constraint: (entity 1 = {offender, hostage})
- aim_at (entity 1, entity 2, means): message with double meaning:
  - sb aims at sb using a weapon
  - sb aims at doing sth
    - entity 1: persons or public institution
    - entity 2: persons or activity
    - means: armament
    - Constraints:
      - if (entity 1 = offender) then (entity 2 = rescue team, hostage, activity)
      - if (entity 1 = public institution) then (entity 2 = activity)
- announce (entity, activity): sb announces an activity
  - entity: person or media or public institution
  - activity: activity
  - Constraint: (entity = {public institution, rescue team, Professional, offenders})
- armed (entity, means): a person is armed with some kind of weapon or a vehicle is "armed" with people
  - entity: persons or vehicle
  - means: armament or persons

    - Constraints:
        - if (entity = {offender or rescue team}) then (means = {explosives, guns})
        - if (entity = vehicle) then (means = {rescue team})
- arrest (entity 1, entity 2, quantity, for_reason): sb arrests sb else for a specific reason
    - entity 1: persons or public institutions
    - entity 2: persons
    - quantity: cardinal or quantitative
    - for_reason: activity
    - Constraints:
        - if (entity 1 = {rescue team or public institutions}) then (entity 2 = {offenders, hostages, Professional})
        - if (entity 1 = null) then (entity 2 = {offenders, hostages})
- arrive (entity, place) a person or a vehicle arrives at a place at a specific moment
    - entity: person or vehicle
    - place: place
    - Constraint: (entity = {offender, vehicle, hostage, governmental executive or rescue team)
- ask_for (entity 1, entity 2, quantity, activity): sb asks for an activity or a number of things
    - entity 1: persons or public institution
    - entity 2: persons or public institution
    - quantity : cardinal or amount
    - activity : activity
    - Constraints:
        - if (entity 1 = {hostage or null}) then (entity 2 = {public institution,journalist, governmental executive or null})
        - if (entity 1 = {public institution or null}) than (entity 2 = {offender, public institution or null})
        - if (entity 1 = relatives or null) then (entity 2 = {governmental executive, rescue team or null})
        - if (entity 1 = demonstrators or null) then (entity 2 = {governmental executive or null})
- assure (entity 1, entity 2, of_what): sb assures sb else of an activity
    - entity 1: persons or public institutions
    - entity 2: persons or public institutions
    - what: activity
    - Constraints:
        - if (entity 1 = offender, Professional) then (entity 2 = {public institutions, Professional})
        - if (entity 1 = public institutions) then (entity 2 = {Professional})
- be_afraid (entity 1, entity 2): sb is afraid of sb else or of an activity
    - entity 1: hostages, public institution
    - entity 2: offender or activity
- block_the_way (entity 1, entity 2, means): sb blocks sb else's way using a vehicle
    - entity 1: persons
    - entity 2: persons or vehicle
    - means: vehicle
    - Constraints:
        - if (entity 1 = {rescue team}) then (entity 2 = {offender, vehicle})
        - if (entity 1 = {Professional}) then (entity 2 = {vehicle})
- deny (entity, activity): a person denies to do sth
    - entity: public institution, offender
    - activity : activity

- encircle (entity, what): sb encircles a place
  - entity: rescue team
  - what: Location of Conduct or vehicle
- end (entity, activity): sb ends an activity
  - entity: persons
  - activity: activity
  - Constraint:
    - (entity = {offender, rescue team or null})
- enter (entity, where): sb enters somewhere
  - entity: persons
  - where: place or vehicle
  - Constraint:
    - (entity = {professional, rescue team, governmental executives, offender or null})
    - (where = {Location of Conduct, vehicle or null})
- escape_from (entity 1, entity 2, quantity): a certain amount of people escape from a person or from a place or a vehicle escapes from a place
  - entity 1: persons
  - entity 2 : persons or place or vehicle
  - quantity: cardinal
  - Constraints:
    - If (entity 1 = {hostage or null}) then (entity 2 = {Location of Conduct, vehicle or null})
    - If (entity 1 = {offender or null}) then (entity 2 = {Location of Conduct or null})
- explode (armament, place, during_what): an armament or a vehicle explodes somewhere during an activity
  - armament: armament or vehicle
  - place: place
  - during_what: activity or temporal concept
- free (entity 1, entity 2, from_place, quantity): sb frees a number of people from a place
  - entity 1: persons
  - entity 2: persons
  - from_place: place
  - quantity: cardinal
  - Constraint:
    - (entity 1 = {offender, rescue team or null})
    - (entity 2 = {hostage or null})
- give_asylum (entity 1, entiy 2, quantity): sb gives asylum to a certain amount of people
  - entity 1: persons
  - entity 2: hostage
  - quantity: cardinal
  - Constraint:
    - (entity 1 = {public institution or null}))
- give_deadline (entity 1, entity 2, how_long, for_what): sb gives sb else a specific amount of time as deadline to do a specific activity
  - entity 1: persons
  - entity 2: persons
  - how_long: temporal concept
  - for_what: activity
  - Constraint:
    - if (entity 1 = {offender}) then (entity 2 = {public institutions})

- head_towards (entity, place): a person or a vehicle moves towards a place
  - entity: offender or hostage or rescue team or vehicle
  - place: place
- help (entity 1, entity 2, activity) sb helps sb else to do sth
  - entity 1: persons
  - entity 2: persons
  - activity: activity
    - Constraints:
      - if (entity 1 = {offender}) then (entity 2 = {offender})
      - if (entity 1 = {public institution or null}) then (entity 2 = {hostage, public institution})
- hijack (entity 1, entity 2): sb hijacks a vehicle at a specific moment
  - entity 1: offenders
  - entity 2: plane, bus
- hold (entity 1, entity 2, quantity, place): sb occupies a building or keeps a number of hostages somewhere or holds armament
  - entity 1: persons
  - entity 2: persons, place or armament
  - quantity: cardinal or quantitative or amount
  - place: place or vehicle
  - Constraints:
    - If (entity 1 = offender or null) then (entity 2 = {hostage, armament or null})
    - If (entity 1 = rescue team or null) then (entity 2 = {amount, armament or null})
    - If (entity 1 = governmental executive or null) then (entity 2 = {hostage, amount or null})
- hospitalised (entity, quantity): a number of people are hospitalized
  - entity: hostage, rescue team, relatives
  - quantity: cardinal or quantitative
- inform (entity 1, entity 2, about_activity): sb informs sb about sth
  - entity 1: persons or public institution
  - entity 2: persons or public institution
  - about_activity: activity
  - Constraints:
    - if (entity 1 = {Professional}) then (entity 2 = {null})
    - if (entity 1 = {hostage}) then (entity 2 = {professional})
    - if (entity 1 = {public institution}) then (entity 2 = {public institution, professional or null})
    - if (entity 1 = {offender}) then (entity 2= {hostage, professional, public institution})
    - if (entity 1= {rescue team}) then (entity 2 = {public institution, hostage, professional})
    - if (entity 1 = {null}) then (entity 2= {public institution})
- interrogate (entity 1, entity 2, about_what): sb interrogates sb else about sth
  - entity 1: rescue team, public institutions
  - entity 2: offender, hostages
  - about_what: activity
- kidnap (entity 1, entity 2, quantity): sb kidnaps sb else
  - entity 1: persons
  - entity 2: persons
  - quantity: cardinal
  - Constraint:
    - if (entity 1 = {offender or null}) then (entity 2 = {hostage})

- kill (entity 1, entity 2, means, quantity, place): sb kills a number of people somewhere using a weapon
  - entity 1: persons
  - entity 2: persons
  - means: armament
  - quantity: cardinal
  - place: place or vehicle
  - Constraints:
    - if(entity 1 = {offender or null}) then (entity 2 = {rescue team, hostage or null }) and (place = {Location of Conduct, vehicle or null})
    - if (entity 1 = {rescue team}) then (entity 2 = {offender, hostage or null }) and (place = {Location of Conduct, vehicle or null})
- lead (entity 1, entity 2, where):
  - sb leads sb else to a place
  - sb is the leader of an activity somewhere
    - entity 1: persons
    - entity 2: persons, activity or vehicle
    - where: place
    - Constraints:
    - if (entity 1 = {offender}) then (entity 2 = {vehicle, offender, activity})
    - if (entity 1 = {rescue team}) then (entity 2 = {rescue team, vehicle})
    - if (entity 1 = {hostage}) then (entity 2 = {rescue team})
- leave (entity, from_place, to_place, how, for_reason): a person or a vehicle leaves a place in a vehicle
  - entity 1: persons or vehicle
  - place: place or vehicle
  - means: vehicle
  - Constraint:
    - (entity 1 = {offender, hostage, rescue team, vehicle})
    - (place = {country, Location of Conduct})
- located (entity, place): sb or sth is located somewhere
  - entity: vehicle, offender, hostage, rescue team, governmental executive
  - place: place
- meet (entity 1, entity 2, where): sb meets sb else somewhere
  - entity 1: persons
  - entity 2: persons
  - where: place
  - Constraints:
    - if (entity 1= {professional}) then (entity 2 = {governmental executive, hostage})
    - if (entity 1 = {governmental executive}) then (entity 2= {governmental executive})
    - if (entity 1 = {hostage}) then (entity 2 = {professional, relatives})
- negotiate (entity 1, entity 2, about): people negotiate about sth
  - entity 1: persons
  - entity 2: persons
  - about: activity
  - Constraints:
    - if (entity 1 = {rescue team, governmental executive or null }) then (entity 2 = {offender or null})
    - if (entity 1 = {offender or null}) then (entity 2 = {governmental executive or null})

- organize (entity 1, entity 2) sb organizes sb or an activity
  - entity 1: persons or public institution
  - entity 2: persons or activity
  - Constraints:
    - if (entity 1 = {offender}) then (entity 2 = {offender or activity})
    - if (entity 1 = {rescue team}) then (entity 2 = {activity})
    - if (entity 1 = {public institution}) then (entity 2 = {activity})
- pay_ransom (entity 1, entity 2, amount): sb pays sb else a specific amount of money for ransom
  - entity 1: persons or public institution
  - entity 2: persons
  - amount: amount
  - Constraint:
    - if (entity 1 = {public institution or null}) then (entity 2 = {offender})
- physical_condition (entity, how): sb is in a specific physical condition
  - entity: hostage, offender
  - how: physical condition
- prevent_from (entity 1, entity 2, activity): sb prevents sb else or a vehicle from doing sth
  - entity 1: persons or public institution
  - entity 2: persons or public institution or vehicle
  - activity: activity
  - Constraints:
    - if (entity 1 = {rescue team}) then (entity 2 = {offender, professional})
    - if (entity 1 = {offenders}) then (entity 2 = {public institution, or rescue team})
    - if (entity 1 = {public institution}) then (entity 2 = {offenders, professional, vehicle})
- put (entity 1, entity 2, place)
  - sb places a person, an armament or vehicle somewhere
  - sb place armament on themselves or inside a vehicle
    - entity 1: persons
    - entity 2: persons, armament or vehicle
    - place: place, persons vehicle
    - Constraint:
      - (entity 1 = {offender}) and (entity 2 = {armament})
- return (entity, from_place, to_place, means): a vehicle or sb returns to a place from a place using a vehicle
  - entity: hostage or governmental executive, vehicle
  - from_place: country or city
  - to_place: country or city
  - means: vehicle
- speak_on_the_phone (entity 1, entity 2): sb speaks to sb else on the phone
  - entity 1: persons or public institutions
  - entity 2: persons or public institutions
  - Constraint:
    - if (entity 1 = {hostage}) then (entity 2 = {professional, relatives, public institutions})
    - if (entity 1 = {public institutions}) then (entity 2 = {public institutions, professional, relatives, offender})
    - if (entity 1 = {offender}) then (entity 2 ={ public institutions, professional, offender, rescue team})
    - if (entity 1 = {rescue team}) then (entity 2 = {offender})

- start (entity , activity): sb starts an activity
  - entity: persons
  - activity: activity
  - Constraint:
    - (entity = {offender, rescue team, public institution or null})
- stay_parked (what, place): a vehicle stays parked at a place
  - what: plane, bus
  - place: locality
- take_control_of (entity 1, entity 2, of_what) sb takes over the control of a place or a vehicle from entity2
  - entity 1: persons
  - entity 2: persons
  - of_what: place or vehicle
  - Constraint:
    - if (entity 1 = {offender}) then (entity 2 = {professional, null})
    - if (entity 1 = {rescue team}) then (entity 2 = {offender, null})
- take_on_responsibility (entity , activity): sb takes on responsibility for having done sth
  - entity: offender
  - activity: activity
- threaten (entity 1, entity 2, activity, in_case): sb threatens to do sth in the case of an activity
  - entity 1: persons or public institution
  - entity 2: persons or public institution
  - activity: activity
  - in_case: activity
  - Constraints:
    - if (entity 1 = {offender or null}) then (entity 2 = {hostage, public institution or null })
    - if (entity 1 = {public institution or null}) then (entity 2 = {offender or null})
- trade (entity 1, entity 2, what, with_what): sb exchanges sth or sb with sb else
  - entity 1: persons
  - entity 2: persons or public institution
  - what: activity or persons
  - with_what: activity or persons
  - Constraints:
    - (entity 1 = {offender}), (entity 2 = {rescue team, public institution})
    - if (what = {hostage}) then (with_what = {amount})
    - if (what = {offender}) then (with_what = {hostage})
- transport (entity 1, entity 2, to_place)
  - sb transports sb else to a place
  - a vehicle transports sb or some things to a place
    - entity 1: persons or vehicle
    - entity 2: persons
    - to_place: place or vehicle
    - Constraints:
      - if (entity 1 = {public institution or null}) then (entity 2 = {rescue team, hostages})
      - if (entity 1 = {vehicle}) then (entity 2= {rescue team, hostages, armament})
      - if (entity 1 = {rescue team}) then (entity 2 = {hostages})
- work_for (entity 1, entity 2) sb works for sb else
  - entity 1: persons

- entity 2: persons or public institution
- Constraint:
  - if (entity 1 = {hostage}) then (entity 2 = {public institution})
  - if (entity 1 = {offender}) then (entity 2 = {offender})

7.4.4 Τρόπος υλοποίησης του αρθρώματος των μηνυμάτων

Εν συντομία τα βήματα σε πολύ αφαιρετικό επίπεδο του αλγορίθμου συμπλήρωσης των ορισμάτων των μηνυμάτων που επινοήθηκε και ονομάστηκε SystemAegeanNcsrMessageArgumentFilling είναι τα εξής:

1. Πάρε το κείμενο από το κάθε κείμενο της συλλογής σε μορφή ακατέργαστων δεδομένων (raw data)
2. Μετέτρεψέ το σε συμβολοσειρά (String)
3. Επέλεξε όλες τις επισημειώσεις τύπου original_message και εφάρμοσε φιλτράρισμα παίρνοντας τελικά μόνο αυτές τις ιδιότητες που έχουν τιμή free
4. Για κάθε επισημείωση τύπου original_message με ιδιότητες που έχουν τιμή free
   4.1. Πάρε το κείμενο της επισημείωσης σε μορφή ακολουθίας από bytes (ByteSequence)
   4.2. Μετέτρεψε το σε συμβολοσειρά
   4.3. Πάρε τα όρια της έκτασης της επισημείωσης
   4.4. Επέλεξε όλες τις επισημειώσεις τύπου Persons και εφάρμοσε φιλτράρισμα παίρνοντας τελικά μόνο αυτές μόνο αυτές τις ιδιότητες που έχουν τιμή Offender ή Rescue_Team
      4.4.1.Για κάθε επισημείωση τύπου Persons με ιδιότητες που έχουν τιμή Offender ή Rescue_Team
         4.4.1.1. Πάρε τα όρια της έκτασης της επισημείωσης
         4.4.1.2. Αν αυτά είναι μεταξύ των ορίων της έκτασης της επισημείωσης τύπου original_message
            4.4.1.2.1. Αποθήκευσε τότε σε μορφή συμβολοσειράς το id της επισημείωσης
            4.4.1.2.2. Βγες από το βρόχο
         4.4.1.3. Αν όχι
            4.4.1.3.1. Θέσε την τιμή null ως συμβολοσειρά
   4.5. Επέλεξε όλες τις επισημειώσεις τύπου Persons και εφάρμοσε φιλτράρισμα παίρνοντας τελικά μόνο αυτές μόνο αυτές τις ιδιότητες που έχουν τιμή Hostage
      4.5.1.Για κάθε επισημείωση τύπου Persons με ιδιότητες που έχουν τιμή Hostage
         4.5.1.1. Πάρε τα όρια της έκτασης της επισημείωσης
         4.5.1.2. Αν αυτά είναι μεταξύ των ορίων της έκτασης της επισημείωσης τύπου original_message
            4.5.1.2.1. Αποθήκευσε τότε σε μορφή συμβολοσειράς το id της επισημείωσης
            4.5.1.2.2. Βγες από το βρόχο
         4.5.1.3. Αν όχι
            4.5.1.3.1. Θέσε την τιμή null ως συμβολοσειρά
   4.6. Επέλεξε όλες τις επισημειώσεις τύπου Place
      4.6.1.Για κάθε επισημείωση τύπου Place
         4.6.1.1. Πάρε τα όρια της έκτασης της επισημείωσης
         4.6.1.2. Αν αυτά είναι μεταξύ των ορίων της έκτασης της επισημείωσης τύπου original_message
            4.6.1.2.1. Αποθήκευσε τότε σε μορφή συμβολοσειράς το id της επισημείωσης
            4.6.1.2.2. Βγες από το βρόχο
         4.6.1.3. Αν όχι
            4.6.1.3.1. Θέσε την τιμή null ως συμβολοσειρά

4.7. Δημιούργησε τις μορφές (patterns) των κανονικών εκφράσεων για την ποσότητα (quantity) σε μορφή συμβολοσειράς
4.8. Μεταγλώττισε τις κανονικές εκφράσεις
4.9. Μετέτρεψε το κείμενο της επισημείωσης original_message σε ακολουθία χαρακτήρων (Charsequence)
4.10. Δημιούργησε Matcher που αντιστοιχίζει γενικά τις κανονικές εκφράσεις με την ακολουθία χαρακτήρων της επισημείωσης original_message
4.11. Για όσο τώρα πραγματοποιείται ταίριασμα της μορφής (pattern) των κανονικών εκφράσεων που δημιουργήσαμε με κάποια επόμενη ποσότητα (υποτίθεται ότι έχει προηγηθεί ένα ταίριασμα)
4.11.1. Εάν η ποσότητα που ταιριάστηκε ταιριάζει με κάποια μορφή (pattern) κανονικής έκφρασης
4.11.1.1. Αποθήκευσε την ποσότητα που ταιριάστηκε σε μορφή συμβολοσειράς
4.11.1.2. Βγες από το βρόχο
4.11.2. Εάν όχι
4.11.2.1. Θέσε την τιμή null σε μορφή συμβολοσειράς
4.12. Δημιούργησε επισημείωση τύπου SystemAegeanNcsrMessageArgumentFilling με ιδιότητες τύπου free, entity1, entity2, from_place και quantity

5. Κάνε την ίδια διαδικασία για όλους τους τύπους μηνυμάτων...