ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Παραγωγή κειμένων φυσικής γλώσσας από οντολογίες βιοϊατρικής με το σύστημα NaturalOWL»

Μαγδαληνή Ευαγγελακάκη

Α.Μ. 3090056

Επιβλέπων Καθηγητής: Ίων Ανδρουτσόπουλος

Βοηθός Επίβλεψης: Γεράσιμος Λάμπουρας

Αθήνα 2014

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1 - ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο της εργασίας

Σκοπός της εργασίας ήταν η παραγωγή κειμένων φυσικής γλώσσας από οντολογίες βιοϊατρικής με το σύστημα NaturalOWL 2.0 [And13], μια μηχανή παραγωγής κειμένων φυσικής γλώσσας για οντολογίες OWL του Σημασιολογικού Ιστού [Ber01], που αναπτύχθηκε από την Ομάδα Επεξεργασίας Φυσικής Γλώσσας του Οικονομικού Πανεπιστημίου Αθηνών. Το κίνητρο πίσω από τη συγκεκριμένη εργασία ήταν η πληθώρα των ιατρικών οντολογιών που χρησιμοποιούνται πλέον σε πολλά συστήματα, ενώ παράλληλα παρατηρείται μια δυσκολία κατανόησης των πληροφοριών που προσφέρουν από το μέσο χρήστη. Έτσι, ερευνήσαμε οντολογίες ιατρικού περιεχομένου, δημιουργήσαμε τους γλωσσικούς πόρους που απαιτούνται από το NaturalOWL 2.0 για να παράγει ευανάγνωστα κείμενα φυσικής γλώσσας που τις περιγράφουν και εκτελέσαμε πειράματα για να διαπιστώσουμε κατά πόσο αυτά τα κείμενα είναι ευανάγνωστα, ορθά και χρήσιμα στην βιοϊατρική κοινότητα.

1.2 Διάρθρωση της εργασίας

Στο κεφάλαιο 2 εισάγονται και εξηγούνται συνοπτικά η θεωρία και οι έννοιες πάνω στις οποίες είναι βασισμένη η παρούσα εργασία και παρουσιάζεται συνοπτικά ο τρόπος λειτουργίας και χρήσης του συστήματος NaturalOWL 2.0. Στο κεφάλαιο 3 παρουσιάζονται οι δύο οντολογίες που ερευνήσαμε, ο τρόπος παραγωγής κειμένων από αυτές, καθώς και τα αποτελέσματα των πειραμάτων μας. Το κεφάλαιο 4 συνοψίζει τα συμπεράσματα της εργασίας και προτείνει μελλοντικές κατευθύνσεις έρευνας.

1.3 Ευχαριστίες

Καταρχάς θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Ίωνα Ανδρουτσόπουλο, για την συνεχή καθοδήγηση και βοήθεια που μου πρόσφερε καθ’ όλη την διάρκεια αυτής της εργασίας.
Επιπλέον ευχαριστώ το Γεράσιμο Λάμπουρα για τις συμβουλές, την τεχνική βοήθεια που μου πρόσφερε, καθώς και για το χρόνο που αφιέρωσε για την ανάγνωση και διόρθωση

του κειμένου της εργασίας.
Ακόμη, ευχαριστώ τους κκ. Γιάννη Αλμυράντη και Δημήτρη Πολυχρονόπουλο του Ινστιτούτου Βιοεπιστημών και Εφαρμογών του ΕΚΕΦΕ «Δημόκριτος» για τη βοήθειά τους στην αξιολόγηση των παραγομένων κειμένων.

ΚΕΦΑΛΑΙΟ 2 - ΤΟ ΣΥΣΤΗΜΑ NATURALOWL

2.1 Σημασιολογικός Ιστός και οντολογίες

Ο Σημασιολογικός Ιστός (Semantic Web) [Ber01] [Sha06] αποτελεί μια επέκταση του Παγκόσμιου Ιστού με πληροφορίες πιο εύκολα «κατανοητές» από υπολογιστές και μηχανισμούς χειρισμού αυτών των πληροφοριών. Για παράδειγμα, η παρακάτω πρόταση γίνεται εύκολα κατανοητή από έναν άνθρωπο, ενώ ένας υπολογιστής πολύ πιο δύσκολα «κατανοεί» το νόημά της:

A symptom of influenza is a sore throat.

Μια οντολογία, με την έννοια που έχει τα τελευταία χρόνια αυτός ο όρος στην πληροφορική, περιγράφει ένα συγκεκριμένο γνωστικό πεδίο (π.χ. προϊόντα υπολογιστών, αρχαιολογικούς χώρους, ασθένειες, κτλ.) ορίζοντας τις εννοιολογικές τάξεις (π.χ. οικογένειες βακτηρίων, ασθένειες), τις οντότητες που ανήκουν σε αυτές (π.χ. συγκεκριμένα βακτήρια), τις σχέσεις μεταξύ τους (π.χ. ότι ένα βακτήριο μεταδίδει κάποια ασθένεια), καθώς και τα αξιώματα των τάξεων (π.χ. τα είδη των οντοτήτων που επιτρέπεται να συμμετέχουν σε κάθε σχέση).

Η γλώσσα οντολογιών OWL (Web Ontology Language), που βασίζεται στο πρότυπο RDF (Resource Description Framework) [Ant04], αποτελεί ένα ευρέως διαδεδομένο πρότυπο για τον ορισμό οντολογιών στον Σημασιολογικό Ιστό. Η OWL2[1] [Gra08] είναι η νεότερη έκδοση της OWL.

Η παρακάτω λογική παράσταση, διατυπωμένη στο συναρτησιακό συντακτικό (functional-style syntax) της OWL2, περιγράφει τη σημασία της πρότασης «A symptom of influenza symptom is a sore throat.» του παραδείγματος παραπάνω:

ClassAssertion(:disease :influenza)
ClassAssertion(:symptom :sore-throat)
ObjectPropertyAssertion(:hasSymptom :influenza :sore-throat)

[1] Δείτε τη σελίδα http://www.w3.org/TR/owl2-primer/ για περισσότερες πληροφορίες για την OWL2.

Αναλύοντας τα παραπάνω αξιώματα, βλέπουμε ότι ορίζονται οι οντότητες «:influenza» και «:sore-throat», όπου ανήκουν στις εννοιολογικές τάξεις «:disease» και «:symptom» αντίστοιχα, και ότι οι οντότητες αυτές συνδέονται μέσω της σχέσης «:hasSymptom». Η παράσταση αυτή γίνεται πιο εύκολα κατανοητή από έναν υπολογιστή, συγκρινόμενη με ένα αντίστοιχο κείμενο φυσικής γλώσσας, αλλά είναι πιο δυσνόητη για έναν μη ειδικευμένο χρήστη.

2.2 Παραγωγή φυσικής γλώσσας και το σύστημα NaturalOWL

Παραγωγή φυσικής γλώσσας - ΠΦΓ (Natural Language Generation - NLG) [Rei00] είναι η αυτόματη δημιουργία ενός κειμένου φυσικής γλώσσας από μια αυστηρά ορισμένη λογική παράσταση πληροφορίας. Τα συστήματα ΠΦΓ παρουσιάζουν μεγάλο ερευνητικό αλλά και εμπορικό ενδιαφέρον, για παράδειγμα σε εφαρμογές όπου εμφανίζεται η ανάγκη για την αυτόματη παραγωγή παράλληλων κειμένων σε πολλές γλώσσες, όπως η παραγωγή πολύγλωσσων εγχειριδίων χρήσεως, δελτίων καιρού και ιστοσελίδων, αλλά και σε περιπτώσεις όπου οι πληροφορίες οντολογιών χρειάζεται να παρουσιαστούν σε χρήστες μη εξοικειωμένους με τυπικές (formal) παραστάσεις γνώσεων.

Συγκεκριμένα, στην εργασία αυτή χρησιμοποιείται το σύστημα παραγωγής φυσικής γλώσσας NaturalOWL. Το NaturalOWL βασίζεται σε ιδέες των ευρωπαϊκών ερευνητικών έργων ILEX [Don01] και M-PIRO [Isa03] και αναπτύχθηκε αρχικά στη διάρκεια προηγούμενης διπλωματικής εργασίας [Gal06]. Στη συνέχεια, χρησιμοποιήθηκε στο ευρωπαϊκό ερευνητικό έργο INDIGO [Kons08], σε εφαρμογές προφορικών διαλόγων με ρομποτικούς ξεναγούς [Ober08] [Vog08] [Kons09] και μελετήθηκαν βελτιώσεις του στη διάρκεια άλλων εργασιών [Kar07] [Mar09]. Η πιο πρόσφατη έκδοση του συστήματος, υποστηρίζει πλήρως τη νέα έκδοση του προτύπου OWL2 και έχει επεκταθεί ώστε να παράγει πιο πολύπλοκα και ενδιαφέροντα κείμενα [And13].

Το NaturalOWL 2.0 παράγει αυτόματα κείμενα που περιγράφουν τάξεις και οντότητες οντολογιών OWL2 στα Αγγλικά και στα Ελληνικά. Για την παραγωγή φυσικής γλώσσας απαιτεί σχετικούς γλωσσικούς πόρους οι οποίοι προ-κατασκευάζονται από ανθρώπους καλούμενους «συγγραφείς» (authors) γλωσσικών πόρων. Το NaturalOWL 2.0 έχει την δυνατότητα να παραγάγει κείμενα και χωρίς τη χρήση προκατασκευασμένων γλωσσικών

πόρων, αλλά η ποιότητα των κειμένων είναι αισθητά χειρότερη. Για παράδειγμα, το παραγόμενο κείμενο για την οντότητα «:DOID_0050204» της οντολογίας Disease χωρίς τη χρήση γλωσσικών πόρων είναι το παρακάτω:

Epstein-Barr virus hepatitis is a kind of viral hepatitis and it located in liver. It results in inflammation. It has symptom abdominal pain, fatigue, fever, headache, jaundice and nausea. It has material basis in Human herpesvirus 4.

Αντίθετα, όταν αξιοποιούνται γλωσσικοί πόροι το παραγόμενο κείμενο είναι πιο ευανάγνωστο και κατανοητό, όπως φαίνεται στο ακόλουθο κείμενο:

Epstein-Barr virus hepatitis is a kind of viral hepatitis that affects the liver. It results in inflammation. Its symptoms are abdominal pain, fatigue, fever, headaches, jaundice and nausea. It is caused by the human herpesvirus 4.

2.2.1 Γλωσσικοί πόροι

Οι γλωσσικοί πόροι που αξιοποιεί το σύστημα NaturalOWL 2.0 αποτελούνται από το λεξικό (lexicon), τα σχέδια προτάσεων (sentence plans), τα ονόματα φυσικής γλώσσας (NL names), τις πληροφορίες σχεδιασμού κειμένου (ordering and structure information) και τα μοντέλα χρηστών (user models). Οι γλωσσικοί πόροι του NaturalOWL 2.0 αποθηκεύονται σε μια ξεχωριστή οντολογία OWL2. Ο συγγραφέας των γλωσσικών πόρων δεν μπορεί να επέμβει στην οντολογία των γλωσσικών πόρων δημιουργώντας νέες τάξεις ή ιδιότητες· μπορεί μόνο να δημιουργεί δικές του οντότητες και αξιώματα για αυτές. Η συγγραφή όλων των απαραίτητων γλωσσικών πόρων γίνεται μέσω του εργαλείου συγγραφής του NaturalOWL 2.0, το οποίο είναι διαθέσιμο ως επέκταση (plug-in) της πλατφόρμας συγγραφής οντολογιών Protégé 4.0[2] του πανεπιστημίου του Stanford.

Το λεξικό (lexicon) περιέχει ουσιαστικά, επίθετα και ρήματα σε όλες τις κλίσεις τους (π.χ. αριθμός, πτώση, γένος, φωνή, χρόνος), τα οποία μπορούν να χρησιμοποιηθούν σε σχέδια προτάσεων και ονόματα φυσικής γλώσσας, όπως εξηγούμε παρακάτω. Το λεξικό

[2] http://protege.stanford.edu/

δεν χρειάζεται να περιέχει λέξεις ή τύπους τους που δεν χρησιμοποιούνται σε σχέδια προτάσεων ή ονόματα φυσικής γλώσσας. Μέχρι στιγμής υποστηρίζει λέξεις της ελληνικής και αγγλικής γλώσσας.

Τα σχέδια προτάσεων (sentence plans) προσδιορίζουν τον τρόπο που αξιώματα της οντολογίας μπορούν να εκφραστούν ως μεμονωμένες προτάσεις φυσικής γλώσσας. Για κάθε ιδιότητα P της οντολογίας, ο συγγραφέας των γλωσσικών πόρων παρέχει ένα ή περισσότερα σχέδια προτάσεων που καθορίζουν την αφηρημένη δομή μιας πρότασης που μπορεί να εκφράσει αξιώματα στα οποία συμμετέχει η ιδιότητα P. Υπάρχει η δυνατότητα να παραχθεί αυτόματα ένα αντίστοιχο σχέδιο πρότασης, χωρίς τη χρήση γλωσσικών πόρων, αλλά η ποιότητα του παραγόμενου κειμένου είναι αισθητώς χειρότερη. Κάθε σχέδιο πρότασης σχηματίζεται από μια ακολουθία από πεδία και πληροφορίες συμπλήρωσης αυτών των πεδίων. Τα δυνατά πεδία είναι τα παρακάτω:

- Property Owner: μια έκφραση που αναφέρεται στην οντότητα ή τάξη που είναι το υποκείμενο του αξιώματος. Για παράδειγμα, για το υποκείμενο «*:influenza»* της δήλωσης *ObjectPropertyAssertion(:hasSymptom :influenza :sore-throat)* μπορεί να παραχθεί η αναφορική έκφραση «this disease» ή «this» ή το όνομα φυσικής γλώσσας «*Influenza»*. Θα περιγράψουμε τα ονόματα φυσικής γλώσσας παρακάτω.

- Property Filler: μια έκφραση που αναφέρεται στο αντικείμενο του αξιώματος. Αν πρόκειται για κάποια οντότητα ή τάξη, τότε παράγεται το όνομα φυσικής γλώσσας που αντιστοιχεί σε αυτή, ενώ αν πρόκειται για τιμή κάποιου τύπου δεδομένων (π.χ. Integer) τότε στο πεδίο μπαίνει η τιμή αυτή ως έχει. Αν το αντικείμενο αποτελεί σύζευξη ή διάζευξη οντοτήτων, τάξεων ή τιμών, τότε παράγεται μια σύζευξη ή διάζευξη των ονομάτων φυσικής γλώσσας τους ή των τιμών αντίστοιχα.

- Lexicon entry: ένα επίθετο, ουσιαστικό ή ρήμα για το οποίο υπάρχει καταχώρηση στο λεξικό. Ο συγγραφέας μπορεί είτε να απαιτήσει προκαθορισμένο γένος, πτώση, αριθμό, κτλ. είτε να ορίσει ότι το επίθετο, ουσιαστικό ή ρήμα πρέπει να συμφωνεί με κάποιο άλλο πεδίο του σχεδίου πρότασης (π.χ. το ρήμα να συμφωνεί σε πρόσωπο και αριθμό με το υποκείμενο).

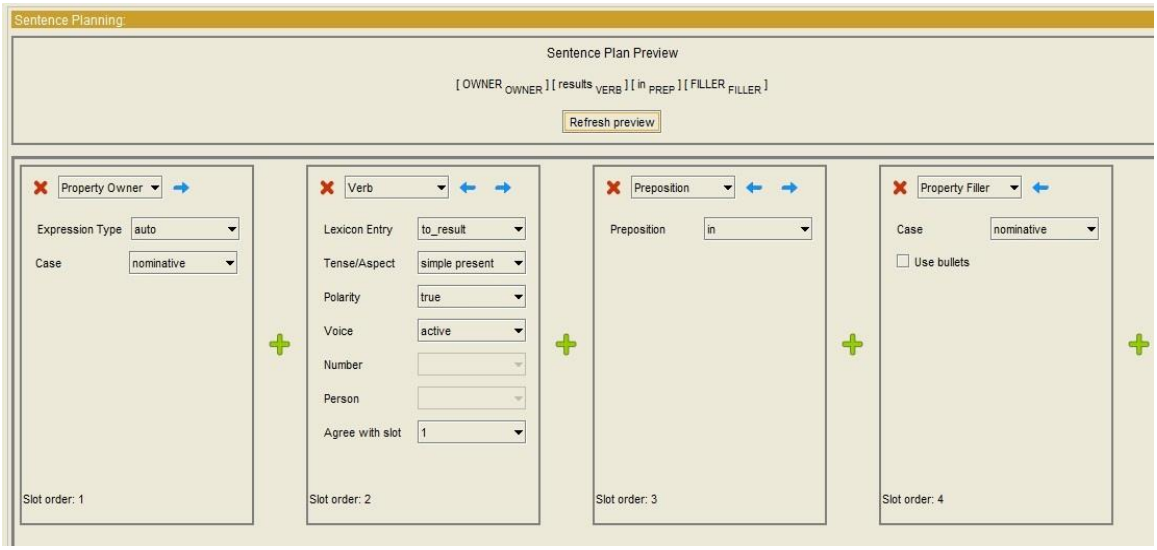
- Preposition: μια πρόθεση από μια προκαθορισμένη λίστα διαθέσιμων προθέσεων.

- String: μια συμβολοσειρά χαρακτήρων (π.χ. «it is assumed that …»).

- Concatenation: μία αλληλουχία τιμών ιδιοτήτων του αντικειμένου του αξιώματος. Για παράδειγμα, ας θεωρήσουμε τα παρακάτω αξιώματα:
  ObjectPropertyAssertion(:hasPrice :tecra8 _:n)
  DataPropertyAssertion(:hasAmount _:n "850"^^xsd:float)
  ObjectPropertyAssertion(:hasCurrency _:n :euroCurrency)

  Κοιτώντας το πρώτο αξίωμα, θέλουμε να παράγουμε μια πρόταση της οποίας το αντικείμενο να αποτελείται από την τιμή τύπου δεδομένων (850), ακολουθούμενη από το όνομα φυσικής γλώσσας του «euroCurrency» («Euros»), δηλαδή την τιμή της ιδιότητας «hasAmount», ακολουθούμενη από την τιμή της ιδιότητας «hasCurrency».

Στην παρακάτω εικόνα φαίνεται ένα σχέδιο πρότασης για την ιδιότητα «:resultsIn», όπως το βλέπει ο συγγραφέας των γλωσσικών πόρων όταν χρησιμοποιεί το plug-in του NaturalOWL 2.0 της πλατφόρμας συγγραφής οντολογιών Protégé.



Εικόνα 1*: Η καρτέλα «Sentence Plans» του plug-in του NaturalOWL, που χρησιμοποιείται από το συγγραφέα των γλωσσικών πόρων κατά τη δημιουργία σχεδίων προτάσεων.*

Αν, για παράδειγμα, το υποκείμενο (S) του αξιώματος είναι η τάξη «:Rift_Valley_fever» και το αντικείμενο (O) είναι η τάξη «:infection», στις θέσεις των πεδίων «Property Owner» και «Property Filler» θα παραχθούν αναφορικές εκφράσεις για τα S και O στην ονομαστική και αιτιατική πτώση αντίστοιχα. Το πεδίο «Verb» αναφέρεται στο ρήμα «to result» του λεξικού, και ορίζει ότι πρέπει να παραχθεί σε χρόνο απλό ενεστώτα και φωνή ενεργητική, χωρίς άρνηση (επιλογή polarity), ενώ στο πεδίο «Agree with slot» ορίζεται ότι πρέπει να συμφωνεί σε αριθμό, πτώση, φωνή και χρόνο με το πρώτο πεδίο του σχεδίου πρότασης («Property Owner» - το υποκείμενο). Δηλαδή, αν το S στο οποίο αναφέρεται το πεδίο «Property Owner» βρίσκεται στον πληθυντικό αριθμό, τότε και το ρήμα θα πρέπει να παραχθεί στον πληθυντικό αριθμό (π.χ. «They result …», αντί «He results …»). Στο πεδίο «Preposition» ορίζεται ότι πρέπει να χρησιμοποιηθεί η πρόθεση «in». Τελικά η πρόταση που θα παραχθεί είναι η ακόλουθη:

«Rift Valley fever results in infections»

Αντίστοιχα, τα ονόματα φυσικής γλώσσας (NL names) προσδιορίζουν την αφηρημένη δομή των φράσεων (συνήθως ονοματικές φράσεις, noun phrases) που μπορούν να χρησιμοποιηθούν για να ονοματίσουν τις οντότητες και τάξεις της οντολογίας. Όπως τα σχέδια προτάσεων, κάθε όνομα φυσικής γλώσσας σχηματίζεται από μια ακολουθία από πεδία και πληροφορίες συμπλήρωσης αυτών των πεδίων, συνδέεται δε με την οντότητα ή την τάξη την οποία ονοματίζει.
Τα δυνατά πεδία των ονομάτων φυσικής γλώσσας είναι τα παρακάτω:

- Article: ένα άρθρο, οριστικό ή αόριστο.

- Lexicon Entry: ένα επίθετο ή ουσιαστικό, για το οποίο υπάρχει καταχώρηση στο λεξικό. Αυτό το πεδίο μπορεί να σημειωθεί ως η πρωτεύουσα λέξη (head) του ονόματος φυσικής γλώσσας, οπότε το γένος, αριθμός και η πτώση του θα θεωρούνται και γένος, αριθμός, πτώση του συνολικού ονόματος.

- Preposition: μια πρόθεση από μια προκαθορισμένη λίστα διαθέσιμων προθέσεων.

- String: μια συμβολοσειρά χαρακτήρων.

Στην παρακάτω εικόνα φαίνεται το όνομα φυσικής γλώσσας της οντότητας «:Diagnostic_Procedure», όπως το βλέπει ο άνθρωπος-συγγραφέας των γλωσσικών πόρων όταν χρησιμοποιεί το plug-in του NaturalOWL 2.0.

Εικόνα 2: *Η καρτέλα «NL Names» του plug-in του NaturalOWL, που χρησιμοποιείται από τον συγγραφέα των γλωσσικών πόρων για τη δημιουργία ονομάτων φυσικής γλώσσας.*

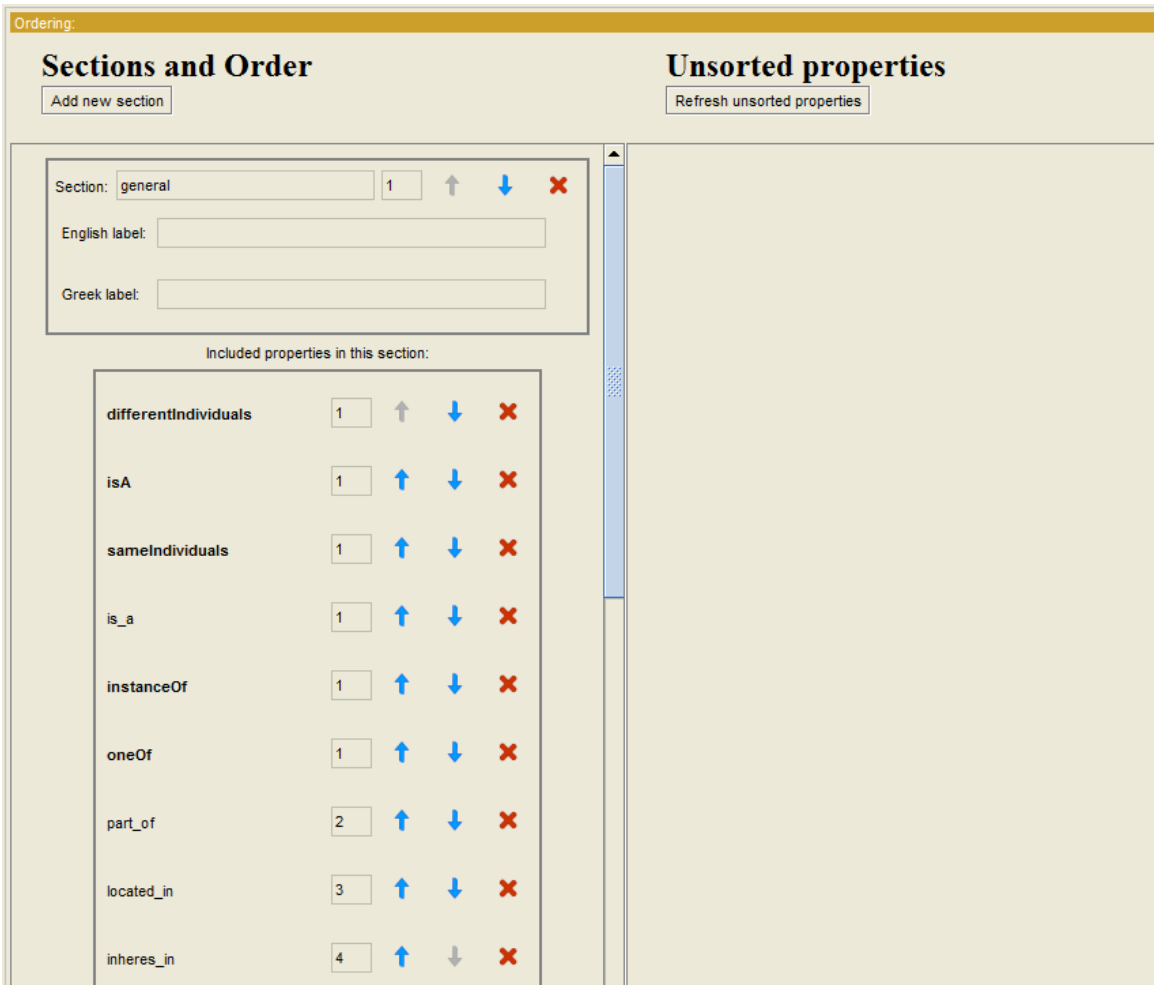
Στο πεδίο «Adjective» έχουμε το επίθετο «diagnostic» του λεξικού και στο πεδίο «Agree with slot» ορίζεται ότι πρέπει να συμφωνεί σε αριθμό, πτώση, φωνή και γένος με το πεδίο «Noun», το οποίο αναφέρεται στο ουσιαστικό «procedure» του λεξικού. Το ουσιαστικό είναι και η πρωτεύουσα λέξη της πρότασης (Head Noun) και στην συγκεκριμένη περίπτωση είναι στον ενικό αριθμό. Το πεδίο «Article» συμφωνεί κι αυτό με το πρωτεύον ουσιαστικό του ονόματος και απαιτείται να είναι αόριστο («a»). Τελικά το όνομα φυσικής γλώσσας που θα παραχθεί για το «:Diagnostic_Procedure» είναι το ακόλουθο:

«a diagnostic procedure»

Εκτός από τα σχέδια προτάσεων και τα ονόματα φυσικής γλώσσας, στους γλωσσικούς πόρους περιλαμβάνονται και πληροφορίες σχεδιασμού κειμένου (ordering and structure information). Όπως φαίνεται και στην παρακάτω εικόνα, στην αντίστοιχη καρτέλα

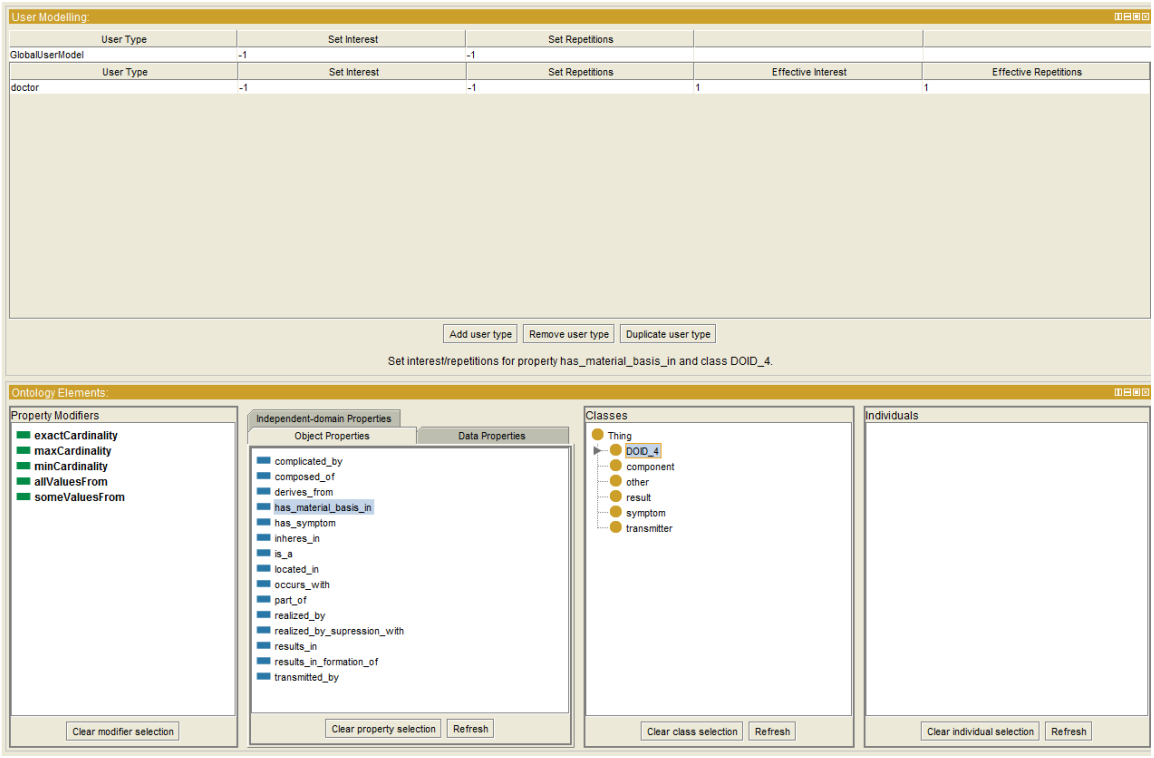
«Sections and Order» του plug-in του NaturalOWL 2.0 ο συγγραφέας των γλωσσικών πόρων ορίζει τις θεματικές ενότητες (που θα αντιστοιχούν σε παραγράφους στο τελικό κείμενο), με ποια σειρά («προτεραιότητα») θα πρέπει να εμφανίζονται οι ενότητες, καθώς και οι ιδιότητες (properties) που πρέπει να αναφέρονται σε κάθε ενότητα. Οι ενότητες ή ιδιότητες που έχουν την ίδια τιμή προτεραιότητας (π.χ. οι πρώτες έξι ιδιότητες που έχουν τιμή 1) μπορούν να εμφανιστούν με οποιαδήποτε σειρά μεταξύ τους.

Στο συγκεκριμένο παράδειγμα έχει οριστεί μία ενότητα «general» που περιλαμβάνει διάφορες ιδιότητες που εκφράζουν γενικές πληροφορίες σχετικά με μία ασθένεια, όπως οι «:located_in» και «composed_of». Η ενότητα «general» έχει και αυτή μία τιμή προτεραιότητας που ορίζει ότι όλες οι προτάσεις που αντιστοιχούν σε ιδιότητες που περιέχει θα εμφανιστούν πριν από προτάσεις άλλων ενοτήτων στο κείμενο.



Εικόνα 3: *Η καρτέλα «Sections and Order» του plug-in του NaturalOWL, που χρησιμοποιείται από το συγγραφέα των γλωσσικών πόρων κατά τη δημιουργία θεματικών ενοτήτων.*

Τέλος, στους γλωσσικούς πόρους περιλαμβάνονται και τα μοντέλα χρηστών (user models). Στην αντίστοιχη καρτέλα «User Modelling» του plug-in, που φαίνεται στην παρακάτω εικόνα, ο συγγραφέας έχει τη δυνατότητα να δημιουργήσει τύπους χρηστών όπως «ενήλικας» ή «παιδί», για τους οποίους το σύστημα πρέπει να παράγει διαφορετικά κείμενα (π.χ. σχέδια προτάσεων με πιο απλές λέξεις και λιγότερο πολύπλοκες προτάσεις όταν παράγονται κείμενα για παιδιά). Μπορεί να οριστεί ο μέγιστος αριθμός αξιωμάτων που επιτρέπεται να περιέχονται σε μια παραγόμενη περιγραφή (κείμενο), καθώς και ο μέγιστος αριθμός αξιωμάτων που επιτρέπεται να συνδυάζονται σε μια πρόταση[3]. Δίνουμε περισσότερες πληροφορίες για την επιλογή προτάσεων και το συνδυασμό τους στην επόμενη ενότητα.



Εικόνα 4: *Η καρτέλα «User Modelling» του plug-in του NaturalOWL, που χρησιμοποιείται από το συγγραφέα των γλωσσικών πόρων κατά τη δημιουργία τύπων χρηστών και άλλων πληροφοριών (ενδιαφέρον/επαναλήψεις).*

Ακόμα, στην ίδια καρτέλα ο συγγραφέας μπορεί να δηλώσει το ενδιαφέρον (interest) και το βαθμό επανάληψης (repetitions) για κάθε χρήστη, ανά αξίωμα της οντολογίας. Το ενδιαφέρον είναι μία τιμή που ορίζει πόσο ενδιαφέρουσα είναι για κάθε χρήστη η

[3] Για την ακρίβεια, οι περιορισμοί αυτοί αναφέρονται σε τριάδες μηνυμάτων, που παρουσιάζονται στην επόμενη ενότητα.

πληροφορία που εκφράζει το κάθε αξίωμα και μπορεί να πάρει τιμές από 0 (να μην αναφερθεί) έως και 3 (πολύ ενδιαφέρουσα). Ο βαθμός επανάληψης ορίζει πόσες φορές πρέπει να επαναληφθεί η πληροφορία του αξιώματος, πριν μπορέσουμε να θεωρήσουμε ότι ο χρήστης την έχει αφομοιώσει ή δεν τον ενδιαφέρει άλλο. Το ενδιαφέρον και ο βαθμός επανάληψης μπορούν να οριστούν για αξιώματα που αφορούν μια ιδιότητα P, ή πιο συγκεκριμένα για αξιώματα στα οποία συμμετέχει η ιδιότητα P και μια συγκεκριμένη τάξη ή οντότητα ως αντικείμενο της.

2.2.2 Στάδια παραγωγής φυσικής γλώσσας του NaturalOWL

Το σύστημα NaturalOWL 2.0 ακολουθεί την αρχιτεκτονική διασωλήνωσης (pipeline), όπου το αποτέλεσμα κάθε σταδίου αποτελεί είσοδο στο επόμενο στάδιο. Αναλυτικά τα στάδια περιγράφονται παρακάτω, με περισσότερη έμφαση στις διαδικασίες που αφορούν τους σκοπούς αυτής της εργασίας. Το σύστημα περιγράφεται πλήρως στο [And13].

Η παραγωγή φυσικής γλώσσας γίνεται σε τρία διαδοχικά στάδια, το σχεδιασμό του εγγράφου (document planning), το μικρο-σχεδιασμό (micro-planning) και την παραγωγή επιφανειακής δομής (surface realization).

Σχεδιασμός εγγράφου (document planning):

Το πρώτο στάδιο του σχεδιασμού του εγγράφου αποτελείται με τη σειρά του από την επιλογή περιεχομένου (content selection) και το σχεδιασμό του κειμένου (text planning). Κατά την επιλογή περιεχομένου, το σύστημα ανακτά από την οντολογία όσα αξιώματα έχουν ως υποκείμενο την συγκεκριμένη τάξη ή οντότητα που θέλουμε να περιγράψουμε. Τα αξιώματα μετατρέπονται σε τριάδες («μηνύματα») της μορφής <S, P, O>, όπου το S είναι το υποκείμενο του αξιώματος (στην περίπτωσή μας η συγκεκριμένη τάξη ή οντότητα που θέλουμε να περιγράψουμε), το P είναι η ιδιότητα της οντολογίας που σχηματίζει το αξίωμα και το O είναι μια οντότητα, τάξη ή τιμή (π.χ. ακέραιος αριθμός, συμβολοσειρά, κτλ.) που σχετίζεται με το S μέσω της ιδιότητας P.

Στο παρακάτω παράδειγμα, το σύστημα ανέκτησε 6 τριάδες για την οντότητα «:nephropathia_epidemica». Όπως φαίνεται, αξιώματα με πολλαπλά αντικείμενα συνδυάζονται σε μια τριάδα:

<:nephropathia_epidemica, :isA, :hemorrhagic_fever_with_renal_syndrome>
<:nephropathia_epidemica, :has_material_basis_in, :Puumala_virus>
<:nephropathia_epidemica, :has_symptom, and(:abdominal_pain,
:back_pain,
:internal_hemorrhage,
:renal_failure,
:headache,
:myalgia,
:nausea,
:vomiting)>
<:nephropathia_epidemica, :located_in, :kidney>
<:nephropathia_epidemica, :results_in, :infection>
<:nephropathia_epidemica, :transmitted_by, :bank_vole>

Από τις τριάδες αυτές επιλέγεται ένα υποσύνολο για να εκφραστεί στο κείμενο, ανάλογα με τις προκαθορισμένες προτιμήσεις του μοντέλου χρήστη στον οποίο απευθύνεται το κείμενο (π.χ. περιορισμούς μεγέθους των τελικών προτάσεων). Τελικά το σύστημα επιλέγει να εκφράσει τις πιο «ενδιαφέρουσες» τριάδες, δηλαδή αυτές με το μεγαλύτερο βαθμό ενδιαφέροντος. Λαμβάνει όμως υπόψη του το ιστορικό αλληλεπίδρασης του συστήματος με το χρήστη, αγνοώντας τις τριάδες που έχουν παρουσιαστεί ήδη στο χρήστη όσες φορές επιτρέπει ο βαθμός επανάληψής τους.

Κατά το σχεδιασμό του κειμένου, οι τριάδες που έχουν προκύψει από την επιλογή περιεχομένου χωρίζονται σε ενότητες και ταξινομούνται με βάση τους γλωσσικούς πόρους σχεδιασμού κειμένου ώστε να σχηματίζεται μια συνεκτική αφήγηση. Ακολουθώντας το προηγούμενο παράδειγμα, οι τριάδες χωρίζονται σε τρεις ενότητες. Η πρώτη ενότητα περιέχει γενικές πληροφορίες για την ασθένεια, στη συγκεκριμένη περίπτωση όσες τριάδες αναφέρονται στις ιδιότητες «:isA» και «:located_in», η δεύτερη περιέχει πληροφορίες για τα αποτελέσματα της ασθένειας, δηλαδή τις τριάδες που

αναφέρονται στις «:has_symptom» και «:results_in», και η τρίτη ενότητα περιέχει τις ιδιότητες που εκφράζουν από πού προέρχεται η ασθένεια και πώς μεταδίδεται, στην περίπτωσή μας τις «:transmitted_by» και «:has_material_basis_in» με αυτή τη σειρά.

General

<:nephropathia_epidemica, :isA, :hemorrhagic_fever_with_renal_syndrome>
<:nephropathia_epidemica, :located_in, :kidney>

Symptoms

<:nephropathia_epidemica, :has_symptom, and(:abdominal_pain,
:back_pain,
:internal_hemorrhage,
:renal_failure,
:headache,
:myalgia,
:nausea,
:vomiting)>
<:nephropathia_epidemica, :results_in, :infection>

Causes

<:nephropathia_epidemica, :transmitted_by, :bank_vole>
<:nephropathia_epidemica, :has_material_basis_in, :Puumala_virus>

Μικρο-σχεδιασμός (micro-planning):

Το επόμενο στάδιο είναι αυτό του μικρο-σχεδιασμού, το οποίο μετατρέπει τις τριάδες που έχουν επιλεγεί από τα προηγούμενα στάδια σε αφηρημένες παραστάσεις προτάσεων. Ο μικρο-σχεδιασμός αποτελείται από τρία υπο-στάδια, τη λεξικοποίηση (lexicalization), την ενοποίηση προτάσεων (sentence aggregation) και την παραγωγή αναφορικών εκφράσεων (referring expression generation).

Κατά την διαδικασία της λεξικοποίησης το σύστημα επιλέγει πώς θα εκφράσει καθεμιά από τις επιλεγμένες τριάδες σε φυσική γλώσσα επιλέγοντας για κάθε συμμετέχουσα ιδιότητα ένα από τα διαθέσιμα σχέδια προτάσεων. Συνεχίζοντας το προηγούμενο παράδειγμα, για την τριάδα <*:nephropathia_epidemica*, :results_in , :infection> θα επιλεγεί το σχέδιο πρότασης που αντιστοιχεί στην ιδιότητα :results_in, δηλαδή το «[*:nephropathia_epidemica*] [result] [in] [:infection]». Κάθε αγκύλη στον παραπάνω συμβολισμό αντιστοιχεί σε ένα πεδίο του σχεδίου πρότασης. Να σημειώσουμε πως για τα αξιώματα που η σημασιολογία τους είναι σταθερή από οντολογία σε οντολογία (π.χ. isA ή SubClassOf) υπάρχουν προκαθορισμένα σχέδια προτάσεων στο NaturalOWL 2.0.

Ακολουθεί η μορφή του παραδείγματός μας μετά το στάδιο της λεξικοποίησης, όπου έχουν επιλεγεί και εφαρμοστεί τα σχέδια προτάσεων σε κάθε τριάδα.

General

[:nephropathia_epidemica] [is] [a] [kind] [of] [:hemorrhagic_fever_with_renal_ syndrome]

[:nephropathia_epidemica] [affect] [kidney]

Symptoms

[:nephropathia_epidemica] [symptom] [is] [and(:abdominal_pain,
:back_pain,
:internal_hemorrhage,
:renal_failure,
:headache,
:myalgia,
:nausea,
:vomiting)]

[:nephropathia_epidemica] [result] [in] [:infection]

Causes

[:nephropathia_epidemica] [transmit][by][:bank_vole]

[:nephropathia_epidemica] [cause][by][:Puumala_virus]

Στο επόμενο βήμα, την ενοποίηση προτάσεων, επιλέγονται (αφηρημένες) προτάσεις που θα συνδυαστούν ώστε να σχηματίσουν μεγαλύτερες προτάσεις, λαμβάνοντας και πάλι υπ' όψιν τις προτιμήσεις του μοντέλου χρήστη. Ένα σύνολο χειροποίητων κανόνων [And13] εφαρμόζονται άπληστα (greedily) στα σχέδια προτάσεων, όπως αυτά έχουν προκύψει από το στάδιο του καθορισμού δομής του κειμένου. Σε όποιες γειτονικές προτάσεις εφαρμόζεται ένας κανόνας, αυτές συγχωνεύονται σε μία μεγαλύτερη πρόταση, με σκοπό την αποφυγή πλεονασμών και τη βελτίωση της αναγνωσιμότητας του κειμένου. Η περιγραφή των κανόνων συνένωσης ξεφεύγει από τους στόχους αυτής της εργασίας και παραλείπεται.

Ένας περιορισμός είναι ότι δεν μπορούν να συνδυαστούν τριάδες των οποίων οι αντίστοιχες ιδιότητες P ανήκουν σε διαφορετικές ενότητες του σχεδιασμού κειμένου. Επιπλέον, κάποια σχέδια προτάσεων ή ονόματα φυσικής γλώσσας μπορεί να απαγορεύουν τον συνδυασμό των τριάδων στα οποία συμμετέχουν με άλλες, για να αποφεύγεται η παραγωγή υπερβολικά μεγάλων ή πολύπλοκων προτάσεων. Ακολουθεί το παράδειγμά μας μετά την ενοποίηση:

General
[:nephropathia_epidemica] [is] [a] [kind] [of] [:hemorrhagic_fever_with_renal_ syndrome] [that] [affect] [kidney]

Symptoms
[:nephropathia_epidemica] [symptom] [is] [:abdominal_pain] [and] [:back_pain] [and] [:internal_hemorrhage] [and] [:renal_failure] [and] [:headache] [and] [:myalgia] [and] [:nausea] [and] [:vomiting]
[:nephropathia_epidemica] [result] [in] [:infection]

Causes
[:nephropathia_epidemica] [transmit] [by] [:bank_vole] [and] [cause] [by] [:Puumala_virus]

Στο τελευταίο υπο-στάδιο του μικρο-σχεδιασμού, γίνεται η παραγωγή αναφορικών εκφράσεων για το υποκείμενο S και αντικείμενο O κάθε τριάδας <S, P, O>. Για

παράδειγμα, ανάλογα με τις τριάδες και τα συμφραζόμενα, μπορεί να είναι καλύτερο να χρησιμοποιήσουμε ένα από τα ονόματα φυσικής γλώσσας που συνδέεται με τα S και O (π.χ. «Influenza»), μια αντωνυμία (π.χ. «It») ή μια δεικτική ονοματική φράση (π.χ. «This disease»). Το NaturalOWL 2.0 χρησιμοποιεί έναν απλό μηχανισμό, ο οποίος επιλέγει κάποια άλλη αναφορική έκφραση αντί του ονόματος φυσικής γλώσσας, μόνο όταν το αντικείμενο της προηγούμενης πρότασης αναφέρεται στην ίδια τάξη ή οντότητα με το υποκείμενο της τρέχουσας πρότασης. Μετά την επιλογή αναφορικών εκφράσεων το παράδειγμα γίνεται:

General

[Nephropathia epidemica] [is] [a] [kind] [of] [hemorrhagic fever with renal syndrome] [that] [affect] [the kidneys]

Symptoms

[It] [symptom] [is] [abdominal pain] [and] [back pain] [and] [internal hemorrhage] [and] [renal failure] [and] [headache] [and] [myalgia] [and] [nausea] [and] [vomiting]
[It] [result] [in] [infection]

Causes

[It] [transmit] [by] [bank vole] [and] [cause] [by] [the Puumala virus]

Παραγωγή επιφανειακής δομής (surface realization):

Στο τελευταίο στάδιο, αυτό της παραγωγής επιφανειακής δομής, οι προτάσεις είναι ήδη ενοποιημένες (σε μεγαλύτερες προτάσεις) και ταξινομημένες (ως προς την επιθυμητή τους σειρά) από τα προηγούμενα στάδια, οπότε απλά οι λέξεις παράγονται στην κατάλληλη μορφή τους (χρόνο, πτώση, αριθμό, κτλ.) και προστίθεται στίξη όπου χρειάζεται, με αποτέλεσμα το τελικό κείμενο. Το τελικό παραγόμενο κείμενο για την οντότητα «:nephropathia_epidemica» είναι:

Nephropathia epidemica is a kind of hemorrhagic fever with renal syndrome that affects the kidneys. Its symptoms are myalgia, nausea, renal failure, vomiting, abdominal pain, headaches, internal hemorrhage and back pain. It results in infections. It is caused by the Puumala virus and it is transmitted by bank voles.

ΚΕΦΑΛΑΙΟ 3 – ΓΛΩΣΣΙΚΟΙ ΠΟΡΟΙ ΒΙΟΪΑΤΡΙΚΩΝ ΟΝΤΟΛΟΓΙΩΝ

Στα πλαίσια αυτής της εργασίας ασχοληθήκαμε με διάφορες βιοϊατρικές οντολογίες, δηλαδή οντολογίες που περιγράφουν ασθένειες, βακτήρια, γονίδια κτλ. Συγκεκριμένα εξετάσαμε την οντολογία Medical Subject Headings (MeSH – ορίζει θεματικές κεφαλίδες για την επισημείωση βιοϊατρικών άρθρων), την Gene Ontology (περιγράφει ιεραρχίες γονιδίων και χαρακτηριστικά τους), την UniProt (περιγράφει πρωτεΐνες και χαρακτηριστικά τους), το Joint Chemical Dictionary (Jochem – αποτελεί ένα λεξικό χημικών στοιχείων και φαρμάκων), την οντολογία του Unified Medical Language System (UMLS – συλλογή από λεξικά για βιοϊατρικούς όρους) και την Disease Ontology (περιγράφει ασθένειες και χαρακτηριστικά τους). Οι οντολογίες που εξετάσαμε (εκτός από την οντολογία του UMLS) αποτελούν ουσιαστικά ταξινομίες τάξεων, αφού περιέχουν μόνο τάξεις και αξιώματα που περιγράφουν ιεραρχικές σχέσεις («is-a») και συνώνυμα όρων («synonym»), όπως τα παρακάτω αξιώματα που προέρχονται από την Gene Ontology.

<*mitochondrion inheritance, isA, organelle inheritance*>
<*mitochondrion inheritance, synonym, mitochondrial inheritance*>

Εκτός από τις σχέσεις αυτές, οι οντολογίες αυτές περιέχουν και κειμενικές περιγραφές που έχουν γραφτεί από ανθρώπους-συγγραφείς. Κάθε κειμενική περιγραφή συνδέεται με την αντίστοιχη τάξη που περιγράφει. Για παράδειγμα στην Gene Ontology, για την τάξη «mitochondrion inheritance» περιλαμβάνεται και η παρακάτω περιγραφή:

«The distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton.»

Όπως αναφέραμε και στο προηγούμενο κεφάλαιο, αυτά τα κείμενα δεν είναι πολύ χρήσιμα σε έναν υπολογιστή, ο οποίος δεν μπορεί εύκολα να εντοπίσει, να εξαγάγει ή να βγάλει συμπεράσματα από τις πληροφορίες που εμφανίζονται στο κείμενο. Επικεντρώσαμε την έρευνα μας στις οντολογίες UMLS και Disease Ontology, οι οποίες

περιέχουν περισσότερες πληροφορίες που μπορούν να χρησιμοποιηθούν κατά την αυτόματη παραγωγή κειμένων.

3.1 Η οντολογία UMLS

Το Unified Medical Language System[4] (UMLS) είναι μια προσπάθεια ανάπτυξης λεξικών και συστημάτων που μπορούν να βοηθήσουν τον εντοπισμό και την κατανόηση βιοϊατρικών όρων από υπολογιστές. Αποτελείται κυρίως από ένα συνδυασμό λεξικών με περισσότερους από ένα εκατομμύριο βιοϊατρικούς όρους, ένα σημασιολογικό δίκτυο ετικετών (παρόμοιο με μια οντολογία) και εργαλεία που παρέχουν λεξικολογικές πληροφορίες βιοϊατρικών όρων για επεξεργασία φυσικής γλώσσας. Το δίκτυο ετικετών είναι διαθέσιμο και ως οντολογία OWL[5] και περιέχει 135 τάξεις και 99 ιδιότητες (properties), που περιγράφουν γενικές τάξεις σχετικές με την υγεία (καθώς και τις σχέσεις ανάμεσα τους), όπως για παράδειγμα τις: «Anatomical_Abnormality», «Physical_Object», «Biologic_Function» ή «Vertebrate».

Για την παραγωγή κειμένων φυσικής γλώσσας από την UMLS μέσω του NaturalOWL 2.0 δημιουργήσαμε γλωσσικούς πόρους που αποτελούνταν από 55 επίθετα, 126 ουσιαστικά και 36 ρήματα στο λεξικό, 94 σχέδια προτάσεων (κάποιες ιδιότητες χρησιμοποιούν το ίδιο σχέδιο πρότασης) και 135 ονόματα φυσικής γλώσσας. Ακόμα, δημιουργήσαμε τρεις διαφορετικές θεματικές ενότητες για την οργάνωση του κειμένου καθώς και την ταξινόμηση των γεγονότων: την ενότητα «General_Info», που περιλαμβάνει ιδιότητες που περιγράφουν βασικές πληροφορίες για μια τάξη ή οντότητα, την «More_Info» που περιλαμβάνει ιδιότητες που περιγράφουν πιο ειδικές πληροφορίες για μια τάξη ή οντότητα και τέλος την ενότητα «Results» η οποία περιλαμβάνει ιδιότητες που περιγράφουν τα αποτελέσματα της εκάστοτε τάξης ή οντότητας. Τέλος, δημιουργήσαμε έναν μοναδικό τύπο χρήστη (doctor) και δηλώσαμε μηδενικές τιμές ενδιαφέροντος (Interest) για κάποιες τριάδες γεγονότων, ώστε να μην εμφανίζονται προφανείς πληροφορίες στο τελικό κείμενο.

[4] http://www.nlm.nih.gov/research/umls/
[5] http://krono.act.uji.es/people/Ernesto/UMLS_SN_OWL

3.2 Η Disease Ontology

Η Disease Ontology αναπτύχθηκε από βιοϊατρικούς ειδικούς με σκοπό την πλήρη περιγραφή διαφόρων ανθρωπίνων ασθενειών και των χαρακτηριστικών τους. Η οντολογία ορίζει 6.286 ασθένειες και 15 σχέσεις. Διατίθεται ελεύθερα[6] στην μορφή σύνταξης οντολογιών OBO (Open Biomedical Ontologies)[7] αλλά η μετατροπή της στη γλώσσα OWL2 είναι απλή. Η εκφραστικότητα της OBO είναι σχετικά παρόμοια με αυτή της OWL2, με μεγαλύτερο βάρος στις ανάγκες της βιολογικής κοινότητας.

Όπως αναφέραμε και στην προηγούμενη ενότητα, η Disease Ontology αποτελεί μια απλή ταξονομία. Περιέχει αποκλειστικά ιεραρχικές σχέσεις («is-a») και σχέσεις συνωνύμων όρων («synonym»). Οι σχέσεις συνωνύμων χρησιμοποιούνται επιπλέον για να συνδέσουν κάθε τάξη με τις «συνώνυμες» τάξεις της όπως αυτές εμφανίζονται σε άλλες οντολογίες, όπως η UMLS και η Gene Ontology. Οι 15 σχέσεις που ορίζονται στην οντολογία εμφανίζονται μόνο στις κειμενικές περιγραφές που συνοδεύουν κάθε τάξη. Για παράδειγμα, ακολουθούν οι πληροφορίες που είναι διαθέσιμες για την τάξη «:Rift_Valley_fever» στην οντολογία OBO της Disease:

Name: Rift Valley Fever (DOID_1328)
is-a: viral infectious disease (DOID_934)
Definition: A viral infectious disease that **results_in** infection, **has_material_basis_in** Rift Valley fever virus, which is **transmitted_by** Aedes mosquitoes. The virus affects domestic animals (cattle, buffalo, sheep, goats, and camels) and humans. The infection **has_symptom** jaundice, **has_symptom** vomiting blood, **has_symptom** passing blood in the feces, **has_symptom** ecchymoses (caused by bleeding in the skin), **has_symptom** bleeding from the nose or gums, **has_symptom** menorrhagia and **has_symptom** bleeding from venepuncture sites.

Σε αντίθεση με τις υπόλοιπες οντολογίες που εξετάσαμε, οι κειμενικές περιγραφές στην Disease Ontology έχουν συγγραφεί ακολουθώντας συγκεκριμένους κανόνες[8], η σύνταξη

[6] http://disease-ontology.org/
[7] Δείτε http://www.obofoundry.org/, http://www.geneontology.org/GO.format.obo-1_2.shtml και http://oboformat.googlecode.com/svn/branches/2011-11-29/doc/obo-syntax.html
[8] http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Style_Guide

των προτάσεών τους είναι σχετικά τυποποιημένη και χρησιμοποιούν κατά κανόνα τα ονόματα των σχέσεων της οντολογίας αντί για ρήματα (π.χ. «results_in»). Σημειώνουμε ότι ενώ όλες οι σχέσεις ορίζονται στην οντολογία, δεν χρησιμοποιούνται παρά μόνο μέσα στις κειμενικές περιγραφές. Επίσης, με εξαίρεση τις ασθένειες, οι οντότητες που αναφέρονται (π.χ. «Aedes mosquitoes») δεν ορίζονται στην οντολογία.

Εκμεταλλευόμενοι την τυποποιημένη σύνταξη των κειμενικών περιγραφών, μπορέσαμε να εφαρμόσουμε σε αυτές κάποια πρότυπα που αναζητούσαν τις εμφανίσεις των ονομάτων των σχέσεων (π.χ. «results_in»). Για κάθε εμφάνιση σχέσης στην περιγραφή εξαγάγαμε ως αντικείμενο της όποια οντότητα ακολουθούσε, εξετάζοντας σημεία στίξης και λέξεις που ορίζουν αλλαγή πρότασης (π.χ. «that», «which»). Κάθε οντότητα που αναφέρεται στην περιγραφή αλλά δεν περιέχεται στην οντολογία την ορίσαμε ως αντικείμενο των τάξεων «disease», «result», «symptom», «component», «transmitter» ή «other», αναλόγως το όνομα σχέσης με το οποίο εμφανίζεται. Οι τάξεις αυτές κατασκευάστηκαν από εμάς για την οργάνωση των νέων οντοτήτων. Η οντολογία μετά ελέγχθηκε εντατικά και οι ασυνέπειες ή λάθη του αλγορίθμου διορθώθηκαν χειρωνακτικά. Η νέα μορφή της οντολογίας περιέχει 6.746 τάξεις, 15 σχέσεις και 1.545 οντότητες. Για παράδειγμα, ο ορισμός της τάξης «Rift_Valley_fever» στη νέα μορφή της οντολογίας είναι ο ακόλουθος:

```
SubClassOf(:DOID_1328
    ObjectIntersectionOf(:DOID_934
        ObjectHasValue(:results_in :infection)
        ObjectHasValue(:has_material_basis_in :Rift_Valley_fever_virus)
        ObjectHasValue(:transmitted_by :Aedes_mosquitoes)
        ObjectHasValue(:has_symptom :jaundice)
        ObjectHasValue(:has_symptom :vomiting_blood)
        ObjectHasValue(:has_symptom :passing_blood_in_the_feces)
        ObjectHasValue(:has_symptom
        :ecchymoses_(caused_by_bleeding_in_the_skin))
        ObjectHasValue(:has_symptom :bleeding_from_the_nose_or_gums)
        ObjectHasValue(:has_symptom :menorrhagia)
        ObjectHasValue(:has_symptom :bleeding_from_venepuncture_sites)))
```

Αφού ολοκληρώθηκε η παραπάνω διαδικασία, από το σύνολο των 6.746 τάξεων της οντολογίας που ορίζουν ασθένειες, αγνοήσαμε τις 5.014 τάξεις για τις οποίες ακόμα ορίζονταν μόνο σχέσεις «is-a», καθώς τα κείμενα που παράγονται από αυτές δεν έχουν ιδιαίτερο ενδιαφέρον. Από τις υπόλοιπες 1732 τάξεις, επιλέξαμε τυχαία 200 για να τις εξετάσουμε κατά την ανάπτυξη των γλωσσικών πόρων (development set) και 100 τάξεις που χρησιμοποιήσαμε στη συνέχεια για την αξιολόγηση της ποιότητας των παραγόμενων κειμένων (test set). Για τις 200 τάξεις του πρώτου είδους δημιουργήθηκαν γλωσσικοί πόροι (εγγραφές στο λεξικό, σχέδια προτάσεων και ονόματα φυσικής γλώσσας). Συγκεκριμένα, δημιουργήθηκαν 115 επίθετα, 263 ουσιαστικά και 7 ρήματα στο λεξικό, 8 σχέδια προτάσεων και 397 ονόματα φυσικής γλώσσας. Για 7 από τις σχέσεις δεν δημιουργήθηκαν σχέδια προτάσεων, αφού δεν συμμετείχαν σε κανένα αξίωμα της οντολογίας που να είναι σχετικό με τις 200 development τάξεις (και δεν εμφανίζονταν στις περιγραφές των τάξεων). Επίσης, για κάποιες τάξεις (π.χ. «:viral hepatitis») δεν ήταν απαραίτητη η δημιουργία αντίστοιχου ονόματος φυσικής γλώσσας καθώς τα ονόματα φυσικής γλώσσας που δημιουργούσε το NaturalOWL 2.0 βασισμένο στις ετικέτες των τάξεων ήταν ικανοποιητικά. Τονίζουμε ότι δεν δημιουργήθηκαν επιπλέον γλωσσικοί πόροι για τις 200 test τάξεις και τα κείμενα που παράγονται από αυτές δεν εξετάστηκαν καθόλου κατά την δημιουργία των γλωσσικών πόρων.

Ακόμα, δημιουργήθηκαν θεματικές ενότητες και πληροφορίες ταξινόμησης (σειράς) και πάλι εξετάζοντας τις 200 development τάξεις. Ο διαχωρισμός των ιδιοτήτων και οι τιμές ταξινόμησης επιλεχθήκαν ώστε να συμφωνούν με τις οδηγίες ταξινόμησης που έχουν δοθεί στους συγγραφείς των περιγραφών της Disease Ontology. Πιο συγκεκριμένα, δημιουργήθηκαν 4 ενότητες: η ενότητα «general», που περιέχει τις ιδιότητες που εκφράζουν γενικές πληροφορίες για την εκάστοτε τάξη (π.χ. «located_in», «part_of»), η ενότητα «causes», που περιγράφει τι μπορεί να προκαλέσει την συγκεκριμένη ασθένεια και περιέχει τις ιδιότητες «has_material_basis_in» και «derives_from», η ενότητα «symptoms», που περιγράφει τα συμπτώματα της ασθένειας και αποτελείται από τις ιδιότητες «has_symptom», «results_in» και «results_in_formation_of», και η ενότητα «complications», που περιλαμβάνει ιδιότητες όπως η «complicated_by».

Τέλος, όπως και στην οντολογία UMLS, ορίσαμε έναν τύπο χρήστη «doctor» και θέσαμε μηδενικές τιμές ενδιαφέροντος (Interest) σε κάποιες τριάδες γεγονότων, ώστε μην εμφανίζονται προφανείς πληροφορίες στο τελικό κείμενο· για παράδειγμα, ώστε να εξαιρείται η παρακάτω τριάδα και η αντίστοιχη πρόταση:

<:Skin_Angiosarcoma, :located_in, :skin>

«*Skin angiosarcoma is located in the skin.»*

3.3 Αποτελέσματα πειραμάτων

3.3.1 Αποτελέσματα οντολογίας UMLS

Οι τάξεις της οντολογίας UMLS αναφέρονται σε πολύ γενικές έννοιες και σχέσεις του βιοϊατρικού πεδίου. Ως αποτέλεσμα, τα παραγόμενα κείμενα που περιγράφουν αυτές τις τάξεις δεν παρουσιάζουν πολύ ενδιαφέρον. Προτιμήσαμε να μην προβούμε σε πειραματική ανάλυση αυτών των κειμένων με ανθρώπους κριτές και επικεντρωθήκαμε σε πειράματα με την Disease Ontology. Ακολουθούν επιλεγμένα παραδείγματα κειμένων από την UMLS με και χωρίς χρήση γλωσσικών πόρων κατά την αυτόματη παραγωγή κειμένου:

- Για την τάξη «Activity» έχουμε:

  <u>Παραγόμενο κείμενο χωρίς χρήση γλωσσικών πόρων:</u>
  Activity is a kind of Event. It has only an evaluation of Qualitative Concept. Activity is performed by only Group.

  <u>Παραγόμενο κείμενο με χρήση γλωσσικών πόρων:</u>
  An activity is a kind of event and it is evaluated by only qualitative concepts. It is performed by only groups.

- Για την τάξη «Therapeutic or Preventive Procedure» έχουμε:

Παραγόμενο κείμενο χωρίς χρήση γλωσσικών πόρων:

Therapeutic or Preventive Procedure is a kind of Health Care Activity. It prevents only Pathologic Function. It affects only Patient or Disabled Group. It is only method of Therapeutic or Preventive Procedure. It complicates Therapeutic or Preventive Procedure. It follows only Diagnostic Procedure.

Παραγόμενο κείμενο με χρήση γλωσσικών πόρων:

A therapeutic or a preventive procedure is a kind of healthcare activity and it follows only diagnostic procedures. It prevents only pathologic functions. It affects only patients or disabled groups.

- Για την τάξη «Patient or Disabled Group» έχουμε:

Παραγόμενο κείμενο χωρίς χρήση γλωσσικών πόρων:

Patient or Disabled Group is a kind of Group. Patient or Disabled Group is treated by only Professional or Occupational Group.

Παραγόμενο κείμενο με χρήση γλωσσικών πόρων:

Patients or disabled groups are a kind of group. They are treated by only professional or occupational groups.

3.3.2 Αποτελέσματα Disease Ontology

Για την διεξαγωγή των πειραμάτων με την Disease Ontology συλλέξαμε τρία σύνολα κείμενων για τις 200 development τάξεις και για τις 200 test τάξεις. Η πρώτη πηγή (Original) αποτελείται από τις κειμενικές περιγραφές των τάξεων όπως αυτές κατασκευάστηκαν από τους συγγραφείς της οντολογίας. Το δεύτερο σύνολο κείμενων (without resources) παράχθηκε από την NaturalOWL 2.0 χωρίς τη χρήση των γλωσσικών πόρων που κατασκευάσαμε κατά την διάρκεια αυτής της εργασίας και το τρίτο σύνολο (with resources) παράχθηκε από την NaturalOWL 2.0 χρησιμοποιώντας όλους τους διαθέσιμους γλωσσικούς πόρους. Ακολουθούν παραδείγματα κειμένων από τα τρία σύνολα για την τάξη «:Gastrointestinal_anthrax» που ανήκει στο development set:

Αρχικό κείμενο (Original):

An anthrax disease that results_in infection located_in mucosa of gastrointestinal tract, has_material_basis_in Bacillus anthracis, which is transmitted_by ingestion of anthrax-infected meat. The infection has_symptom lesions, has_symptom vomiting of blood, has_symptom severe diarrhea, has_symptom loss of appetite.

Παραγόμενο κείμενο χωρίς χρήση γλωσσικών πόρων (without Resources):

Gastrointestinal anthrax is a kind of anthrax disease. It has symptom loss of appetite, vomiting of blood, severe diarrhea and lesions. It results in infection. It located in mucosa of gastrointestinal tract. It has material basis in Bacillus anthracis. It transmitted by ingestion of anthrax-infected meat.

Παραγόμενο κείμενο με χρήση γλωσσικών πόρων (with Resources):

Gastrointestinal anthrax is a kind of anthrax disease that results in infections and affects the mucosa of the gastrointestinal tract. Its symptoms are severe diarrhea, lesions, loss of appetite and vomiting blood. It is transmitted by the ingestion of anthrax-infected meat and it is caused by Bacillus anthracis.

Αρχικά συλλέξαμε κείμενα από τα προαναφερθέντα σύνολα για 10 τυχαία επιλεγμένες τάξεις από τις development και 10 τάξεις από τις test. Τα 60 κείμενα (10 x 3 + 10 x 3) που προέκυψαν δόθηκαν σε δύο ανθρώπους κριτές – ειδικούς βιοϊατρικής. Οι κριτές εξέταζαν τα κείμενα σε τριάδες, όπου κάθε κείμενο της τριάδας περιέγραφε την ίδια τάξη αλλά προερχόταν από διαφορετικό σύνολο. Τα κείμενα της κάθε τριάδας ήταν σε τυχαία σειρά και οι κριτές δεν γνώριζαν από ποιο σύνολο προερχόταν το καθένα. Επίσης, τα αντίστοιχα αξιώματα της οντολογίας δεν δίνονταν στους χρήστες. Σε αυτό το στάδιο και οι δύο χρήστες είδαν τις ίδιες τριάδες κειμένων και τους ζητήθηκε να συγκρίνουν τα κείμενα κάθε τριάδας μεταξύ τους και να τα βαθμολογήσουν ως προς τα παρακάτω κριτήρια. Χρησιμοποιήθηκε μια κλίμακα από 1 έως 5, όπου 1 σήμαινε «διαφωνώ απόλυτα» και 5 σήμαινε «συμφωνώ απόλυτα».

Αναγνωσιμότητα (Readability):

Κάθε ξεχωριστή πρόταση του κειμένου είναι γραμματικά σωστή και φυσική. Κάθε πρόταση είναι ευνόητη δεδομένου ότι ο αναγνώστης είναι οικείος με ιατρικούς όρους.

Δομή (Structure):

Η σειρά των προτάσεων είναι σωστή με βάση της οδηγίες που είχαν δοθεί και στους συγγραφείς της Disease Ontology για τις περιγραφές των τάξεων.

Πληροφορία (Informativeness):

Το κείμενο της τριάδας που παρέχει τις περισσότερες πληροφορίες βαθμολογείται με τον μέγιστο βαθμό, και τα υπόλοιπα αναλόγως του πόσο λιγότερες πληροφορίες εκφράζουν.

Εξετάσαμε τις απαντήσεις των δύο κριτών στα παραπάνω κριτήρια και υπολογίσαμε την συμφωνία τους με βάση τις μετρικές Pearson, Spearman και Kendall. Παρακάτω φαίνονται τα αποτελέσματα ανά κριτήριο για τις development και test τάξεις αντίστοιχα.

Development			
	Readability	**Structure**	**Informativeness**
PEARSON	0.72	0.58	0.45
SPEARMAN	1	1	0.87
KENDALL	1	1	0.82

Test			
	Readability	**Structure**	**Informativeness**
PEARSON	0.38	0.63	0.55
SPEARMAN	1	1	1
KENDALL	1	1	1

Θεωρώντας ότι η συμφωνία ανάμεσα στους κριτές είναι αρκετά καλή, συλλέξαμε επιπλέον κείμενα για 90 τυχαία επιλεγμένες τάξεις από τις development και 90 από τις test (αγνοώντας τις τάξεις που είχαμε ήδη επιλέξει). Τα 540 κείμενα (90 x 3 + 90 x 3) που προκύψαν δόθηκαν και πάλι στους δύο ανθρώπους κριτές. Αυτή την φορά κάθε τριάδα εξετάστηκε μόνο από ένα κριτή. Τα αποτελέσματα και διαστήματα εμπιστοσύνης (95%) για τις development και test τάξεις φαίνονται στους παρακάτω πίνακες. Οι τιμές που είναι με έντονα γράμματα αντιστοιχούν στα καλύτερα αποτελέσματα ανά κριτήριο, ενώ οι σημειωμένες με αστερίσκο τιμές παρουσιάζουν στατιστικά σημαντική διαφορά από τις υπόλοιπες του κριτηρίου, όπως έχει υπολογιστεί μέσω two-tailed t-test ($\alpha = 0.05$).

Εκτελέσαμε t-tests μόνο σε περιπτώσεις όπου τα διαστήματα εμπιστοσύνης επικαλύπτονταν.

Development Data			
	Without Resources	**Original**	**With Resources**
Readability	3.85 ± 0.21	2.78 ± 0.21	**4.96** ± 0.13
Structure	4.22 ± 0.23	3.64 ± 0.22	**4.73** ± 0.10
Informativeness	4.73 ± 0.06*	**4.91** ± 0.15	4.72 ± 0.13*

Test Data			
	Without Resources	**Original**	**With Resources**
Readability	3.66 ± 0.25	2.55 ± 0.24	**4.89** ± 0.16
Structure	4.11 ± 0.26	3.79 ± 0.22	**4.78** ± 0.06
Informativeness	4.65 ± 0.12*	**4.98** ± 0.15	4.65 ± 0.16*

Παρατηρώντας τις βαθμολογίες των κριτών στις development τάξεις βλέπουμε ότι η χρήση των γλωσσικών πόρων που κατασκευάσαμε βοηθάει έντονα στην αναγνωσιμότητα των κειμένων, ειδικά όταν συγκρίνουμε με τις αρχικές (Original) κειμενικές περιγραφές της οντολογίας. Παρόμοια συμπεριφορά βλέπουμε και στην δομή, που είναι λογικό αφού τα αυτόματα συστήματα μπορούν εύκολα να διατηρούν την σειρά των γεγονότων που ορίζουν οι οδηγίες. Τα κείμενα που περιέχουν την περισσότερη πληροφορία ανήκουν στις πρωτότυπες κειμενικές περιγραφές. Αυτό οφείλεται στο ότι η διαδικασία εξαγωγής αξιωμάτων από τις πρωτότυπες κειμενικές περιγραφές αδυνατεί να εξαγάγει κάποιες πληροφορίες (είτε λόγω σφαλμάτων κατά τη συγγραφή των αρχικών κειμενικών περιγραφών είτε επειδή περιέχουν πληροφορίες που δεν μπορούν να παρασταθούν στη νέα μορφή της οντολογίας). Αντίστοιχα συμπεράσματα προκύπτουν και από τα αποτελέσματα στις test τάξεις, κάτι που υποδηλώνει πως η αυτόματη παραγωγή φυσικής γλώσσας έχει καλά αποτελέσματα και σε τάξεις που δεν έχουν ληφθεί υπόψη κατά την κατασκευή των γλωσσικών πόρων. Αν π.χ. προστεθούν νέες τάξεις στην οντολογία δεν είναι απαραίτητη και η συγγραφή επιπρόσθετων γλωσσικών πόρων.

Ακολουθούν επιλεγμένα παραδείγματα κειμένων από την Disease Ontology:

- Για την τάξη «Hypochondrogenesis» (test set) έχουμε τα τρία κείμενα:

  Αρχικό κείμενο (Original):
  An osteochondrodysplasia that has_material_basis_in a mutation in the COL2A1 gene which affects bone growth and results_in a small body, hydrops fetalis, and abnormal ossification located_in vertebral column or located_in pelvis. The disease has_symptom enlarged abdomen.

  Παραγόμενο κείμενο χωρίς χρήση γλωσσικών πόρων (without Resources):
  Hypochondrogenesis is a kind of osteochondrodysplasia. It results in a small body. It located in vertebral column and pelvis. It has material basis in a mutation in the COL2A1 gene. It has symptom enlarged abdomen.

  Παραγόμενο κείμενο με χρήση γλωσσικών πόρων (with Resources):
  Hypochondrogenesis is a kind of osteochondrodysplasia that results in a small body and affects the vertebral column and pelvis. Its symptom is enlarged abdomen. It is caused by a mutation in the COL2A1 gene.

- Για την τάξη «Foodborne botulism» (development set) έχουμε τα τρία κείμενα:

  Αρχικό κείμενο (Original):
  A botulism that involves intoxication caused by botulinum neurotoxins (BoNTA, B, E and F), which are transmitted_by ingestion of food contaminated with preformed toxins, has_material_basis_in Clostridium botulinum A, has_material_basis_in Clostridium botulinum B, has_material_basis_in Clostridium botulinum E and has_material_basis_in Clostridium botulinum F. The infection has_symptom blurred vision, has_symptom diplopia, has_symptom dysarthria, has_symptom dysphonia, has_symptom dysphagia and has_symptom descending muscle paralysis.
  Παραγόμενο κείμενο χωρίς χρήση γλωσσικών πόρων (without Resources):
  Foodborne botulism is a kind of botulism. It has material basis in Clostridium

botulinum F, Clostridium botulinum A, Clostridium botulinum B and Clostridium botulinum E. It has symptom dysphonia, dysphagia, descending muscle paralysis, dysarthria, diplopia and blurred vision. It transmitted by ingestion of food contaminated with preformed toxins.

Παραγόμενο κείμενο με χρήση γλωσσικών πόρων (with Resources):

Foodborne botulism's symptoms are dysphagia, dysarthria, diplopia, blurred vision, descending muscle paralysis and dysphonia. It is transmitted by ingestion of food contaminated with preformed toxins and it is caused by Clostridium botulinum A, Clostridium botulinum B, Clostridium botulinum E and Clostridium botulinum F.

ΚΕΦΑΛΑΙΟ 4 – ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΔΟΥΛΕΙΑ

Ο σκοπός αυτής της εργασίας ήταν η μελέτη βιοϊατρικών οντολογιών και η συγγραφή γλωσσικών πόρων με σκοπό την παραγωγή κειμένων φυσικής γλώσσας που περιγράφουν τις οντολογίες μέσω του συστήματος παραγωγής φυσικής γλώσσας NaturalOWL 2.0.

Αρχικά αναζητήσαμε και εξετάσαμε διάφορες βιοϊατρικές οντολογίες, ώστε να εντοπίσουμε οντολογίες των οποίων οι αυτόματα παραγόμενες περιγραφές να είναι δυνητικώς ενδιαφέρουσες, ευανάγνωστες και χρήσιμες για ειδικούς βιοϊατρικής μη εξοικειωμένους με οντολογίες του Σημασιολογικού Ιστού. Καταλήξαμε στις οντολογίες UMLS και Disease Ontology, για τις οποίες δημιουργήσαμε όλους τους γλωσσικούς πόρους που είναι απαραίτητοι για την παραγωγή ποιοτικών κειμένων με το NaturalOWL 2.0. Στην συνέχεια δημιουργήσαμε αυτόματα κείμενα φυσικής γλώσσας που περιγράφουν τάξεις των δύο οντολογιών. Για την Disease Ontology χρειάστηκε να εξαγάγουμε πρώτα αξιώματα από τις κειμενικές περιγραφές που εμπεριείχε η αρχική μορφή της οντολογίας. Πειράματα έδειξαν ότι τα κείμενα που παράγει το NaturalOWL 2.0 με τους γλωσσικούς πόρους που δημιουργήσαμε είναι καλύτερα από τα κείμενα που παράγει το ίδιο σύστημα χωρίς γλωσσικούς πόρους και καλύτερα εν γένει από τις αρχικές κειμενικές περιγραφές που εμπεριείχε η οντολογία.

Στο μέλλον, θα θέλαμε να εξετάσουμε και άλλες οντολογίες του βιοϊατρικού αλλά και άλλων γνωστικών πεδίων. Ακόμα, ενδιαφέρον θα είχε η συγγραφή των αντίστοιχων γλωσσικών πόρων για την παραγωγή των κειμένων στα Ελληνικά. Τέλος, θα είχε ενδιαφέρον η παραγωγή κειμένων από τις πολυάριθμες (δισεκατομμύρια) τριάδες RDF του Linked Life Data.[9]

[9] http://linkedlifedata.com/

ΑΝΑΦΟΡΕΣ

[And13] I. Androutsopoulos, G. Lampouras, D. Galanis, "Generating Natural Language Descriptions from OWL Ontologies: the NaturalOWL System", Journal of Artificial Intelligence Research Volume 48, pages 671-715, 2013.

[Ant04] G. Antoniou, F. Van Harmelen, "A Semantic Web Primer", MIT Press, 2004.

[Ber01] T Berners-Lee, J Hendler, O Lassila, "The Semantic Web" – Scientific American, 2001.

[Don01] M. O'Donnell, C. Mellish, J. Oberlander, A. Knott, ILEX: an architecture for a dynamic hypertext generation system, Nat. Language Engineering 7 (3) (2001) 225-250.

[Gal06] D. Galanis, "Development of a natural language generation engine for OWL ontologies", MSc thesis, Department of Informatics, Athens University of Economics and Business, 2006.

[Gra08] B. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, U. Sattler, "OWL 2: The Next Step for OWL", Web Semantics: Science, Services and Agents on the World Wide Web 6 (4), 309-322, 2008.

[Isa03] A. Isard, J. Oberlander, I. Androutsopoulos, C. Matheson, Speaking the users' languages, IEEE Intelligent Systems 18 (1) (2003) 40-45.

[Kar07] G. Karakatsiotis, "Automatic generation of comparisons in a natural language generation system", MSc thesis, Department of Informatics, Athens University of Economics and Business, 2007.

[Kons08] S. Konstantopoulos, I. Androutsopoulos, H. Baltzakis, V. Karkaletsis, C.

Matheson, A. Tegos, P. Trahanias, INDIGO: Interaction with personality and dialogue enabled robots, in: 18th European Conf. on Artificial Intelligence, (demos), Patras, Greece, 2008.

[Kons09] S. Konstantopoulos, A. Tegos, D. Bilidas, I. Androutsopoulos, G. Lampouras, P. Malakasiotis, C. Matheson, O. Deroo, Adaptive natural language interaction, in: 12th Conf. of the European Chapter of ACL (demos), Athens, Greece, 2009.

[Mar09] K. Markantoni, "Improvements to the NaturalOWL natural language generation system", BSc thesis, Department of Informatics, Athens University of Economics and Business, 2009.

[Ober08] J. Oberlander, G. Karakatsiotis, A. Isard, I. Androutsopoulos, Building an adaptive museum gallery in Second Life, in: Museums and the Web, Montreal, Quebec, Canada, 2008.

[Rei00] E. Reiter and R. Dale, Building Natural Language Generation Systems. Cambridge University Press, 2000.

[Sha06] Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web revisited. IEEE Intell. Systems, 21, 96-101

[Vog08] D. Vogiatzis, D. Galanis, V. Karkaletsis, I. Androutsopoulos, C. Spyropoulos, A conversant robotic guide to art collections, in: 2nd Workshop on Language Technology for Cultural Heritage Data, Language Resources and Evaluation Conf. 2008.