This project is about attachment and semantic interpretation of noun compounds extracted from a large medical text collection, using semantic information drawn from an ontology (MeSH). I used a classification system such as decision tree for the classification of the attachments and for the bootstrapping of common patterns of noun compounds that can be assigned the same semantic interpretation. In the first Section, I describe the data I analyzed and the system I developed for the extraction of the noun compounds and for the labeling of the nouns with information from the MeSH thesaurus. Section 2 is a description of the classification results obtained with the decision tree and finally, Section 3 covers the semantic interpretation analysis.

1) The Data and the Extraction of Noun Compounds

The text data I used is a huge (781182 lines and 8619699 words) collection of titles and abstracts of medical journals found in Medline, which is a service of the United States National Library of Medicine that provides references and abstracts from 4300 biomedical journals. To give a sense of the data, I report in the Table below the first two titles and abstracts.

<table>
<thead>
<tr>
<th>I 1</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>87049087</td>
<td>.S</td>
</tr>
<tr>
<td>Allied Health Personnel/<em>; Electric Countershock/</em>; Emergencies; Emergency Medical Technicians/<em>; Human; Prognosis; Recurrence; Support, U.S. Gov't, P.H.S.; Time Factors; Transportation of Patients; Ventricular Fibrillation/</em>;TH.</td>
<td>.T</td>
</tr>
<tr>
<td>Refibrillation managed by EMT-Ds: incidence and outcome without paramedic back-up.</td>
<td>.P</td>
</tr>
<tr>
<td>JOURNAL ARTICLE.</td>
<td>.W</td>
</tr>
</tbody>
</table>

Some patients converted from ventricular fibrillation to organized rhythms by defibrillation-trained ambulance technicians (EMT-Ds) will refibrillate before hospital arrival. The authors analyzed 271 cases of ventricular fibrillation managed by EMT-Ds working without paramedic back-up. Of 111 patients initially converted to organized rhythms, 19 (17%) refibrillated, 11 (58%) of whom were reconverted to perfusing rhythms, including nine of 11 (82%) who had...
spontaneous pulses prior to refibrillation. Among patients initially converted to organized rhythms, hospital admission rates were lower for patients who refibrillated than for patients who did not (53% versus 76%, $P = NS$), although discharge rates were virtually identical (37% and 35%, respectively). Scene-to-hospital transport times were not predictively associated with either the frequency of refibrillation or patient outcome. Defibrillation-trained EMTs can effectively manage refibrillation with additional shocks and are not at a significant disadvantage when paramedic back-up is not available.

A Stults KR; Brown DD.
.I 2
.U 87049088
.M Antidepressive Agents, Tricyclic/*PO; Arrhythmia/PP; California; Electrocardiography; Emergencies; Emergency Service, Hospital/*; Female; Human; Length of Stay; Male; Prognosis; Retrospective Studies; Tachycardia, Sinus/DI/TH.
.T Tricyclic antidepressant overdose: emergency department findings as predictors of clinical course.
.P JOURNAL ARTICLE.
.W There is controversy regarding the appropriate utilization of health care resources in the management of tricyclic antidepressant overdosage. Antidepressant overdose patients presenting to the emergency department (ED) are routinely admitted to intensive care units, but only a small proportion develop cardiac arrhythmias or other complications requiring such an environment. The authors reviewed the findings in 165 patients presenting to an ED with antidepressant overdose. They found that major manifestations of toxicity on ED evaluation (altered mental status, seizures, arrhythmias, and conduction defects) were commonly associated with a complicated hospital course. Patients with the isolated findings of sinus tachycardia or QTc prolongation had no complications. No patient experienced a serious toxic event without major evidence of toxicity on ED evaluation and continued evidence of toxicity during the hospital course. These data support the concept that proper ED evaluation can identify a large body of patients with trivial ingestions who may not require hospital observation.
.A Foulke GE; Albertson TE; Walby WF.
.I 3
.U 87049089

Table 1

Figure 1 is a logical view of the system for the extraction of the noun compounds from the Medline file and the labeling with MeSH information. The preprocessing extracts only the titles and abstracts (indicated by ".T" and ".W", see Table 1). The titles and abstract are then tagged using Brill's tagger [Brill]1. The tagged text is the input to a program that extracts noun-compounds. The program can extract compounds with any numbers of constituents but for this project I focused on three-noun compounds.

In the Medline file described above, I found 74966 three-noun compounds.

---

1 In the future I may want to optimize the tagger (i.e. re-train it) to deal with this particular kind of data.
At this point I used the information from MeSH. MeSH (Medical Subject Headings) is the National Library of Medicine's controlled vocabulary thesaurus. MeSH consists of a set of terms or subject headings that are arranged in both an alphabetic and a hierarchical structure. In Table 2, the 15 main tree structures and in Table 3 one example of a node expanded.

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
Table 2

1. Anatomy [A]
   - Body Regions [A01] +
   - Musculoskeletal System [A02] +
   - Digestive System [A03] +
   - Respiratory System [A04] +
   - Urogenital System [A05] +
   - Endocrine System [A06] +
   - Cardiovascular System [A07] +
   - Nervous System [A08] +
   - Sense Organs [A09] +
   - Tissues [A10] +
   - Cells [A11] +
   - Fluids and Secretions [A12] +
   - Animal Structures [A13] +
   - Stomatognathic System [A14] +
   - Hemic and Immune Systems [A15] +
   - Embryonic Structures [A16] +

2. Organisms [B]

3. Diseases [C]

4. Chemicals and Drugs [D]

5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]

6. Psychiatry and Psychology [F]

7. Biological Sciences [G]

8. Physical Sciences [H]

9. Anthropology, Education, Sociology and Social Phenomena [I]

10. Technology and Food and Beverages [J]

11. Humanities [K]

12. Information Science [L]

13. Persons [M]

14. Health Care [N]

15. Geographic Locations [Z]

Table 3

The MeSH vocabulary is continually updated by subject specialists in various areas. Each year hundreds of new concepts are added and thousands of modifications are made. 2000 MeSH includes more than 19,000 main headings, 110,000 Supplementary Concept Records (formerly Supplementary Chemical Records), and an entry vocabulary of over 300,000 terms. [MeSH]
To build the graph structure from a text file describing the MeSH hierarchy, I used the software another student (David Blei) wrote but I had to change part of it for the extraction of the information I needed for my problem. MeSH is read and a graph is built; each node contains all the information about that particular “concept”.

Of course, not all of the 74966 nouns found previously represent concepts in MeSH. From these, I extracted only the ones with at least 1, 2 and all three nouns in MeSH. There are 46137 with at least one noun in MeSH, 15217 with at least 2 and 2056 where all the three nouns are concepts in MeSH. For this project I considered only the latter case. After deleting the duplicates (and, by hand, some of the plurals/singular) I ended up with 934 noun compounds for which all the three nouns are in MeSH.

Here below an example of a MeshNode (migraine) with the information from MeSH. For each concept, we know the definition, all the synonyms (the entry terms for this concepts), the children, the parent, the path to the root, the “in-links” (the concepts that mention this concept in their scope note) the “out-links” (the concepts that this concept mentions in its scope note), if it is a leaf node and the tree positions.

<table>
<thead>
<tr>
<th>Concept: migraine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Def : a subtype of vascular headaches characterized by periodic unilateral pulsatile headaches which begin in childhood adolescence or early adult life and recur with diminishing frequency during advancing years</td>
</tr>
<tr>
<td>TreePositions</td>
</tr>
<tr>
<td>C10.228.140.546.800.525</td>
</tr>
<tr>
<td>C10.228.140.300.800.542</td>
</tr>
<tr>
<td>C14.907.253.937.542</td>
</tr>
<tr>
<td>Synonyms</td>
</tr>
<tr>
<td>migraine_variant</td>
</tr>
<tr>
<td>migraine_variants</td>
</tr>
<tr>
<td>migraine_abdominal</td>
</tr>
<tr>
<td>syndromes_cervical_migraine</td>
</tr>
<tr>
<td>migraine_acute_confusional</td>
</tr>
<tr>
<td>migraines_acute_confusional</td>
</tr>
<tr>
<td>sick_headache</td>
</tr>
<tr>
<td>headaches_sick</td>
</tr>
<tr>
<td>migraine_hemicrania</td>
</tr>
<tr>
<td>headache_sick</td>
</tr>
<tr>
<td>migraines_hemicrania</td>
</tr>
<tr>
<td>hemicranias</td>
</tr>
<tr>
<td>acute_confusional_migraines</td>
</tr>
<tr>
<td>hemicrania_migraines</td>
</tr>
<tr>
<td>... (others)</td>
</tr>
<tr>
<td>Parent of &lt;migraine&gt; = &lt;headache_disorders&gt;</td>
</tr>
<tr>
<td>PathToRoot</td>
</tr>
<tr>
<td>&lt;Top&gt;</td>
</tr>
<tr>
<td>&lt;Diseases&gt;</td>
</tr>
<tr>
<td>&lt;nervous_system_diseases&gt;</td>
</tr>
<tr>
<td>&lt;central_nervous_system_diseases&gt;</td>
</tr>
<tr>
<td>&lt;brain_diseases&gt;</td>
</tr>
<tr>
<td>&lt;headache_disorders&gt;</td>
</tr>
<tr>
<td>&lt;migraine&gt;</td>
</tr>
</tbody>
</table>
The tree positions are the unique identifiers for the concepts and, as they are the most representative features of the MeshNodes, I use them for the labeling of the nouns. For example, “treatment” corresponds to the tree position E02. E is the main tree and 02 represents the 02 part of the E tree, which is Therapeutics.

Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
Diagnosis [E01] +
Therapeutics [E02] +
Anesthesia and Analgesia [E03] +
Surgical Procedures, Operative [E04] +
Investigative Techniques [E05] +
Dentistry [E06] +
Equipment and Supplies [E07] +

The longer the tree position, the longer the path from the root and the more precise the description. For example migraine is C10.228.140.546.800.525, that is, a C (a disease), C10 (Nervous System Diseases), C10.228 (Central Nervous System Diseases) and so on. Migraine, in reality has more than one tree position, (as do many concepts) i.e. it appears in more than one place in the hierarchy. For example:

“migraine headache recurrence”

<table>
<thead>
<tr>
<th>migraine</th>
<th>headache</th>
<th>recurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>C10.228.140.546.800.525</td>
<td>C23.888.592.612.441</td>
<td>C23.550.291.937</td>
</tr>
<tr>
<td>C10.228.140.300.800.542</td>
<td>C10.597.617.470</td>
<td></td>
</tr>
<tr>
<td>C14.907.253.937.542</td>
<td>C23.888.646.487</td>
<td></td>
</tr>
</tbody>
</table>

The different alternatives in the tree positions can play different roles in different contexts, or have several attributes. In the example above, Migraine is under both Vascular Diseases and Central Nervous System Diseases.
For the moment I am considering only one tree position (the first one) but this is definitely something I want to keep in mind for the future.
I therefore label the noun compound "migraine headache recurrence" with C10.228.140.546.800.525  C23.888.592.612.441  C23.550.291.937

In Table 4 a short extract of the noun compounds with the labeling
migraine headache recurrence
C10.228.140.546.800.525 C23.888.592.612.441 C23.550.291.937

blood plasma perfusion
A12.207.152 A15.145.693 E05.680

migraine headache pain
C10.228.140.546.800.525 C23.888.592.612.441 G11.561.796.444

brain stem neurons
A08.186.211 E05.595.402.541.250 A08.663

rat liver mitochondria
B02.649.865.635.560 A03.620 A11.368.702.564

plasma cell membranes
A15.145.693 A11 A10.615

mink disease virus
B02.649.147.625 C23.550.288 B04

heating aluminum fluoride
G03.230.150.300 D01.268.557.050 D01.248.497.158.380

marrow biopsy needles
A15.382.216 E01.450.230.100 E07.612

marrow aspiration needle
A15.382.216 G09.772.521.700.145 E07.612

marrow cell suspensions
A15.382.216 A11 D27.720.280.165.810

bone marrow cells
A10.165.265 A15.382.216 A11

burn scar contractures
C21.866.200 C17.800.120 C05.550.323

stem cell lines
E05.595.402.541.250 A11 G05.331.599.110.708.330.800.400

blood urea nitrogen
A12.207.152 D02.948 D01.362.625

rat prl gene
B02.649.865.635.560 D14.600.600.630 G05.275

breast cancer cells
A01.236 C04 A11

mouse cell lines
B02.649.865.635.500 A11 G05.331.599.110.708.330.800.400

cancer cell lines
In the next Sections I explain how I did the attachments and tried to find a semantic interpretation for the series of 934 of such noun compounds labeled in this way.

2) Attachment

We are ultimately interested in the semantic interpretation of noun compounds and in seeing if compounds with the same pattern (in the form, in our case, of MeSH labels) have the same interpretation. However, starting from scratch, and trying all the combinations to see if they can be assigned a pattern is unfeasible. Just looking at the tree structure and not at deeper levels, there are 15 “starting points” (see Table 1). I actually have only 8 (A-G) in my data but still the combinations are 8^3! Moreover, it is difficult to look for interpretations without having a clear idea of the kinds of interpretations we may find.

My approach was to look at the attachments and see if the classification of noun compounds into left or right attachments could “bootstrap” some kind of pattern to be further analyzed for the semantic interpretation.

I also think there is some useful information in the type of the attachment; the phrases “acute migraine treatment” and “intra-nasal migraine treatment” are a good example of why we need to know the attachment for the interpretation. I think that in this example the interpretation is unambiguous once we know the MeSH category labels and the attachment.

For the classification, I had to label some of the noun compounds by hand.

I use the following labels: la for left attachment [[N N] N] (like door bell manufacturer) an ra for the right attachment [N [N N]] (woman aid worker). Of the 934 noun compounds available, I labeled 204.

2 This kind of data requires medical knowledge, which I don't have, and in lots of the cases I didn't know what attachment to assign. I labelled the noun compounds for which I was more certain of the attachment and I tried at least to be consistent in the decisions but in the future, I would like to get some professional help for both the attachments and the interpretation.
The classification algorithm I use is the decision tree. [Mitchell] lists the characteristics of the problems that decision trees are best suited to:

- Instances are best represented by attribute-value pairs
- The target function has discrete output values
- Disjunctive descriptions may be required
- The training data may contain errors
- The training data may contain missing attribute values

The problem I am tackling seems to suit well this description. Moreover, we are not only interested in the classification, but also in finding out what are the most discriminative features; also, in general, decision trees represent a disjunction of conjunctions of constraints on the attribute values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute test, and the tree itself to a disjunction of these conjunctions [Mitchell]. The idea here is to use a conjunction as a pattern for the noun compounds and look at the corresponding interpretation. I'll come back to this in Session 3.

I divided the labeled compounds into 152 for the training and 52 for the testing. I used Quinlan's algorithm and C4.5 software [Quinlan].

The feature vectors contain the main tree of the 3 nouns (discrete attribute: A, B, C, D, E, F, G) and the their tree positions (continuous values) in this order:

1. first noun tree
2. first tree position
3. second noun tree
4. second tree position
5. third noun tree
6. third tree position

For example, the noun compound “migraine headache recurrence” that we saw earlier is described by the following feature vector:

C, 10.228.140.546.800.525, C, 23.888.592.612.441, C, 23.550.291.937 la

I was also interested in finding out what level of description (that is, in our case, what level of the tree) is the most general and at the same time accurate, or in other words, what is the smallest set of descriptions (i.e. the highest level of the tree) that covers the biggest number of examples. To exploit this, I described the data using different levels of the MeSH Hierarchy. In Table 5 the feature vectors for the different levels in the case of the example “migraine headache recurrence”.

<table>
<thead>
<tr>
<th>Only Tree</th>
<th>Feature vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>C, C, C</td>
<td></td>
</tr>
</tbody>
</table>

3 la is the label
I run the algorithm for all these levels. In the Table 6, an example of the output of the decision tree:

<table>
<thead>
<tr>
<th>Level</th>
<th>C, 10, C, 23, C, 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 2</td>
<td>C, 10.228, C, 23.888, C, 23.550</td>
</tr>
<tr>
<td>Level 3</td>
<td>C, 10.228.140, C, 23.888.592, C, 23.550.291</td>
</tr>
<tr>
<td>Level 4</td>
<td>C, 10.228.140.546, C, 23.888.592.612, C, 23.550.291.937</td>
</tr>
</tbody>
</table>

**Table 5**

<table>
<thead>
<tr>
<th>Evaluation on training data (152 items):</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>53</td>
</tr>
</tbody>
</table>

**Evaluation on test data (52 items):**

<table>
<thead>
<tr>
<th>Size</th>
<th>Errors</th>
<th>Size</th>
<th>Errors</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>8(15.4%)</td>
<td>42</td>
<td>8(15.4%)</td>
<td>(18.9%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(a) (b)</th>
<th>&lt;classified as</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>(a): class ra</td>
</tr>
<tr>
<td>3</td>
<td>(b): class la</td>
</tr>
</tbody>
</table>

**Table 6**

This says that the original tree of 53 nodes misclassifies 17 of the 152 training data. The pruned of 42 nodes tree misclassifies 18 but the program predicts that it will have a higher error rate of 18.9% on unseen cases.

We can also see the results for the testing data. The original tree misclassifies 8 of the 52 testing data (error rate of 15.4%) exactly as does the simplified. Here we’re doing better than the prediction of 18.9%.

The final part of the output is a confusion matrix for the simplified tree on test cases, showing how the misclassification were distributed. There were 32 test cases of class ra, 27 of which are correctly classified as ra while 5 are classified as la. Similarly, 17 of the 20 test cases of class la are classified correctly by the pruned tree and 3 are classified as ra.

In Table7, the results of the classification for different levels of the descriptions. For a description of the algorithm see [Quinlan].
Table 7

| Levels | Pruning  
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(size tree before and after pruning)</td>
</tr>
<tr>
<td>Only Tree</td>
<td>-c default 43-29</td>
</tr>
<tr>
<td></td>
<td>-c 50 43-36</td>
</tr>
<tr>
<td>Level 1</td>
<td>-c 53-19</td>
</tr>
<tr>
<td>*</td>
<td>-c 70 53-42</td>
</tr>
<tr>
<td>Level 2</td>
<td>-c 54-36</td>
</tr>
<tr>
<td>Level 3</td>
<td>-c 59-25</td>
</tr>
<tr>
<td></td>
<td>-c 30 59-34</td>
</tr>
<tr>
<td>Level 4</td>
<td>-c 43-25</td>
</tr>
<tr>
<td></td>
<td>-c 50 43-32</td>
</tr>
</tbody>
</table>

Level 2 performs better for the training set whereas Level 1 performs better on the testing data (line with asterisk); for this reason, I choose this tree for the remaining of the analysis. The results of Table 6 refer to this particular case.

I cannot judge the performance of this classification without comparing it with other methods, which I didn’t.

Also, we see here that Level 1 and Level 2 are good levels of description. However, we can’t infer from here that they are the best levels of description for this kind of data (they are only for the limited number of training examples). Deeper levels need more data and to test weather Level 1 and Level 2 are intrinsically good levels, we should test and train the system with more data and see if these levels still gives a good performance (or which one does).

In general, the purpose of constructing classification models is not limited to the development of accurate predictors, although this is certainly a key concern. Another principal aim is that the model should be intelligible to human beings. I mentioned before that we want to take advantage of the expressiveness of the decision tree for our quest of semantic interpretation.

One way to understand the model constructed by the decision tree is to look at the decision tree itself, or, better, at the simplified tree.

Let’s have a look at the (simplified) decision tree found in the case of Level 1 for which we obtained the best classification results on testing data.

Options:
- Trees evaluated on unseen cases
- Pruning confidence level 70%

\(^4\) -c n determines the pruning (the default is 25%) where small values cause more heavier pruning the bigger value. The number in italics are the tree sizes before and after pruning.
File stem <Data/my_nc/level_1>

ERROR: attribute third tree position has only one value

Read 152 cases (6 attributes) from Data/my_nc/level_1.data

Simplified Decision Tree:

first noun tree = B: ra (33.0/3.7)
first noun tree = E: ra (2.0/1.6)
first noun tree = F: la (0.0)
first noun tree = G: la (4.0/0.3)
first noun tree = A:
   | second noun tree = B: la (0.0)
   | second noun tree = D: la (4.0/0.3)
   | second noun tree = E: la (10.0/0.4)
   | second noun tree = F: la (0.0)
   | second noun tree = G: la (6.0/1.6)
   | second noun tree = A:
      | first tree position <= 4 : ra (7.0/1.6)
      | first tree position > 4 : la (36.0/5.8)
   second noun tree = C:
      | third noun tree = A: ra (9.0/0.3)
      | third noun tree = B: la (0.0)
      | third noun tree = D: la (1.0/0.3)
      | third noun tree = E: la (5.0/0.3)
      | third noun tree = F: la (0.0)
      | third noun tree = G: ra (2.0/1.6)
      | third noun tree = C:
         | third tree position <= 21 : ra (5.0/2.6)
         | third tree position > 21 : la (5.0/0.3)
first noun tree = C:
   second noun tree = B: la (7.0/2.7)
   second noun tree = C: la (5.0/1.6)
   second noun tree = D: la (0.0)
   second noun tree = E: la (0.0)
   second noun tree = F: la (0.0)
   second noun tree = G: la (1.0/0.3)
   second noun tree = A:
      | third noun tree = A: ra (0.0)
      | third noun tree = B: ra (0.0)
      | third noun tree = C: ra (2.0/0.3)
      | third noun tree = D: la (1.0/0.3)
      | third noun tree = E: ra (1.0/0.3)
      | third noun tree = F: ra (0.0)
      | third noun tree = G: ra (2.0/1.6)
first noun tree = D:
   first tree position <= 6 : la (2.0/0.3)
   first tree position > 6 : ra (2.0/0.3)

Tree saved
Table 8

Lets analyze the tree in the cases of N>0 and lets show it in the more readable graph form.

![Tree Diagram]

**Figure 2**

This illustrates how the tree of the first noun is the most discriminative feature and the tree of the second noun (in two cases out of three) the second most discriminative feature.

If the first noun belongs to tree B (Organisms) the attachment is right and the first noun (the organism) will be a modifier of the third noun, like in phrases such as: “mouse ear skin” where mouse: B02.649 (B02 is Vertebrates), ear: A09.246, skin A01.835. Similarly, if the first noun is in

---

5 In number at the left of the class (N/E) (ex: first noun tree = B: ra (33.0/3.7) ), N is the number of training cases covered by the leaf and E is the number of predicted errors if a set of unseen cases were classified by the tree

6 N = 0 means that there are not training cases covered by the leaf and the class assigned is the default class (in our case la) which is the class that contains the most training data not covered by any rules
the E part of the hierarchy (Analytical, Diagnostic and Therapeutic Techniques and Equipment),
the attachment is right and the noun modifies the third noun, like in “puncture wound infections”.
On the other hand if the first noun is a G (Biological Sciences) the attachment is left and the first
name is a modifier of the second noun, like in “growth hormone secretion” (G07.553, D27.505.440,
A12.200) or in “sex chromosome abnormality” (G08.520, A11.223, C16.131).

If the first noun is Anatomy [A], we have to go further down the tree; if the second noun tree is A
again, we have to look at the first tree position: if it is smaller than 4, we have a right attachment,
left otherwise. For example, “bone marrow metastases” (A10.165, A15.382, C04.697) and
“bladder cell membrane” (A05.810.161, A11 A10.615) attach on the left and “forearm bone
fractures” (A01.378, A10.165, C21.866) attaches on the right.
Tree position <= 4 for A includes: Body Regions [A01], Musculoskeletal System [A02], Digestive
System [A03] and Respiratory System [A04], whereas, some of the other subtrees in A that are
bigger than 4 are: Endocrine System [A06], Nervous System [A08], Tissues [A10], Cells [A11],
Fluids and Secretions [A12], Sense Organs [A09] (see Table 3). The two clusters appear to be
different in nature and we can perhaps assume that this difference in their nature determines a
different attachment.

As we can see from the tree, when the noun is either an A or a C (Diseases) we need further rules
for a correct classification.

However, I don't know if this is the case because A and C are more ambiguous (and therefore they
need more decisions) or because I simply happen to have more data for them.
In the next table, for each tree, the number of nouns belonging to that tree in the training-testing set.
and in the whole list of (934) compounds.

<table>
<thead>
<tr>
<th>Tree</th>
<th>training-testing set</th>
<th>Whole list</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>291</td>
<td>1103</td>
</tr>
<tr>
<td>B</td>
<td>63</td>
<td>284</td>
</tr>
<tr>
<td>C</td>
<td>124</td>
<td>450</td>
</tr>
<tr>
<td>D</td>
<td>39</td>
<td>477</td>
</tr>
<tr>
<td>E</td>
<td>48</td>
<td>234</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>46</td>
<td>251</td>
</tr>
</tbody>
</table>

This table shows that I have more training and testing example for A and C; the proportions are
roughly the same for the whole list, except for tree D. In the future, I certainly need to get and
analyze more examples with the D nouns and in general more data.

When classification tasks become more intricate, even simplified trees can grow to unwieldy
proportions. It is true in general that decision tree for real-word problems are very big and, although
accurate, too complex to be understood by anyone.

I briefly mention here C4.5’s attempt of surmounting this comprehension barrier.
C4.5 re-expresses a classification model as production rules, a format that appears to be more
intelligible than trees.
The programs use a simplified form of production rule \( L \rightarrow R \) in which the left-hand side \( L \) is a conjunction of attribute-based tests and the right-hand side \( R \) is a class. C4.5RULES examines the (original) decision tree produced by the C4.5 program and derives from it a set of production rules of the form above. (Note that the RULES program doesn’t only translate the tree into rules (one for each leaf of the tree) for better readability but also checks that the antecedents of individual rules may contain irrelevant conditions and generalize the rules by deleting these superfluous conditions without affecting the accuracy [C4.5].

In the case of the tree of Table 8 and Figure 1, the rules produced are:

```
C4.5 [release 8] rule generator    Fri May 12
14:14:26 2000
-------------------------------
Options:
    Rulesets evaluated on unseen cases
    File stem <Data/my_nc/level_1>

ERROR: attribute third tree position has only one value

Read 152 cases (6 attributes) from
Data/my_nc/level_1

------------------
Processing tree 0

Final rules from tree 0:

Rule 8:
    second noun tree = C
    third noun tree = A
    \rightarrow class ra [87.1%]

Rule 16:
    first noun tree = B
    \rightarrow class ra [85.1%]

Rule 1:
    first tree position <= 4
    second noun tree = A
    \rightarrow class ra [81.6%]

Rule 20:
    first noun tree = D
    first tree position > 6
    \rightarrow class ra [50.0%]

Rule 14:
    second noun tree = E
    \rightarrow class la [87.1%]
```
Rule 10:
second noun tree = C
third tree position > 21
-> class la [84.1%]

Rule 12:
first noun tree = A
third noun tree = E
-> class la [82.2%]

Rule 2:
first tree position > 4
second noun tree = A
-> class la [81.6%]

Rule 13:
second noun tree = D
-> class la [73.1%]

Default class: la

At this point, having the tree and the corresponding rules, we can use the program that allows us to interactively interact with the program. The interactive program request information from the user by prompting for the values of attributes. This can be a handy tool for the classification. Here below some example of interactions.

```
info% ./Src/consult -f Data/my_nc/level_1
C4.5 [release 8] decision tree interpreter Fri May 12 17:02:43 2000
------------------------------------------
ERROR: attribute third tree position has only one value
first noun tree: A
second noun tree: A
first tree position: 3
Decision:
  ra CF = 0.86 [ 0.77 - 1.00 ]
Retry, new case or quit [r,n,q]: n
-------------------------------------------
first noun tree: D
first tree position: 3
Decision:
  la CF = 1.00 [ 0.84 - 1.00 ]
```

The user can also reply with “?” if the attribute value is not known.

```
first noun tree: ?
second noun tree: C
third noun tree: E
first tree position: ?
Decision:
  la CF = 0.76 [ 0.67 - 0.79 ]
```
3) Semantic interpretation

In a previous session I hypothesized that the paths from the root to the leaves of the decision tree can show some kind of patterns, hopefully semantically meaningful, or that, at least, the decision tree can provide some idea of what are most important features and act as a bootstrapping for the discovery of common patterns of noun-compounds that can be assigned the same semantic interpretation.

In other words, the hypothesis is that the syntactic analysis can help narrowing down the semantic search.

The results obtained, seem to confirm this hypothesis.

For example, let's look at some of the paths in the tree of Figure 2.

I sorted from the list of noun compounds only the ones with the combinations corresponding at particular paths on the decision tree.

1. ACA

This corresponds to the labeling: <anatomy> <disease> <anatomy>

The training-testing nouns with this combination are:

<table>
<thead>
<tr>
<th>Noun Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast cancer cells</td>
</tr>
<tr>
<td>A01.236 C04 A11 ra</td>
</tr>
<tr>
<td>bladder cancer cells</td>
</tr>
<tr>
<td>A05.810.161 C04 A11 ra</td>
</tr>
<tr>
<td>colon carcinoma cells</td>
</tr>
<tr>
<td>A03.492.411.495.356 C04.557.470.200 A11 ra</td>
</tr>
<tr>
<td>prostate tumor cells</td>
</tr>
<tr>
<td>A10.336.707 C04 A11 ra</td>
</tr>
<tr>
<td>bladder tumor cells</td>
</tr>
<tr>
<td>A05.810.161 C04 A11 ra</td>
</tr>
<tr>
<td>prostate cancer tissue</td>
</tr>
<tr>
<td>A10.336.707 C04 A10 ra</td>
</tr>
<tr>
<td>lung cancer cells</td>
</tr>
<tr>
<td>A04.411 C04 A11 ra</td>
</tr>
<tr>
<td>colon cancer cells</td>
</tr>
<tr>
<td>A03.492.411.495.356 C04 A11 ra</td>
</tr>
<tr>
<td>brain tumor tissue</td>
</tr>
<tr>
<td>A08.186.211 C04 A10 ra</td>
</tr>
<tr>
<td>colon cancer tissues</td>
</tr>
<tr>
<td>A03.492.411.495.356 C04 A10 ra</td>
</tr>
</tbody>
</table>
We can definitely see a common semantic pattern here:

<Part-of-body> noun3 has a <disease> noun2 in <location> noun1

Here, one can notice that the last A is always either A10 (Tissue) or A11 (Cells). Thinking that this may be an abnormality of the training-testing set (and wanting to know weather there are combinations at different levels of the third tree A that have a different meaning) I sorted the whole list by this pattern and got back the following noun compounds:
It is interesting to note that in all the noun compounds (form the list of 934 compounds) with pattern ACA the last A is A10 or A11.

We can therefore specialize our interpretation:
<cell | tissue> noun3 has a <disease> noun2 in <location> noun1

Also, most of the C are 04 (Neoplasms) so we could also choose to further specialize the interpretation. The trade-off of generalizing vs. specializing (also for the unambiguous cases) is an interesting issue to be exploited. In general, the goal is to use the highest level in the ontology for which the interpretation in unambiguous, but in doing so, one can lose some important information.

2. ACE

Let's now look at ACE for the training-testing set and the whole list

<table>
<thead>
<tr>
<th>training-testing set</th>
<th>Whole list</th>
</tr>
</thead>
<tbody>
<tr>
<td>muscle disease diagnosis</td>
<td>muscle disease diagnosis</td>
</tr>
<tr>
<td>breast cancer prognosis</td>
<td>breast cancer prognosis</td>
</tr>
<tr>
<td>A01.236 C04 E01.789  la</td>
<td>A01.236 C04 E01.789  la</td>
</tr>
<tr>
<td>breast cancer treatment</td>
<td>breast cancer treatment</td>
</tr>
<tr>
<td>A01.236 C04 E02  la</td>
<td>A01.236 C04 E02  la</td>
</tr>
<tr>
<td>hip fracture treatment</td>
<td>hip fracture treatment</td>
</tr>
<tr>
<td>A01.378.592.467 C21.866.405 E02  la</td>
<td>A01.378.592.467 C21.866.405 E02  la</td>
</tr>
<tr>
<td>cell cancer treatment</td>
<td>cell cancer treatment</td>
</tr>
<tr>
<td>A11 C04 E02  la</td>
<td>A11 C04 E02  la</td>
</tr>
<tr>
<td>brain tumor treatment</td>
<td>brain tumor treatment</td>
</tr>
<tr>
<td>A08.186.211 C04 E02  la</td>
<td>A08.186.211 C04 E02  la</td>
</tr>
</tbody>
</table>

ACE is <anatomy> <disease> <Analytical, Diagnostic and Therapeutic Techniques and Equipment>
Noting that E is always E01, E02 or E04 we can assign to this pattern the following semantic interpretation:

\[
< \text{Diagnosis [E01]} | \text{Therapeutics [E02]} | \text{Surgical Procedures [E04]> noun3 on (for, regarding..) <disease> noun2 , where disease is in <location> noun1}
\]

3. ACG

\(<\text{anatomy}> <\text{disease}> <\text{Biological Sciences}>\)

<table>
<thead>
<tr>
<th>Training-testing set</th>
<th>Whole list</th>
</tr>
</thead>
<tbody>
<tr>
<td>brain tumor surgery</td>
<td>brain tumor surgery</td>
</tr>
<tr>
<td>A08.186.211 C04 G02.403.810.762 ra</td>
<td>A08.186.211 C04 G02.403.810.762 ra</td>
</tr>
<tr>
<td>breast cancer specialists</td>
<td>breast cancer specialists</td>
</tr>
<tr>
<td>A01.236 C04 G02.811 la</td>
<td>A01.236 C04 G02.811 la</td>
</tr>
<tr>
<td>colon tumor lines</td>
<td>colon tumor lines</td>
</tr>
<tr>
<td>A03.492.411.495.356 C04 G05.331.599.110.708.330.800.400</td>
<td>A05.810.161 C04 G07.553.481</td>
</tr>
<tr>
<td>bladder cancer growth</td>
<td>bladder cancer growth</td>
</tr>
<tr>
<td>A05.810.161 C04 G07.553.481</td>
<td></td>
</tr>
<tr>
<td>cell pertussis vaccination</td>
<td>cell pertussis vaccination</td>
</tr>
<tr>
<td>A11 C08.730.969 G03.770.670.310.890</td>
<td></td>
</tr>
<tr>
<td>t-cell leukemia lines</td>
<td>t-cell leukemia lines</td>
</tr>
<tr>
<td>A15.382.520.520.604 C04.557.337 G05.331.599.110.708.330.800.400</td>
<td>A03.492.411.495.356 C04.557.470.200.025 G05.331.599.110.708.330.800.400</td>
</tr>
<tr>
<td>colon adenocarcinoma lines</td>
<td>colon adenocarcinoma lines</td>
</tr>
</tbody>
</table>

Here the ACG patterns in training-testing set can be assigned the interpretation

\(<\text{Health Occupations [G02]> noun3 regarding <disease> noun2, disease in <location> noun1}\>

The G02 tree contains terms such Mortuary Practice [G02.438], Nursing [G02.478], Nutrition [G02.513], Optometry [G02.553], Sociology, Medical [G02.790], Psychology, Medical [G02.720] etc that are very different from the G05 terms (genetics) or from the [G07], Physiological Processes. Evidently, the ACG pattern needs to be further divided. It may be possible that including the other terms into the training-testing set, the decision tree algorithm would find a different leaf.

4. ACC

The pattern ACC \(<\text{anatomy}> <\text{disease}> <\text{disease}>\) is further specialized into third tree position >21 and <=21

<table>
<thead>
<tr>
<th>ACC with third tree position &gt;21</th>
<th>ACC with third tree position &lt; 21</th>
</tr>
</thead>
<tbody>
<tr>
<td>heart disease death</td>
<td>lung cancer metastases</td>
</tr>
</tbody>
</table>
ACC with third tree position >21 can be assigned the semantic interpretation:

<disease> noun 3 caused by <disease> noun2 in <location> noun1

The pattern ACC with third tree position < 21 has a very different meaning from the previous one. Here, the semantic interpretation could be:

<abnormal (tissue) condition> noun3 in the form of <disease> noun2 in <location> noun1

A very similar pattern is found for

5. CAC

<table>
<thead>
<tr>
<th>disease</th>
<th>anatomy</th>
<th>techniques and equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>tumor cell necrosis</td>
<td>C04 A11 C23.550.717</td>
<td>ra</td>
</tr>
<tr>
<td>cancer liver metastases</td>
<td>C04 A03.620 C04.697.650</td>
<td>ra</td>
</tr>
</tbody>
</table>

6. CAD

For CAD, <disease> <anatomy> <...Techniques and Equipment> I found:

<table>
<thead>
<tr>
<th>disease</th>
<th>anatomy</th>
<th>techniques and equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>tumor cell reimplantation</td>
<td>C04 A11 E04.936.494</td>
<td>ra</td>
</tr>
<tr>
<td>tumor cell injection</td>
<td>C04 A11 E05.300.530</td>
<td></td>
</tr>
<tr>
<td>cadaver kidney transplants</td>
<td>C23.550.260.224 A05.810.453 E07.945</td>
<td>ra</td>
</tr>
<tr>
<td>cadaver kidney transplantation</td>
<td>C23.550.260.224 A05.810.453 E04.936</td>
<td>ra</td>
</tr>
<tr>
<td>emergency liver transplantation</td>
<td>C23.550.291.781 A03.620 E04.936</td>
<td></td>
</tr>
</tbody>
</table>
There are here two different patterns with two different semantic interpretations:
For first noun tree = C04 the semantic interpretation can be:

<technique> noun3 on <part of body> noun2 with <disease> noun1

For first noun tree = C23 the semantic interpretation can be:

<technique> noun3 on <part of body> noun2 and <technique> noun3 is of <type> noun1

Here I got the label for "tumor cell reimplantation" wrong and probably labeling the other compounds I would have gotten two different paths.

I didn’t find common semantic patterns for all the paths of the tree. For example, the path AA with first tree position<4 gives a lengthy list of noun compounds with no apparent common interpretation and needs to be further divided.

4) Conclusions and future work

I think these are encouraging initial results. Finding a mapping from noun compounds with a certain pattern to semantic interpretation is important for information extraction and inference. There are still a lot of issues to explore. I fist need to do a more extensive and in depth analysis of the data and results presented here. Then, I'd like to consider the noun compounds with only two and one noun in MeSH. For those I won't have the MeSH category labels and I will need a different way to represent them (it's very likely that the semantics of the phrases depends also on the terms not in MeSH but it will be interesting to see up to what degree this is true).
Finding the right level of representation is also very important; the level may be different for different combination of terms, and it can also be different for different nouns in the same compound.
Another thing I’d like to do is to use this initial semantic classification to train a classifier to find the right semantic interpretation.
Also, it would be interesting to see whether these results can be useful for the interpretation of prepositional phrases. Moreover, we need to find a good formal representation for the semantic interpretation and ultimately, to see how we can use all this for the finding/linking of information across texts.

4) References

