A possible solution of the Final Term exam

The medical domain is one of the most challenging for textual NLP but also it is one of the domains that has attracted more attention to the NLP research and development communities.

Roughly two kind of texts are usually used: medical literature (papers, books, ...) and medical reports. The later are clearly more challenging. Million of medical reports are daily produced by different medical professionals (doctors, nurses, radiologists, ...) in different steps of health treatment.

Different types of medical reports differ in their length, degree of formality, terminology used, ...  We will focus on radiology reports that use to be short texts resulting from radiology explorations (X rays, scanning, tomography, ...) that describe the image(s) resulting from the exploration. Consider the following examples:

1) On top of left arm side area in the subcutaneous cell, we observe small rounded shaped  huts with heterogeneous calcifications. Measures 1.9 x 1.2 x 2.7 cm in diameter. Find  seems compatible with BCGitis.

2) Liver: normal size and ecoestructure. Hepatic artery, portal vein and suprahepatics without alterations. Intra- and extrahepatic biliary not extensive. Gallbladder: acalculous.  Pancreas: ecoestructure and size normal. Spleen: normal size and ecoestructure. Longitudinal diameter: 8 (cm) retroperitoneum vascular without alterations. No adenomegalias detected.

3) Static  study:  both hips centered. Osteocartilaginosis  adequate coverage. Left femoral nuclei with incipient ossification. Dynamic study both hips stable. CONCLUSION: ultrasound characteristics of normal hips.

4) A magnetic resonance imaging study will be scheduled as an outpatient later to rule out a small vascular malformation.

Answer the following questions:

1. Characterize the sublanguage used in radiology reports. Comment briefly the problems (compared with general texts) that occurs in this genre of documents.

2. List the kind of information and knowledge that could be extracted from these documents. Use example 1) above for illustrating your answer.

3. How to deal with the information extraction tasks presented in 2)? Which linguistic tasks are involved? Describe the tasks focusing on the challenges derived from the characteristics presented in 1).

4. Which knowledge sources are needed for 3)? How could be obtained? Quantify roughly the size of these knowledge sources.

5. Discuss about a multilingual setting. Could some of the resources in 4) for a language be used (perhaps with some limited human intervention) for other languages.

All the five questions are equally weighted (2 points each)

1. **Characterize the sublanguage used in radiology reports. Comment briefly the problems (compared with general texts) that occurs in this genre of documents.**

- Some of the sentences are well formed (for instance example 4 above) and, so, can be processed with standard NLP processors (Freeling, Stanford, OpenNLP, Senna, ...). In most cases, however, reports consist on a concatenation of chunks (mostly nominal) difficult to be processed without genre-specific processors ( for instance "Liver:" in example 2, "Static study:" in example 3).

- Capitalization is not coherent, some tokens are fully capitalized, some others semi-capitalized, some other just the first letter, some others lowercased. Example: BCGitis, CONCLUSION.

- Use of headers, as "Liver:", "Pancreas:" in example 2.

- Terminological items (that have to be detected and classified). Some terms are embedded into other more complex:

    - Body parts: "top of left arm side area", "arm", "Hepatic artery", "artery", "subcutaneous cells".

    - Diseases: "BCGitis", "vascular malformation"

    - Finds: "huts", "calcification", "ossification", "alterations"

    - Imaging terms: "magnetic resonance imaging", "ultrasound characteristics".

- Some terms include non terminological qualifiers or modifiers (adjectives, prepositional modifiers, nominal compounds, ...:

    - "Small rounded shaped nuts"

- Frequent use of negations with highly variate scope:

    - "without alterations", "not extensive", "No adenomegalias detected"

- Sometimes use of hedges. Both negations and hedges prevent the extraction of facts.
    - “seems compatible with”.
- Frequent use of measures, units, ...
    - “1.9 x 1.2 x 2.7 cm in diameter”
    - “Longitudinal diameter: 8 (cm)”
- Not in the examples provided but in many others, personal information (name of the patient, address, ...) is included and has to be located and removed (the text has to be anonimized for allowing its use. An anonimizer just masks the occurrences of personal information by general tags (“John Smith, young of 23 years from Michigan” should be changed into, for instance, “<PER>, young of <AGE> years from <LOCATION>”

2. **List the kind of information and knowledge that could be extracted from these documents. Use example 1) above for illustrating your answer.**

- The information to be extracted depends, obviously, of the intended use of the resource. Usually the final users should be medical professionals. We can consider two families of applications:
    - Providing to the medical professionals the relevant content of the radiologic report in a more friendly way: summarizing, highlighting the relevant parts, displaying warnings, .., in order to help in the human diagnosis.
    - Automatically extracting useful and relevant facts that could be used by an expert system that proposes an automatic diagnostic .
- Information that could be extracted:
    - Semantic tagging (equivalent to NERC in general texts)
        - A tagset should be defined. For instance, BP = body part, DS = disease, DR = drug or pharmaceutical product, F = clinical finding, M = measure, IM = imaging:
        - For instance, for the first example the following mentions could be obtained and tagged (I tag only the widest scope of each mention):

    - On [top of left arm side area][BP] in the [subcutaneous cell][BP], we observe [small rounded shaped  huts][F] with [heterogeneous calcifications][F]. Measures [1.9 x 1.2 x 2.7 cm in diameter][M]. Find  seems compatible with [BCGitis][DS].

  - Relations between the entities tagged:
    - Seems_compatible([small rounded shaped  huts][F], [BCGitis][DS]).

  - Terminology extraction: collection of terms occurring in the whole collection of reports.

  - Measures could be extracted, normalized, and represented in an interpretable way (by human and machine):
    - <entity, body part, type of measure, value, unit>
    - For instance:
      - <[subcutaneous cell][BP], [small rounded shaped huts][F],  diameter , 1.9 x 1.2 x 2.7 , cm> (example 1)
      - < [Spleen] [BP], [Spleen] [BP],  longitudinal diameter , 8 , cm> (example 2)

  - Facts (usually predicate + arguments):
    - “both hips centered”  =>  centered(hip_left), centered(hip_right)
    - “Spleen normal size” => size(Spleen, normal)
    - “Spleen normal echostructure” => echostructure (Spleen, normal)

3. **How to deal with the information extraction tasks presented in 2)? Which linguistic tasks are involved? Describe the tasks focusing on the challenges derived from the characteristics presented in 1).**

- Obviously we need a tokenizer able to deal with the specific kind of terms described in 1), special difficulty is placed by measures & units.

- A semantic tagger (similar to a NERC) but adapted to the tagset described in 2) is needed.
- No parsing information seems to be needed.
- We need chunkers (at least a nominal chunk) trained from a collection of reports huge enough. Training the chunker using PTB or Ancora is useless, we need to retrain it with domain specific material.
    - Chunks could be (and use to be) nested:
        - “top of left arm side area”, “left arm side area”, ““left arm”, “arm”, ...
        - Several kinds of modifiers should be detected:
            - Adjectives (“heterogeneous calcifications”)
            - prepositional modifiers (“top of the left arms”)
            - nominal compounds (“portal veins”)
- measure/unit extractor and normalizer
    - Perhaps using FST technology
- A co-referencer is needed:
    - [small rounded shaped huts][F] and “Find” co-refer.
    - Other forms of co-reference do not occur in the examples but surely could occur in other documents (pronominal, ellipsis, ...)
- Relation extraction has to be faced. Thus, probably, a Semantic role labeller (SRL) would be used. Relevant predicates (linguistically realized as verbs or de-verbal nominalizations) should be detected.
- A system for detecting negation and its scope is needed for distinguishing between asserted facts and negated ones.
- A system for detecting hedges and its scope is needed for distinguishing between asserted facts and opinions or conjectures.
- Perhaps in a multilingual setting a language identifier could be used (for instance, in Catalan health centers, reports are written in Spanish, Catalan, or a merge of both languages ).

4. **Which knowledge sources are needed for 3)? How could be obtained? Quantify roughly the size of these knowledge sources.**

Radiology reports are written in NL, so, some general purpose NLP tools and resources could be used. The specificity of the domain (Medicine) and genre (radiology reports) prevent us from a direct use of such resources that should be tuned.

Focusing in the specific resources we will need gazetteers for covering the vocabularies of BP = body part, DS = disease, and DR = drug or pharmaceutical product. Ontologies, terminologies and lexicons for these entities exist for English, Snomed, UMLS, RadLEx, BioPortal, MedLine, ...

Snomed contains for English (there are smaller versions for other languages, including Spanish) more than 100 Kw, UMLS contains about 1 Mw and RadLex, specialized in radiology, about 40,000 entries. Medline points to several millions of medical papers. BioPortal allows the access to 300 ontologies in the medical/genetic domains.

For the other tags (F, M, IM) lexicons have to be collected automatically or manually from huge enough collections of documents. Also negation and hedges triggers should be collected from the texts. I guess that about 10,000 words should be enough for this task. Relevant predicates could be selected in a similar way.

Mapping verbs to their de-verbal nominalizations (e.g. extract => extraction) can be done easily with the help of WordNet, resource that, fortunately, has a rich coverage of the medical terminology.

Chunkers and co-reference taggers should be tuned to the new domain (usually adding a small collection of domain and genre specific documents to the original training material is enough).

5. **Discuss about a multilingual setting. Could some of the resources in 4) for a language be used (perhaps with some limited human intervention) for other languages.**

A direct mapping of the tools and resources from English to other languages has little sense. Most of the gazetteers exist only for English (with some exceptions, there is a SNomed version for Spanish). Also Medline and UMLS contain information for other (few) languages. A good (but limited) source of multilingual links is Wikipedia (through the interwiki links) or dbpedia. Also WordNet can be used for the cases when a mapping between the corresponding wordnet and the English WordNet exists. Anyway, all these possibilities are far to be completed because of the general scope (not domain specific) of the sources. Using Machine Translation tools (like Google) does not work well for this domain/genre. So I am afraid that most of the work has to be repeated for each of the languages implied.