Layered MorphoSaurus Lexicon Extension

Problem

- Confuse and arbitrary synonym classes of non-medical concepts
- High ambiguity of general (non-terminological) language
- Maintenance cost not justified by search engine performance
- Risk of precision loss due to general language terms

Solution 1 (radical)

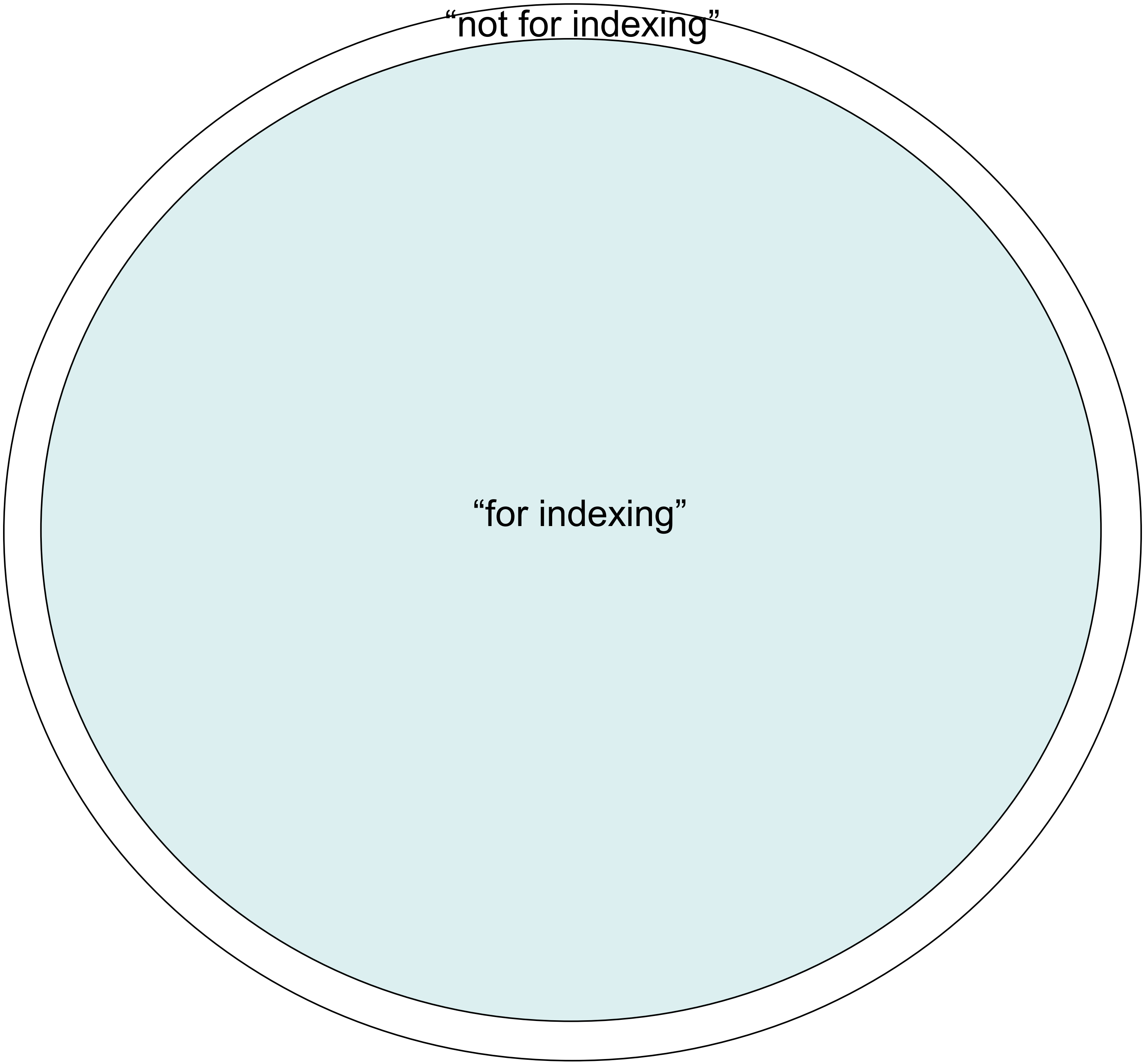
- Abandon present synonym class architecture, consider only stem variations as synonym
- Example:
  - remove: {derm, haut}, {hyper, high}
  - maintain: {diagnos, diagnost}, {bruch, bruech}
- Expected outcome:
  - Monolingual IR: Precision + Recall -
  - Cross-Language IR: seriously hampered
- Make up strategy: Multiword Thesaurus
  - maps MID sequences: [m1 m2 m3] -> [m7 m8 m9]

Solution 2 (semiradical)

- No alteration of lexicon structure
- Customization of thesaurus export:
  - Option 1: as is (e.g. for cross-language retrieval)
  - Option 2: automatically generate new Eq classes on the fly:
    - ignore “has-sense”
    - crack existing eq classes: Example:
      MID1 = {a, b, c, d, e, f, g}
      split into MID1’ = {a}; MID1’’ = {b}; MID1’’’ = {c,d,e}; MID1’’’’ = {f,g};
      being d and e variants of the stem c, and g a variant of stem f
    - Criterion for stem variants:
      - lexemes are in the same eq class (before splitting)
      - lexemes have a Levenshtein edit distance below threshold
- Advantage:
  - choice between Full and Lite version maintained
  - completely automated generation of Lite out of Full

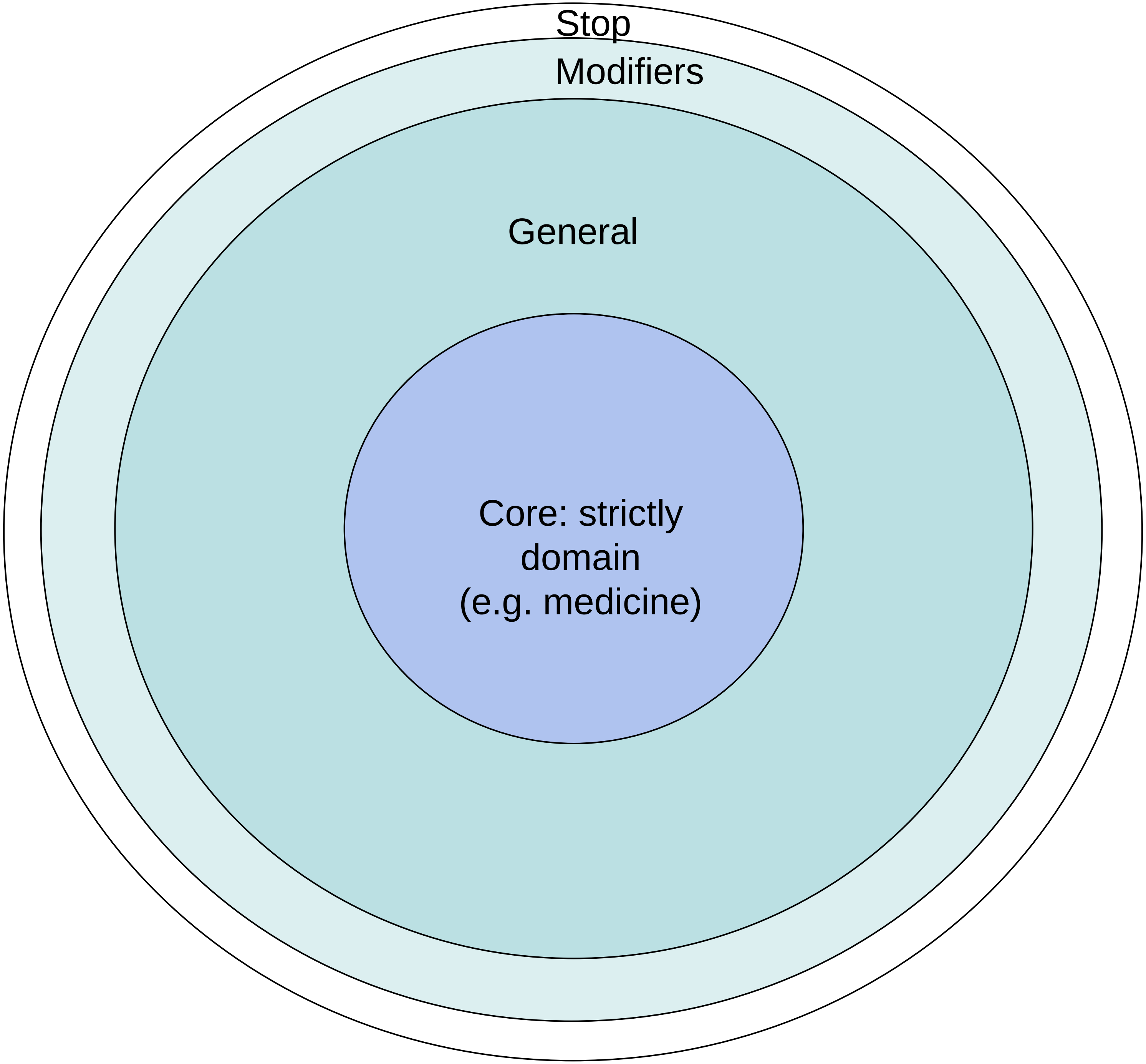
Layering of the lexicon

- Hypothesis: MIDs play different roles in a domain specific IR context
- So far we have two layers:
  - “for indexing”
  - “not for indexing”



“not for indexing”

“for indexing”


Stop

Modifiers

General

Core: strictly domain (e.g. medicine)

MID Characterization

S = Stop	Irrelevant for document indexing and retrieval	Personal pronouns, auxiliary verbs, some prefixes, most derivation suffixes
M = Modifier	Meaningful and discriminative in local context only Depend on other words Never constitute solely a user query, very low idf	negation particles, many adjectives, quantifiers, graduation, modality
G = General	General language terms that cannot be assigned to any specific domain terminology	Most verbs and nouns that are found in a normal lexicon
C = Core	Domain specific terms Domain queries should contain at least one C term	Generally nouns, can only be found in a domain specific lexicon

How to classify MIDs (or subwords ?) by layers

- S: already done (“not for indexing”)
- C candidates: MIDs from UMLS (subset) indexing
- G $\cup$ M candidates: MIDs from WordNet indexing
- Separation of M: manually check frequent, nonmedical MIDs extracted from nonmedical corpus

Differentiated treatment in IR context

- M: completely ignore outside local context
  - Hyperkalemia -> #highgrade[M] #potassium[C] #blood[C]
  - retrieve document with:
    “elevated potassium…..blood” ->
    #highgrade[M] #potassium[C] ……………. #blood[C]
  - ignore document with
    “moderate hypernatriemia but normal potassium.….blood”
    #moderate[M] #highgrade[M] #sodium[C] #blood[C]
    #normal[M] #potassium[C] ….. #blood[C]:
    #potassium[C] outside the scope of # highgrade[M]
- G: similar treatment, broader scope (window), if outside scope: downranking but not excluding

Differentiated lexicon redesign by layer

- Layer M: allow big and unspecific classes
- Layer G: apply Solution 1 or 2
- Layer C: continue fine-grained lexicon modelling including semantic relations
- Much more possibilities of adjustment of retrieval system by requirements
  - Whether to apply solutions 1 or 2
  - On which thesaurus layers
  - Whether or not apply phrase search or near operator when dealing with “M” classes.