Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Two hierarchical text categorization approaches for BioASQ semantic indexing challenge

Francisco J. Ribadas
Víctor M. Darriba

Compilers and Languages Group
Universidade de Vigo (Spain)
http://www.grupocole.org/

ribadas@uvigo.es
darriba@uvigo.es

Luis M. de Campos
Alfonso E. Romero

Research Group of Uncertainty Treatment in Artificial Intelligence
Universidad de Granada (Spain)
http://decsai.ugr.es/gte/

lci@decsai.ugr.es
aeromero@cs.rhul.ac.uk

BioASQ challenge 2013

Valencia, September 2013

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

1. Motivation and objectives

2. Description of our systems
   - HACE framework
   - Rebayct

3. Our participation in BioASQ
   - Adaptations
   - Preprocessing
   - Tested configurations

4. Conclusions and future work
   - Conclusions
   - Future work

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Motivation

- Joint work of CoLe group (Univ. of Vigo) and UTAI group (Univ. of Granada)
- Previous independent work on related small/medium size problems
  - Legal documents: { parliamentary initiatives (UTAI); public grants and subsidies (CoLe) }
  - Medium size thesauri (EUROVOC + custom thesaurus)
  - Both dealing with Spanish texts (also Galician for CoLe)
  - Minimal linguistic processing (no tagging, no lemmatization, no NER)
- Thesarus topic assigment as a hierarchical text categorization problem
  - Top-down scheme using a local classifier per node approach (CoLe)
  - Bayesian network induced from thesaurus hierarchy (UTAI)

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Objectives

1. Test the scalability of our proposals with large real-world data
   - BioASQ Task 1A: *Large-Scale Online Biomedical Semantic Indexing*
   - Large hierarchy of descriptors and large training set $\rightarrow$ size and time restrictions
2. Evaluate the suitability of a pure text categorization approach for semantic indexing with MeSH $\rightarrow$ minimal linguistic processing
   - Very different domain $\rightarrow$ complex terminology

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Origins of our systems

HACE framework

Text categorization on a public grants/subsidies collection

- Small size custom thesaurus ($\approx$ 1800 descriptors)
- Medium/large size documents
- A few labeling inconsistence in training documents
- Additional requirement: return ”many” results (search for a high recall)
  - Human curators will postprocess system output

Rebayct

Text categorization on a parliamentary initiatives collection

- EUROVOC thesaurus ($\approx$ 4000 descriptors)
- Very small size documents (1-2 paragraphs)
- Additional requirement: return ”many” results (search for a high recall)
  - Human curators will postprocess system output

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

HACE framework (I)

Generic framework for hierarchical categorization *(under development)*

Top-down *Local Classifier per Node Approach*

- Local binary classifier trained for each node in the hierarchy
- Is current node (or its descendants) pertinent as label?
- Pachinko-like top-down traversal of local classifiers

Able to deal with tree and DAG structured taxonomies

Plug-in architecture with several components for:

- selecting sets of positive examples with a bottom-up procedure
- selecting sets of negative examples
- feature selection at each local model (IG, Chi squared, ...)
- classification algorithm to perform the ”routing” decisions at each local model
- dealing with unbalanced classes (weighting, boundary negative examples, split negative example set in an ensemble of classifiers)

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

HACE framework (II)

Specific features for large scale hierarchical text categorization

- Textual features computation backed by a Lucene index
- Bottom-up positive example selection (from positive examples sets in descendant nodes)
  - avoid unmanageable training sets on top levels
  - random selection among descendant positive examples
  - k-means clustering based selection $\rightarrow$ selecting examples close to centroids
- Guided top-down search using a simplified k-nearest neighbours
  - query the Lucene index to get a set of promising labels from most similar documents
  - top-down search starts at grandparents nodes $\rightarrow$ avoids premature discard of useful paths

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

HACE framework (III)

Contextual routing decisions (not tested in BioASQ data)

- **Objective:** try to reduce false negatives (mainly in top levels)
- Two classifier per node model { content based / context based
- Exploiting bottom-up information (*metafeatures*) coming from content based routing decisions performed by descendant nodes (and optionally by ancestor and sibling nodes)
- Roughly inspired by *classifier chain* approaches in multilabel classification
  - Adds to content based features a set of *metafeatures* about decisions of surrounding models
- Moderate performance improvements + high training/classification cost

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Rebayct (I)

Builds a Bayesian network using:

1. thesaurus hierarchical structure
2. terms (tokens) taken from { descriptor labels / non-descriptor labels }
3. terms (tokens) taken from training documents

Elements

- Concept nodes (representing thesaurus concepts/nodes)
- Descriptor and Non-Descriptor nodes (representing descriptor and non descriptor labels)
- Term nodes (representing words [tokens])

Every concept node C linked with three virtual nodes

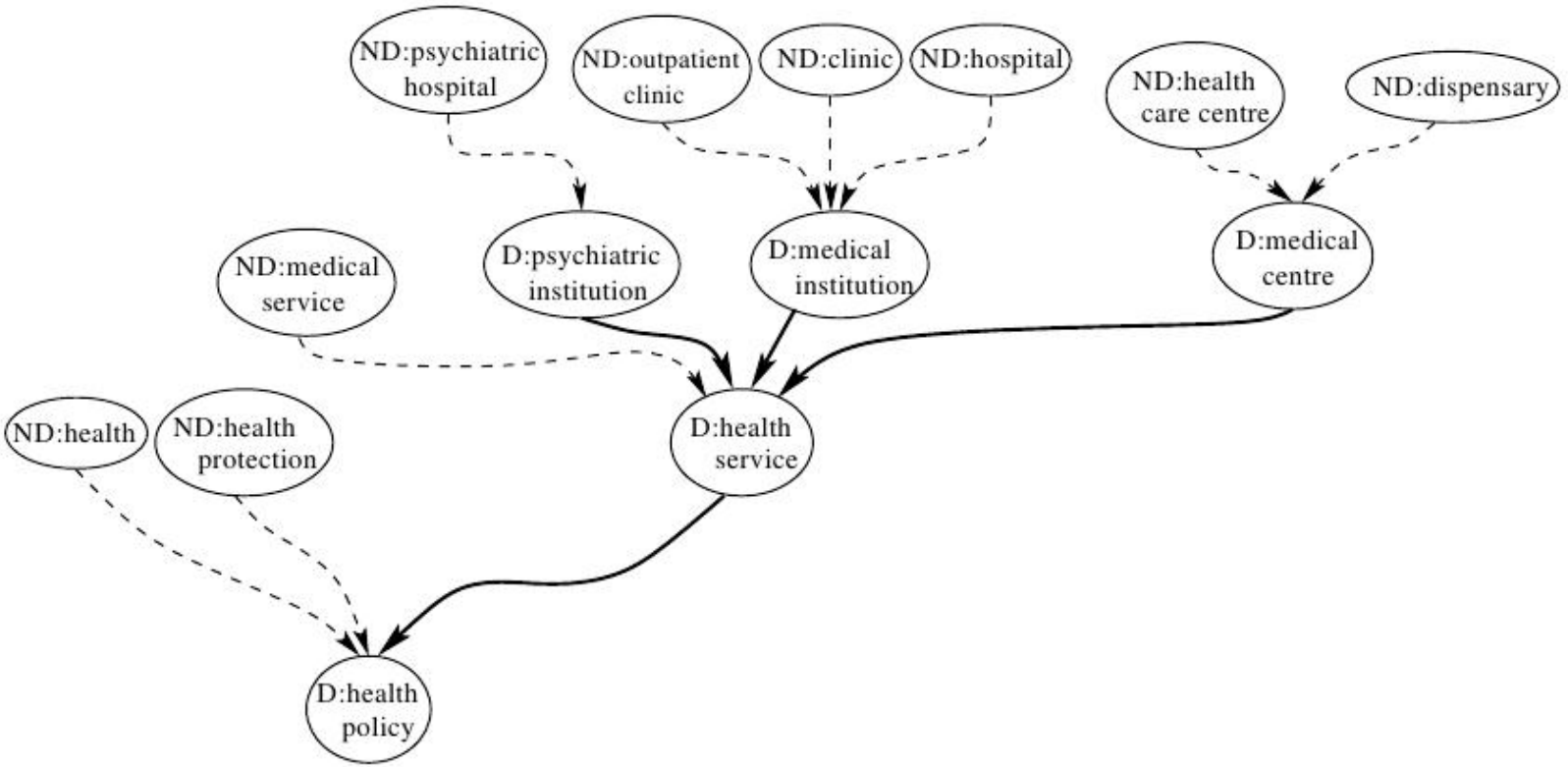
- H_C: info. from BT (*Broader Term*) relationships in the thesaurus
- E_C: info. from descriptor and non-descriptor labels (synonymy)
- T_C: info. from training documents

Efficient OR-gate model to define conditional probabilities.

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Rebayct (II)

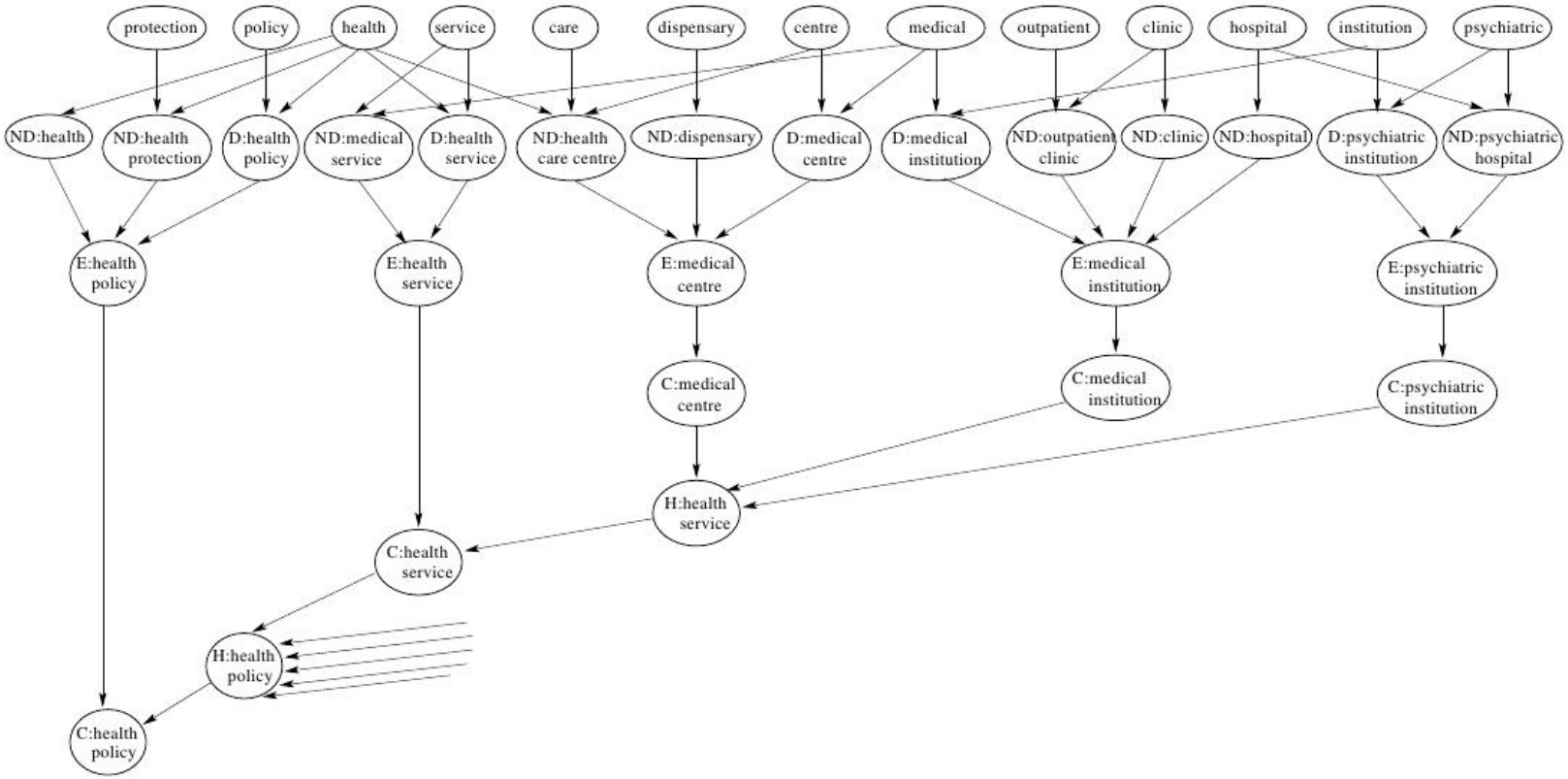
Thesaurus fragment (D: descriptors, ND: non-descriptors)



Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Rebayct (II)

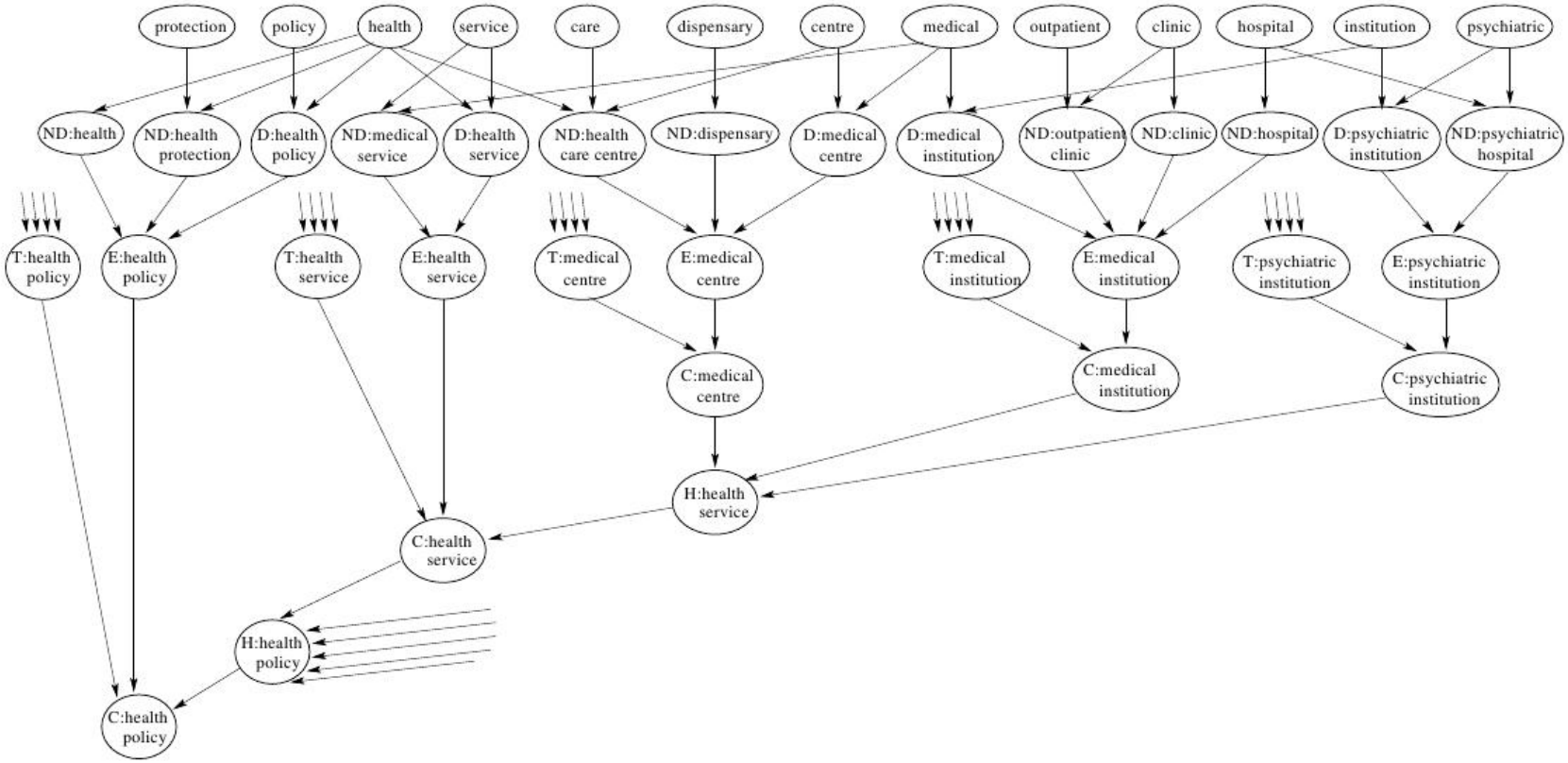
Bayesian network build from thesaurus hierarchy and descriptor and non-descriptor equivalence relationships



Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Rebayct (II)

Bayesian network after adding terms from training documents



Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Adaptations for BioASQ challenge

Own concept taxonomy with a DAG structure extracted from 2013 XML version of MeSH

- Hierarchical relationships created from *TreeNumber* elements
- *TreeNumbers* describe the places a MeSH descriptor occupies inside the 16 concept taxonomies

Results

- DAG with 26,702 nodes (excludes 151 descriptors from subhierarchy V)
- 36,647 parent-child relationship with only two cycles
  - descriptors D009014 (*Morals*) + D004989 (*Ethics*)
  - descriptors D006885 (*Hydroxybutyrates*) + D020155 (*3-Hydroxybutyric Acid*)
- 108,117 related terms (synonyms or lexical variants) [non-descriptors for Rebayct]

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Problem: spurious relationships in the final DAG taxonomy

Example: *eye* as { part of *face* / a *sense organ* } leads to consider *eyebrows* as an element of a *sense organ*

```
Body Regions;A01
  ...
  Head;A01.456
    Ear;A01.456.313
    Face;A01.456.505
      Cheek;A01.456.505.173
      Chin;A01.456.505.259
      Eye;A01.456.505.420
        Eyebrows;A01.456.505.420.338
        Eyelids;A01.456.505.420.504
          Eyelashes;A01.456.505.420.504.421
      Forehead;A01.456.505.580
      Mouth;A01.456.505.631
        Lip;A01.456.505.631.515
      Nasolabial Fold;A01.456.505.682
      Nose;A01.456.505.733
      Parotid Region;A01.456.505.750
    Scalp;A01.456.810
    ...
```

```
Sense Organs;A09
  Ear;A09.246
    (44 descendants)
  Eye;A09.371
    Anterior Eye Segment;A09.371.060
      (20 descendants)
    Anterior Capsule of the Lens;A09.371.061
    Axial Length, Eye;A09.371.199
    Eyelids;A09.371.337
      Conjunctiva;A09.371.337.168
      Eyelashes;A09.371.337.338
      Meibomian Glands;A09.371.337.614
    Lacrimal Apparatus;A09.371.463
    ...
    Retina;A09.371.729
      (21 descendants)
    Sclera;A09.371.784
      Tenon Capsule;A09.371.839
    Uvea;A09.371.894
      (6 descendants)
  Nose;A09.531
  ...
```

No disambiguation info. on training data to avoid it

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Preprocessing for BioASQ challenge

Only elementary text processing on train and validation documents.

- stop-word removal using a standard English stop-word list
- default English stemmer from the Snowball project

Also: alternative collection extracting word bigrams from descriptor labels and document text after stop-word removal

- simple way to capture some complex terms (but far from perfect)

Rebayct scalability limitations

- Extract a reduced training set of 1,242,670 documents (10 %)
- $\approx$ 50 more representative documents for every descriptor taken from Lucene index
- Split into 5 groups (248.534 training instances) to train

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Tested configurations (I)

HACE framework

Previous parameter tunning phase with a small dataset from subhierarchy *"[C] Diseases"*

1. Effectiveness of bottom-up positive example selection
   - random document selection *vs.* *k*-means based document selection
   - up to 500, 1000 and 2000 selected instances per node
2. Effectiveness of guided top-down classification
3. Usefulness of word bigram based features *vs.* single token features

Results

- *k*-means based document selection has better performance, but training time is almost twice
- better results using greater amounts of positive instances
- guided top-down search improved classification time and quality
- word bigrams did not appear to help

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Tested configurations (II)

Rebayct

1. Effectiveness of aggregating the results of 5 models *vs.* single model performance
2. Usefulness of word bigrams as instance features using a single Rebayct model

Conclusions

- marginal improvements when results of 5 Rebayct models were combined
- great improvement due to word bigram representation

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

BioASQ Task 1A participation

Submitted configurations

- HACE1: HACE framework with
  - *k*-means bottom-up positive example selection (2000 documents per node)
  - IG as local feature selection (100 features)
  - SVM as local content based classifier
- HACE2: same as HACE1 employing a guided top-down search approach
- HACE2-NE: same as HACE2 using word bigrams as textual features
- REBAYCT: combination of 5 Rebayct models trained with 5 splits of the reduced training set
- REBAYCT2: single Rebayct model using word bigram alternative collection

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Conclusions

- Two different hierarchical text categorization systems evaluated in 2013 BioASQ challenge
- Quite far from top performance systems in the challenge, but some improvements were done from the original systems employed in first batch
- With minor changes our systems were able to deal with a problem larger than the ones that originated them
- Tested configurations and BioASQ challenge results give us some insights for improvement
- Large training data sets make unnecessary to employ sophisticated machine learning approaches?
  - Training text contribution in Rebayct classifications was more important than structural and descriptor label contributions
  - Guided top-down search in HACE employs a very simple kind of k-nn prefiltering.

Objectives
Our systems
HACE
Rebayct
Our participation
Adaptations
Preprocessing
Configurations
Conclusions
Conclusions
Future work

Future work

- Advanced NLP processing of documents and descriptor labels (POS tagging, NER, ...)

HACE framework

- Make a more deep parameter tunning
- Exploit the guided top-down search approach
- Exploit (and optimize) the context based ”routing” approach

Rebayct

- Evolve to a sort of active learning approach with better training document selection
- Exploit word bigrams and powerful text processing approaches to improve quality of input data