CoLe and UTAI participation at the 2014 BioASQ semantic indexing challenge

Francisco J. Ribadas
Victor 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 2014 **QA-CLEF 2014**

Sheffield, September 2014

1. Motivation and objectives

2. Description of base systems
   - HACE framework
   - REBAYCT

3. Our participation at BioASQ 2014
   - Document preprocessing
   - Combining ensembles of classification models

4. Results and conclusions
   - Conclusions

Motivation and objectives

Second participation at BioASQ semantic indexing challenge (**task 2A**)

- joint work of CoLe group (Univ. Vigo) and UTAI group (Univ. Granada)

Thesarus topic assignment as a hierarchical text categorization problem

- top-down scheme with local classifier per node approach (CoLe)
- Bayesian network induced from thesaurus hierarchy (UTAI)

Objectives in 2nd BioASQ challenge

Same base systems, try to improve document preprocessing and label postprocessing

- test more powerful linguistic processing on input documents
- evaluate strategies to combine ensembles of classification models

HACE framework (I)

Generic framework for hierarchical categorization able to deal with tree and DAG structured taxonomies

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

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
- classification algorithm to perform ”routing” decisions at each local model
- dealing with imbalanced classes

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 sets of positive examples at descendant nodes)
- 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 → avoids premature discard of useful paths

In both BioASQ editions we have employed a fairly aggressive configuration for term and document selection

- our main aim with this framework is to be able to train local models in memory using
  - small sets of highly descriptive terms
  - small positive documents sets representing current node descendants

REBAYCT (I)

Builds a Bayesian network using:

1. thesaurus hierarchical structure
2. terms (tokens) taken from $\left\{\begin{array}{l}\text{descriptor labels}\\ \text{non-descriptor labels}\end{array}\right.$
3. terms (tokens) taken from training documents

Elements

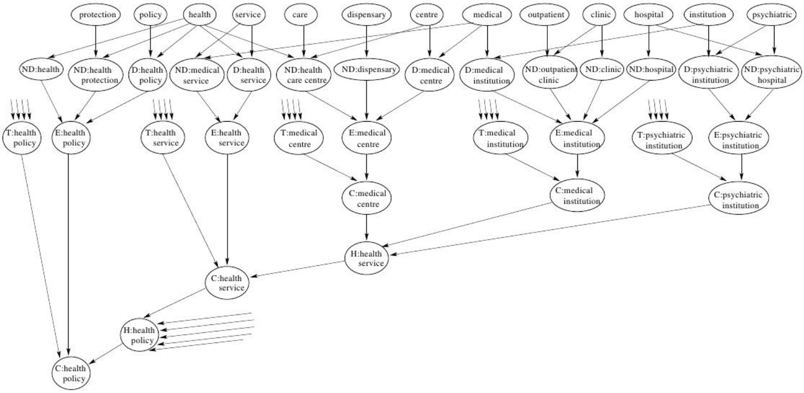
- Concept nodes (representing thesaurus concepts/nodes)
- Descriptor and Non-Descriptor nodes (representing descriptor and non descriptor labels)
- Term nodes (representing words [tokens])

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

Rebayct (II)

Bayesian network after adding terms from training documents



Document preprocessing

- Previous BioASQ participation: simple document preprocessing (stop-word removal + stemming)
- Complex domain (specific language + rich terminology) suggest the suitability of employing sophisticated linguistic processing
- Balance huge categorization task *vs.* NLP processing cost and complexity
- **Our approach:** approximate powerful NLP processing using cheaper approaches
  1. local abbreviation expansion
  2. term selection based on morphosyntactic information
  3. approximate multiword terms with selected bigrams

Expanding local abbreviations

- High writing quality in BioASQ collection documents
    - MEDLINE abstracts written and reviewed by experts
    - consistent usage of a set of well established conventions for scientific writing
- Use of abbreviations and acronyms with local meaning within a paper
    - fairly simple and easy to detect patterns of use
    - usually closely related with paper main contents
- Schwartz and Hearst algorithm (`http://biotext.berkeley.edu/software.html`)
    - authors report report 96 % precision and 82 % recall in custom MEDLINE evaluation set
    - in our preprocessing phase abbreviation expansion in 10.1 % of processed documents (from 4.8 M docs. in training set B)

short form	long form	
IFP	*inflammatory fibroid polyps*	5 occurrences within the abstract
ACTH	*adrenocorticotropic hormone*	5 occurrences within the abstract
PEM	*protein energy malnutrition*	2 occurrences within the abstract
QoL	*quality of life*$^{(*)}$	14 occurrences within the abstract
	(*) An actual MeSH descriptor (ID D011788, tree number F01.829.458)	

Morphosyntactic processing

- Initial aim: test the suitability of complex linguistic processing (dependence parsing, NER, ...)
- Final experiments employed only simple NLP tools
  - dealing with morphosyntactic variation using a lemmatizer to identify lexical roots
  - stop-word removal replaced by a content-word selection based on part-of-speech (PoS) tags
- Linguistic analysis using ClearNLP toolset (`http://clearnlp.wikispaces.com/`)
  - provides dictionaries and trained PoS tagging models built from medical corpora
  - PoS tags to filter content-words: *Noum*, *Verb*, *Adjective*, *Unknown word*.

Bigram association measures

- Use of word bigrams gave us some performance improvement in our first BioASQ participation
  - but it increases memory and time requirements both in training and labeling
- **Basic premise:** in complex domains multiword terms usually provide higher discriminative power than the simple terms that constitute them
- **Our approach:** identify approximately multiword terms without deep natural language analysis
  - compute association measures on bigrams taken from filtered and normalized tokens
  - Mutual Information using Ngram Statistics Package (`http://ngram.sourceforge.net/`)
  - top ranked bigrams could be considered as multiword collocations
- Some examples of top bigrams:

`escherichia coli`
`amino acids`
`gamma globulin`
`guinea pig`

`electron microscopy`
`cell wall`
`staphylococcus aureus`
`deoxyribonucleic acid`

`molecular weight`
`electron microscope`
`amino acid`
`ribonucleic acid`

`f`
`e`
`b`
`p`

Combining ensembles of classification models

- **Basic idea:** mixing label proposals coming from an ”expert committee” of models could improve the overall quality of the individual predictions
- Tested strategies:
  1. build a model for each MeSH subhierarchy
  2. build several models ensuring maximum diversity

Model per subhierarchy approach (I)

- MeSH thesaurus can be arranged into 16 overlaping subhierarchies according to the *TreeNumber* list attached to every MeSH descriptor
    - replicating common subtrees those subhierarchies can be assumed to be independent

subhierarchy	# of desc.
(A) Anatomy	2,927
(B) Organisms	5,196
(C) Diseases	11,303
(D) Chemicals and Drugs	20,992
(E) Analytical, Diagnostic and Therapeutic Techniques and Equipment	4,764
(F) Psychiatry and Psychology	1,150
(G) Biological Sciences	3,428
(H) Physical Sciences	513

subhierarchy	# of desc.
(I) Anthropology, Education, Sociology and Social Phenomena	651
(J) Technology and Food and Beverages	601
(K) Humanities	218
(L) Information Science	519
(M) Persons	258
(N) Health Care	2,350
(V) Publication Characteristics	188
(Z) Geographic Locations	553

- The categorization task in the BioASQ challenge can be splitted across a set of partial models specialized in each one of the MeSH subhierarchies

Model per subhierarchy approach (II)

- **Critical points:** when labeling a given document ...
  1. ... which partial models should be checked?
  2. ... how many descriptor labels should be taken from each submodel prediction?
  3. ... how those predicted labels should be ranked?
- **Our approach:** $k-$nn based voting scheme
  - a Lucene index is queried to select the top k similar documents
  - candidate subhierarchies are identified from label assignment in those documents
  - doc. similarity scores are averaged to assign a weight to each subhierarchy
  - voting scores determine: {
    - submodels to be checked
    - number of top predicted descriptors to be retrieved
    - weights to be employed in final ranking
- Only REBAYCT can take advantage of this method
  - HACE implicitly performs an equivalent decision during first stages of document labeling.

Introducing diversity into the trained models

- REBAYCT limitations led to employ a bagging based ensemble in first BioASQ challenge
  - small performance improvement due to combination of several REBAYCT models
  - inspection of partial results showed a high coincidence level among top predicted labels
- **Our objective:** ensure more diversity into the ensemble submodels
- Iterative submodel generation (boosting inspired)
  - starts with a REBAYCT model built only from thesaurus structure (with no training documents)
  - training docs. where current model performance is poor (threshold on F-meaure) are retained as ”interesting” cases to train next model
  - training docs. where performance is acceptable are not taken into account to train next model
  - once enough ”interesting” cases are available, current REBAYCT model is updated using them
  - test performance
    - new model better than current model → continue
    - new model worse than current model → retain current model + restart iteration

BioASQ Task 2A participation

Submitted configurations

- HACE1: HACE framework with
  - *k*-means bottom-up positive example selection (up to 5000 documents per node)
  - IG as local feature selection (up to 500 top ranked features)
  - SVM as local content based classifier
  - guided top-down search approach
  - bigrams from the top 1 % of the ranked list of word bigrams
- HACE2: same configuration as **hace1** without using the guided top-down search approach
- HACE2-NE: same configuration as **hace1** using the top 5 % of the ranked list of word bigrams.
- REBAYCT: 15 REBAYCT models using *model per subhierachy* approach
  - bigrams from the top 1 % of the ranked list of word bigrams
- REBAYCT2: 20 REBAYCT models using *iterative submodel generation* approach
  - bigrams from the top 1 % of the ranked list of word bigrams

Conclusions

- Performance of our systems in BioASQ official runs was disappointing
  - average performance fell into the bottom third of the ranked list of participating systems
  - small performance improvements in results obtained by the HACE framework
  - relative great improvements in REBAYCT runs (mainly due to ensemble based strategies) but still not competitive
- Lemmatization and content-word filtering based on PoS produced minor improvements over stemming and stop-word removal
- Association measures on word bigrams were apparently able to catch useful multiword terms in the biomedical domain

Conclusions

- Both ensemble based approach lead to performance improvements, but more work and tuning need to be done
- Our question from previous BioASQ conclusions remains unanswered ...

  Does 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
  - tests with simple $k-$nn over our Lucene index were not so bad