SMBM'05, Hinxton, 10th-13th April 2005

Semantic Mining in BioMedicine 2005

02/05/2005

Title: Induced extension of gene ontology from biomedical resources with flexible identification of candidate terms (Jin-Bok Lee, Jung-Bae Kim, Jong C. Park)

- Method to predict more detailed terms than those in GO
- They identify meta-level rules among the components of GO terms
- eg: When analysing GO terms that belong to *chemokine* binding, they use meta-level rules to the effect that *C-C chemokine*, *C-X-C chemokine* and *C-X3-C chemokine* all belong to the *chemokine* family.

[Induced extension of gene ontology from biomedical resources with flexible identification of candidate terms]

- After predicting new terms, the system validates them with information collected from biomedical literature
  - system parses relevant sentences to identify syntactic dependencies among components of a given term
  - system provides validity checking for each newly introduced term

[Induced extension of gene ontology from biomedical resources with flexible identification of candidate terms]

- developed a toolset for
  - manipulating extended GO terms
  - searching newly introduced GO terms through PubMed
  - annotating relevant terms in retrieved articles
- the toolset can be used to
  - construct an annotated corpus
  - to extend a given ontology

Title: Combining light-weight retrieval strategies for robust text categorization (Patrick Ruch)

- Text categorization system designed to automatically assign biomedical categories to any input text
- largely data independent, unlike usual text categorization systems
- just needs a small set of instances for tuning the system
- produces a ranked list of categories which can be interactively filtered by user

[Combining light-weight retrieval strategies for robust text categorization]

- METHODS:
  - To evaluate robustness ⇒ test system on 2 different biomedical terminologies
  - combines
    - pattern matcher
    - vector space retrieval engine
  - uses
    - stems
    - linguistically-motivated indexing units

[Combining light-weight retrieval strategies for robust text categorization]

- RESULTS:
  - shows effectiveness of phrase indexing for both GO and MeSH categories
  - categorization power of tool depends on controlled vocabulary
  - precision at high ranks ranges from 90% for MeSH to less than 20% for GO
  - synonyms are useful for GO categorization, not for MeSH
- CONCLUSION:
  - effective categorization strategy when training data are not available,
  - but effectiveness is directly related to the controlled vocabulary.

Title: Discovering paradigm shift patterns in biomedical abstracts: Application to Neurodegenerative Diseases (Frederique Lisacek, Christine Chichester, Aaron Kaplan, Agnes Sandor)

- They propose a methodology for rapidly arriving at a synthetic view of the literature on a particular research reports that contain key elements in the field
- Instead of processing factual info in abstracts, they search for linguistic cues that indicate actual or potential breakthroughs
- they hypothethise that relevant papers are the ones that refer to contradictions, inconsistencies, etc.

[Discovering paradigm shift patterns in biomedical abstracts: Application to Neurodegenerative Diseases]

- They consider that the complex concept of paradigm shift is expressed via the composition of constituents notions:
- for example: "Further[TIME] research[IDEA] on the zinc paradox[CONTRAST] in AD is needed ..."
- They use a total of 7 constituent notions + rules indicating which of the possible combinations of these notions indicates a paradigm shift.

[Discovering paradigm shift patterns in biomedical abstracts: Application to Neurodegenerative Diseases]

- With the 7 constituent notions are associated a total of some 600 words, which were selected by hand with some automatic support for proposing candidates.
- Final result: list of abstracts that contain paradigm shift expressions, ranked in order of the concentration of topic keywords in the abstract
- $+$ to help increase the accuracy of the ranking: additional keywords based on frequency of co-occurence with the initial set
- Test case for evaluation: Neurodegenerative Diseases - high precision.

Title: Large-scale extraction of protein/gene relations for model organisms (Jasmin Saric, Lars Juhl Jensen, Rossita Ouzounova, Isabel Rojas, Peer Bork)

- Previously developed a rule-based approach for extracting information on the regulation of gene expression in yeast ⇒ now expanding system to also extract info on other equally important regulatory mechanisms (eg phosphorylation)
- new results for extraction of relational information from biomedical text:
- improved system to both capture new types of linguistic constructions and biological information
- precision stable with a slight increase in recall

[Large-scale extraction of protein/gene relations for model organisms]

- for almost 1 million PubMed abstracts related to 4 model organisms, they manage to extract regulatory networks and binary phosphorylations comprising 3319 relation chunks.

- Accuracy:
  - 83-90% for gene expression relations,
  - 86-95% for (de-)phosphorylation relations.

[Large-scale extraction of protein/gene relations for model organisms]

- To achieve this, they made use of an organism-specific resource of gene/protein names larger than usually used
- These names were included in the lexicon when retraining the POS tagger on the GENIA corpus.
- For this domain, accuracy of 96.4% attained on POS tags.
- Rules were developed for yeast and successfully applied to both abstracts and full-text related to other organisms with comparable accuracy.

Title: Biomedical information extraction with predicate-argument structure patterns (Akane Yakushiji, Yusuke Miyao, Yuka Tateisi, Jun'ichi Tsuji)

- They propose a new method for automatic construction of application-specific extraction rules, which uses PASs produced by a full parser
- By dividing labor between generic linguistic rules in the parser and application-specific extraction rules to be constructed from scratch ⇒ Method facilitates the acquisition of extraction rules from a relatively small annotated corpus
- Experiment: extraction of protein-protein interaction ⇒ Results show that performance is promising and comparable with those obtained by manual-mode extraction rules or those obtained by rules generalised by ML techniques

[Biomedical information extraction with predicate-argument structure patterns]

- methods: text annotated with desired info ⇒ pattern constructor ⇒ PAS patterns

- Construction of an extraction rule in 3 steps:
  - Pattern extraction by full parsing
  - Pattern division into components
  - Pattern filtering

[Biomedical information extraction with predicate-argument structure patterns]

- to extract new protein-protein interaction, they match obtained patterns to parsing results of sentences in new input text
- pattern matching is done by PAS matching
- results on extraction of protein-protein interaction on 199 Medline abstracts: 37.3% precision and 45.3% recall without any manual intervention