

# Syntactic-semantic analysis for information extraction in biomedicine

Sérgio Matos<sup>1</sup>, Anabela Barreiro<sup>2</sup>

<sup>1</sup>IEETA, Universidade de Aveiro

<sup>2</sup>Centro de Linguística, Universidade do Porto

[aleixomatos@ua.pt](mailto:aleixomatos@ua.pt); [barreiro\\_anabela@hotmail.com](mailto:barreiro_anabela@hotmail.com)

June 2009



# Outline

- Background
- Text Mining and Information Extraction in Biomedicine
- Objectives
- Implementation
- Results
- Conclusions

# Background

- Genomics and Proteomics are fast-growing fields
- Literature grows exponentially
  - MEDLINE/PubMed ~ 18m citations
- Researchers need to contextualize their theories and findings
  - Interactions between genes/proteins
  - Involvement in biological processes and in disease
  - And many other factors...
  - How to keep up-to-date with new knowledge in the field?

# Background

- Manually curated biomedical databases are a good source of information
  - Publications are reviewed and important information added to DBs (e.g. protein interactions)
  - Impossible to keep DBs up-to-date due to increased volume of publications
- Text Mining can be useful for
  - Information retrieval (IR)
  - Information extraction (IE)
  - DB curators and end-users (researchers)

# Text Mining and Information Extraction in Biomedicine

- Text mining deals with the automated processing of texts to derive high quality information
- Information Extraction can be seen as one application of TM
- Different processing levels



- |                           |  |
|---------------------------|--|
| • Entity Recognition (ER) | genes, proteins, etc.                        |
| • Normalization           | ATF2 - GeneID 1386<br>ATF-2 – Uniprot P15336 |
| • Relation extraction     | PPI, gene/disease                            |
| • Event extraction        | gene expression, regulation                  |

+ semantics + domain knowledge

# Text Mining and Information Extraction in Biomedicine

- Good results for NER, but limited to a few entity types
  - 80%-90% for recognition of genes/proteins
  - Need to include more entities, like chemical compounds, diseases, experimental conditions
- Relation extraction has focused mostly on PPI
- Inter-concept relations not too explored
  - e.g. gene/disease, drug/target
  - mostly based on co-occurrence statistics

# Text Mining and Information Extraction in Biomedicine

- Recent interest towards extraction of events
  - BioNLP shared task and BioCreaTive II.5
- ... and other entities / facts
  - e.g. Experimental conditions, lab techniques, measurements
- ... Discourse analysis
  - “indicating/suggesting that...”, “in contrast...”
- Full-text vs. Abstracts
  - Complexity in grammar

# Linguistic Resources for Biomedical TM

- UMLS Metathesaurus
  - various terms, all linked to same concept (e.g. ‘Hypertension’)
  - semantic information provided by the UMLS Semantic Network
- BioLexicon
  - Includes domain relevant verbs (localize, bind, express, ...)
- Lexical resources can be created from available online DBs
  - NCBI Entrez Gene for gene names
  - UniProt for proteins
  - OMIM for diseases
  - Various ontologies

# Objectives

- Extract phrases indicating a biomolecular event from scientific text
- Biomolecular events include various types
  - Examples
    - “phosphorylation of TRAF2”
    - “localization of beta-catenin”
    - “TRADD interacts with TES2”
- BioNLP'09 Shared Task on Event Extraction
  - <http://www-tsuji.is.s.u-tokyo.ac.jp/GENIA/SharedTask/>

# Objectives

- Six event types considered
  - Localization, Binding, Gene expression, Transcription, Protein Catabolism, Phosphorylation
- Training data
  - Annotation of genes/proteins occurring in each input text, including the text section (start and end characters)
  - Annotation of the events, including the event type, the participating entities and the corresponding trigger word (with start and end times)
- Test data
  - Annotation of participating genes/proteins is given
  - Create annotation of events for the given entities

# Implementation

- General approach
  - Create syntactic grammars to detect phrases that indicate events
  - Grammars are based only on NEs and domain verbs (and derived names)
- Requisites
  - Grammars outputs should indicate the event type
- Solution
  - Event types can be associated with the trigger word using the semantic properties in NooJ dictionaries
  - Event types associated with each trigger word are derived from training data

# Implementation

- Resources
  - Entity dictionary
    - Create dictionary with list of entities occurring in the texts

# Implementation

Lemma	PoS	FLX	Semantic properties	ID	TAXID
human	N	TABLE	ORGANISM		9606
Homo sapiens	N		ORGANISM		9606
Mus musculus	N		ORGANISM		10090
Breast cancer type 1 susceptibility protein	N		PROTEIN	P3839 8	9606
BRCA1	N		PROTEIN	P3839 8	9606
BRCA1	N		PROTEIN	P4875 4	10090
BRCA1	N		GENE	672	9606
RNF53	N		GENE	672	9606

# Implementation

- Resources
  - Entity dictionary
    - Create dictionary with list of entities occurring in the texts
  - BioLexicon verb dictionary
    - Adapted to include event type
      - From the training data, extract the verbs associated with events
      - Add a semantic property to the dictionary entry indicating the event type
      - Example: “express,V+EventType=Gene\_Expression”
    - Added inflectional and derivation rules
      - The inflected and derivated forms inherit the verb’s semantic properties

# Implementation

- Verb dictionary

---

Lemma	PoS	DRV	FLX	EventType
express	V	ION:TABLE	ABOLISH	Gene_expression
ligate	V	TION:TABLE	SMILE	Binding
stimulate	V	TION:TABLE	SMILE	Positive_regulation

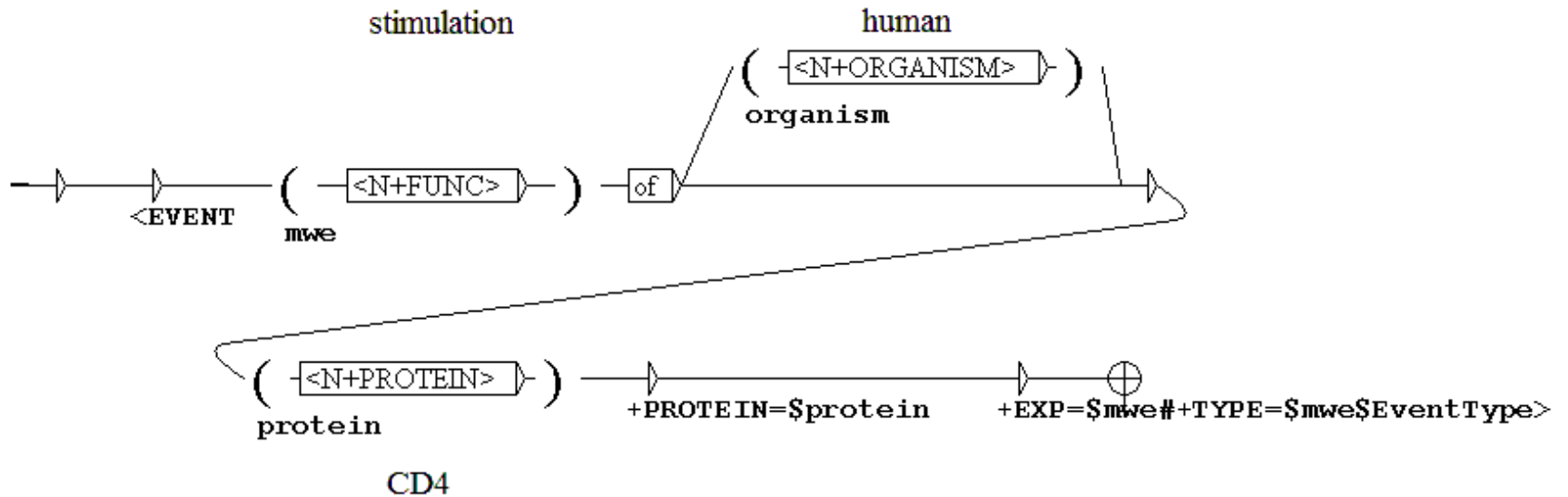
---

# Implementation

- Syntactic grammars
  - Sentences from training set used to generate surface patterns
  - Manual procedure
  - Seven grammars created
  - Example:

“stimulation of human CD4”

# Implementation



Stimulation of human CD4

<EVENT+PROTEIN=CD4+EXP=Stimulation+TYPE=Positive\_regulation>

# Results

- Example patterns extracted from texts

---

Pattern	Concordance in text
<entity> [<entity_type>] <nominalization>	<i>HSP gene expression</i>
<nominalization> “of” [<entity_type>] <entity>	<i>upregulation of Fas</i>
<entity> [<entity_type>] <be> [“not”] [<adverb>] <verb>	<i>IL-2R stimulation was totally inhibited</i>
<verb> <preposition> <entity>	<i>binding of TRAF2</i>
<verb> <nominalization> “of” <entity>	<i>suppressing activation of STAT6</i>

---

# Results

- Average results

<b>Event type</b>	<b>Recall</b>	<b>Precision</b>	<b>F-score</b>
Localization	35.63	70.45	47.33
Binding	13.54	34.06	19.38
Gene Expression	46.40	78.45	58.31
Transcription	33.58	41.07	36.95
Protein Catabolism	35.71	62.50	45.45
Phosphorylation	49.63	79.76	61.19
<b>Average</b>	<b>36.76</b>	<b>65.58</b>	<b>47.11</b>

# Conclusions

- NooJ syntactic grammars for IE
  - Simple and flexible approach
  - Takes advantage of semantic properties and inflectional and derivational morphology in NooJ dictionaries
- Pattern identification
  - Manual method is limited
  - How to generate new patterns automatically ?
- Gene regulatory events
  - Described by complex constructions
  - Can syntactic grammars be used for this type of events ?

# References and Acknowledgments

- BioLexicon was developed within the BOOTStrep project
  - <http://www.nactem.ac.uk/biolexicon/>
  - <http://www.bootstrep.eu/bin/view/Extern/WebHome>
- Data set from the BioNLP'09 Shared Task on Event Extraction
  - <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/SharedTask/>

Sérgio Matos is funded by Fundação para a Ciência e Tecnologia (FCT) under the Ciência2007 programme .