Semantic Retrieval Tools

Construction, Types, and Uses in BioNLP

Ekta Sood[1]

[1] University of Stuttgart
Institute for Natural Language Processing

29 January 2018

Outline

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information
  - use case specific

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information
  - use case specific
  - genes, diseases, etc

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information
  - use case specific
  - genes, diseases, etc
- Enhancing search and access to relevant info = important task for NLP researchers

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information
  - use case specific
  - genes, diseases, etc
- Enhancing search and access to relevant info = important task for NLP researchers
  - large size of unstructured & ambiguous literature and rapid growth

The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information
  - use case specific
  - genes, diseases, etc
- Enhancing search and access to relevant info = important task for NLP researchers
  - large size of unstructured & ambiguous literature and rapid growth
- IR Tools : event extraction, entity-relation extraction, PPI, gene mutations, etc

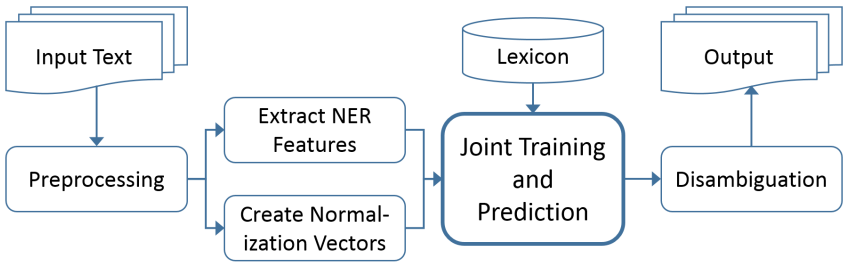
The Needs

BioNLP : utilizes NLP methods to extract info from scientific literature to find knowledge relevant to research

- Biomedical scientists who want access to (text repositories) and retrieve meaningful information
  - use case specific
  - genes, diseases, etc
- Enhancing search and access to relevant info = important task for NLP researchers
  - large size of unstructured & ambiguous literature and rapid growth
- IR Tools : event extraction, entity-relation extraction, PPI, gene mutations, etc
  - use cases vary depending on the scientists position

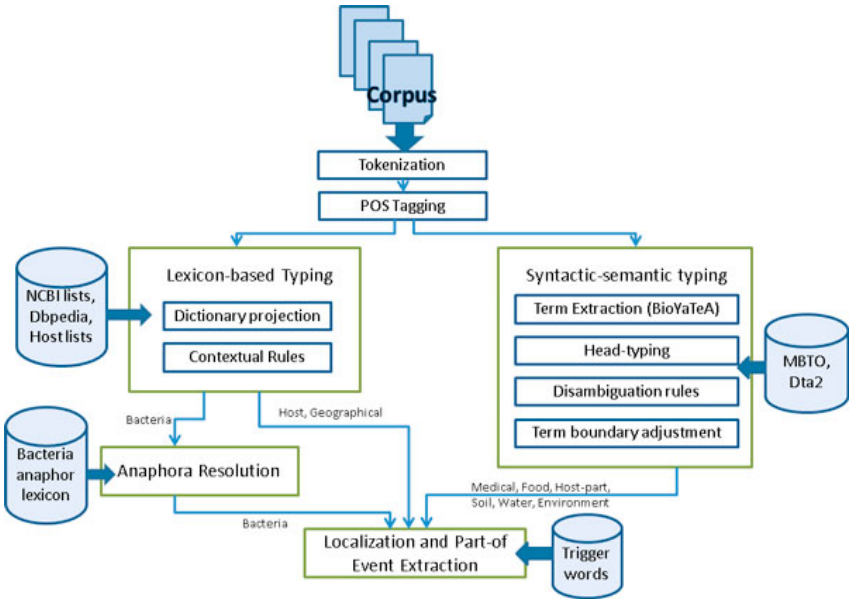
Workflow of Use Case : Named Entity Recognition

FIGURE – TaggerOne : Joint Named Entity Recognition and Normalization



Workflow of Use Case : Event Extraction

FIGURE – Bacteria Biotypes : Event Extraction



Motivation

Curate the vast amount of data : build upon existing databases to increase search specificity and decrease ambiguity

Motivation

Curate the vast amount of data : build upon existing databases to increase search specificity and decrease ambiguity

Problem 1 : Large amount of data

- Large : rapid growth of published texts
- Expensive to manually search for specific info

Motivation

Curate the vast amount of data : build upon existing databases to increase search specificity and decrease ambiguity

Problem 1 : Large amount of data

- Large : rapid growth of published texts
- Expensive to manually search for specific info

Problem 2 : Ambiguity

- Ambiguous : names of biological entities (genes, chemicals, etc)
- Outputs large amounts of unspecific texts

Motivation

Curate the vast amount of data : build upon existing databases to increase search specificity and decrease ambiguity

Problem 1 : Large amount of data

- Large : rapid growth of published texts
- Expensive to manually search for specific info

Problem 2 : Ambiguity

- Ambiguous : names of biological entities (genes, chemicals, etc)
- Outputs large amounts of unspecific texts

Problem 3 : Unstructured

- Data is often unstructured

Motivation

Curate the vast amount of data : build upon existing databases to increase search specificity and decrease ambiguity

Problem 1 : Large amount of data

- Large : rapid growth of published texts
- Expensive to manually search for specific info

Problem 2 : Ambiguity

- Ambiguous : names of biological entities (genes, chemicals, etc)
- Outputs large amounts of unspecific texts

Problem 3 : Unstructured

- Data is often unstructured

Abundance

Unspecific

Unstructured

Retrieval Tools : Pros and Cons

PubMed – Medline

- Pros : largest, popular, 15M abstracts
- Cons : finding relevant literature = difficult
- Use case : interface –type in a query and get a list of results from Medline

Retrieval Tools : Pros and Cons

PubMed – Medline

- Pros : largest, popular, 15M abstracts
- Cons : finding relevant literature = difficult
- Use case : interface –type in a query and get a list of results from Medline

iHop – PubMed Sentences

- Pros : access to subset of sentences – proteins connected to specific terms
- Con : non-protein entities not considered
- Use case : protein mentions

Retrieval Tools : Pros and Cons

PubMed – Medline

- Pros : largest, popular, 15M abstracts
- Cons : finding relevant literature = difficult
- Use case : interface –type in a query and get a list of results from Medline

iHop – PubMed Sentences

- Pros : access to subset of sentences – proteins connected to specific terms
- Con : non-protein entities not considered
- Use case : protein mentions

Alibaba – PubMed Queries

- Pro : aggregates info from results & visualizes with graphs
- Con : focus is not on document level & suboptimal results due to NER and RE methods
- Use case : combines results from documents which describe a similar topic and generates graphs extracted from texts

Retrieval Tools : Pros and Cons

EbiMed – PubMed Queries

- Pros : retrieves co-occurring entities & ranks by freq
- Cons : does not focus on documents
- Use case : combines results from documents which describe a similar topic

Retrieval Tools : Pros and Cons

EbiMed – PubMed Queries

- Pros : retrieves co-occurring entities & ranks by freq
- Cons : does not focus on documents
- Use case : combines results from documents which describe a similar topic

UKPMC – PubMed Central – uses Whatiszit ...

- Pros : using PM Central to highlight entities in abstracts
- Con : functionality isn't for full texts & searches for entity names instead of identifiers
- Use case : recognize and highlight entities in abstracts

Retrieval Tools : Pros and Cons

EbiMed – PubMed Queries

- Pros : retrieves co-occurring entities & ranks by freq
- Cons : does not focus on documents
- Use case : combines results from documents which describe a similar topic

UKPMC – PubMed Central – uses Whatiszit ...

- Pros : using PM Central to highlight entities in abstracts
- Con : functionality isn't for full texts & searches for entity names instead of identifiers
- Use case : recognize and highlight entities in abstracts

SCAIView – Advanced semantic search engine

- Pro : info relevant to general researchers & can deal with complex queries (SOA tools)
- Con : uses dictionary based NER
- Use case : Retrieves entities in full texts

Retrieval Tools : Pros and Cons

XplorMed – PubMed with MeSH

- Pros : maps PM results to 8 MeSh categories & extracts keywords & co-occurrences
- Cons : too general & misses important info
- Use case : better search for general literature

Retrieval Tools : Pros and Cons

XplorMed – PubMed with MeSH

- Pros : maps PM results to 8 MeSh categories & extracts keywords & co-occurrences
- Cons : too general & misses important info
- Use case : better search for general literature

Vivisimo – Pubmed with Ontology

- Pros : automatically derives ontology from search results
- Con : Ontologies are not well structured (not hand-curated)
- CHANGE THIS Use case : dictionaries, built on ontologies, facilitate integration of results

Retrieval Tools : Pros and Cons

XplorMed – PubMed with MeSH

- Pros : maps PM results to 8 MeSh categories & extracts keywords & co-occurrences
- Cons : too general & misses important info
- Use case : better search for general literature

Vivisimo – Pubmed with Ontology

- Pros : automatically derives ontology from search results
- Con : Ontologies are not well structured (not hand-curated)
- CHANGE THIS Use case : dictionaries, built on ontologies, facilitate integration of results

Textpresso – Pubmed with Ontology

- Pros : 30 categories for classification
- Con : only 30 categories for classification & linguistic limitations
- Use case : results show subsections of papers relevant to query

Retrieval Tools : Pros and Cons

GeneView – Pubmed and Pubmed Central Articles

- Pros : semantically enrich articles, user access, inter-operational components
- Cons : difficult to maintain
- Use case : researcher who often use gene mapping data, but are not tech savvy

Retrieval Tools : Pros and Cons

GeneView – Pubmed and Pubmed Central Articles

- Pros : semantically enrich articles, user access, inter-operational components
- Cons : difficult to maintain
- Use case : researcher who often use gene mapping data, but are not tech savvy

GoPubMed–Pubmed with GO

- Pros : improves search by exploiting Gene Ontology (structure of GO is key)
- Con : limited to 100 papers
- Use case :Enhanced search efficiency

Retrieval Tools : Pros and Cons

GeneView – Pubmed and Pubmed Central Articles

- Pros : semantically enrich articles, user access, inter-operational components
- Cons : difficult to maintain
- Use case : researcher who often use gene mapping data, but are not tech savvy

GoPubMed–Pubmed with GO

- Pros : improves search by exploiting Gene Ontology (structure of GO is key)
- Con : limited to 100 papers
- Use case :Enhanced search efficiency

Let's Zoom In !

- For the next few slides, we will review the GeneView system in detail

GeneView - Semantic Search Engine

Architecture of GV

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific
- Specific annotations on PM abstracts and open full texts

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific
- Specific annotations on PM abstracts and open full texts
  - generated by various algorithms

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific
- Specific annotations on PM abstracts and open full texts
  - generated by various algorithms
- Entity and relation types

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific
- Specific annotations on PM abstracts and open full texts
  - generated by various algorithms
- Entity and relation types
  - identifies and normalizes

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific
- Specific annotations on PM abstracts and open full texts
  - generated by various algorithms
- Entity and relation types
  - identifies and normalizes
- Indexes abstracts and full texts

GeneView - Semantic Search Engine

Architecture of GV

- Bundles best available algs
  - processing of text = use specific
- Specific annotations on PM abstracts and open full texts
  - generated by various algorithms
- Entity and relation types
  - identifies and normalizes
- Indexes abstracts and full texts
- Generates structured info

Graphical Depiction of GV Architecture

GeneView - Semantic Search Engine

NER and RE

GeneView - Semantic Search Engine

NER and RE

- recognizes entity mentions from 10 classes

GeneView - Semantic Search Engine

NER and RE

- recognizes entity mentions from 10 classes
  - chemicals, cell types, diseases, drugs, enzymes, genes, histone modifications, SNPs, species, and tissues

GeneView - Semantic Search Engine

NER and RE

- recognizes entity mentions from 10 classes
  - chemicals, cell types, diseases, drugs, enzymes, genes, histone modifications, SNPs, species, and tissues
- recognizes 3 relationship types

GeneView - Semantic Search Engine

NER and RE

- recognizes entity mentions from 10 classes
  - chemicals, cell types, diseases, drugs, enzymes, genes, histone modifications, SNPs, species, and tissues
- recognizes 3 relationship types
  - PPIs, Regulatory, and DDIs

GeneView - Semantic Search Engine

NER and RE

- recognizes entity mentions from 10 classes
  - chemicals, cell types, diseases, drugs, enzymes, genes, histone modifications, SNPs, species, and tissues
- recognizes 3 relationship types
  - PPIs, Regulatory, and DDIs
- 210M entity mentions (since 2013)

GeneView - Semantic Search Engine

NER and RE

- recognizes entity mentions from 10 classes
  - chemicals, cell types, diseases, drugs, enzymes, genes, histone modifications, SNPs, species, and tissues
- recognizes 3 relationship types
  - PPIs, Regulatory, and DDIs
- 210M entity mentions (since 2013)
- 8.3M relationships (since 2013)

GeneView - Semantic Search Engine

System Description - IE Workflow

GeneView - Semantic Search Engine

System Description - IE Workflow

- Pre-process texts

GeneView - Semantic Search Engine

System Description - IE Workflow

- Pre-process texts
- NER Tools

GeneView - Semantic Search Engine

System Description - IE Workflow

- Pre-process texts
- NER Tools
  - Tools use : CRFs and dictionaries

GeneView - Semantic Search Engine

System Description - IE Workflow

- Pre-process texts
- NER Tools
  - Tools use : CRFs and dictionaries
  - Tools : GNAT, MutationFinder, ChemSpot, LINNAES, Brno histone, other NE (cell types, diseases, etc)

GeneView - Semantic Search Engine

System Description - IE Workflow

- Pre-process texts
- NER Tools
  - Tools use : CRFs and dictionaries
  - Tools : GNAT, MutationFinder, ChemSpot, LINNAES, Brno histone, other NE (cell types, diseases, etc)
- RE Tools

GeneView - Semantic Search Engine

System Description - IE Workflow

- Pre-process texts
- NER Tools
  - Tools use : CRFs and dictionaries
  - Tools : GNAT, MutationFinder, ChemSpot, LINNAES, Brno histone, other NE (cell types, diseases, etc)
- RE Tools
  - SVM

Information Extraction Workflow

GeneView - Semantic Search Engine

System Description - 3 Machines

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database
  - stores annotations and metadata

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database
  - stores annotations and metadata

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database
  - stores annotations and metadata
  - specific information of entities : type, ID, position in text

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database
  - stores annotations and metadata
  - specific information of entities : type, ID, position in text
  - meta-data : authors, MeSH terms, figures

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database
  - stores annotations and metadata
  - specific information of entities : type, ID, position in text
  - meta-data : authors, MeSH terms, figures
- Web-based Interface : shows the stored information to the user

GeneView - Semantic Search Engine

System Description - 3 Machines

- Lucence
  - text storage
  - query processing
  - ranking engine
- Relations database
  - stores annotations and metadata
  - specific information of entities : type, ID, position in text
  - meta-data : authors, MeSH terms, figures
- Web-based Interface : shows the stored information to the user
  - web-server communicates queries with Lucence

GeneView - Semantic Search Engine

Features

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems
  - general filtering

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems
  - general filtering
    - document relevance

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems
  - general filtering
    - document relevance
  - general ranking

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems
  - general filtering
    - document relevance
  - general ranking
    - publication date, PMID, etc

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems
  - general filtering
    - document relevance
  - general ranking
    - publication date, PMID, etc
  - ranking by entity counts

GeneView - Semantic Search Engine

Features

- Number of features not available with other systems
  - general filtering
    - document relevance
  - general ranking
    - publication date, PMID, etc
  - ranking by entity counts
  - personalized ranking (users define their own gene lists)

GeneView - Semantic Search Engine

Systematic Statistics

GeneView - Semantic Search Engine

Systematic Statistics

- 20M abstracts

GeneView - Semantic Search Engine

Systematic Statistics

- 20M abstracts
- 271K full texts

GeneView - Semantic Search Engine

Systematic Statistics

- 20M abstracts
- 271K full texts
- 194 M entities

GeneView - Semantic Search Engine

Systematic Statistics

- 20M abstracts
- 271K full texts
- 194 M entities
- 3921267 classified PPIs

GeneView - Semantic Search Engine

Systematic Statistics

- 20M abstracts
- 271K full texts
- 194 M entities
- 3921267 classified PPIs
  - 15197637 total co-occurring protein mutations

GeneView – Summary

- **GeneView** :

GeneView – Summary

- **GeneView** :
  - uses IE tools to automatically annotate :

GeneView – Summary

- **GeneView** :
  - uses IE tools to automatically annotate :
    - 10 NER tools for entity mentions and 3 relation extraction methods

GeneView – Summary

- **GeneView** :
  - uses IE tools to automatically annotate :
    - 10 NER tools for entity mentions and 3 relation extraction methods
  - Indexing for streamlining search

GeneView – Summary

- **GeneView** :
  - uses IE tools to automatically annotate :
    - 10 NER tools for entity mentions and 3 relation extraction methods
  - Indexing for streamlining search
  - FIX ranks and filters
  - Pros and Cons

GeneView – Summary

- **GeneView** :
  - uses IE tools to automatically annotate :
    - 10 NER tools for entity mentions and 3 relation extraction methods
  - Indexing for streamlining search
  - FIX ranks and filters
  - Pros and Cons
    - **solves problem of ambiguity as it has in depth entity search options**

GeneView – Summary

- **GeneView** :
  - uses IE tools to automatically annotate :
    - 10 NER tools for entity mentions and 3 relation extraction methods
  - Indexing for streamlining search
  - FIX ranks and filters
  - Pros and Cons
    - **solves problem of ambiguity as it has in depth entity search options**
    - **expensive to update ..120 days**

To Summarize

- **Bio Medical Web Based Tools** :

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth
  - unstructured texts

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth
  - unstructured texts
  - ambiguity in entity mentions and relations

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth
  - unstructured texts
  - ambiguity in entity mentions and relations
- **Tools utilize information extraction methods** :

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth
  - unstructured texts
  - ambiguity in entity mentions and relations
- **Tools utilize information extraction methods** :
  - NER

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth
  - unstructured texts
  - ambiguity in entity mentions and relations
- **Tools utilize information extraction methods** :
  - NER
  - term extraction

To Summarize

- **Bio Medical Web Based Tools** :
  - need to curate vast amt of literature
  - used across many domains
  - aids to limit manual work
  - stream lines information gain in order to optimize time of the user
- **Creating beneficial tools is difficult** :
  - rapid literature growth
  - unstructured texts
  - ambiguity in entity mentions and relations
- **Tools utilize information extraction methods** :
  - NER
  - term extraction
  - gene ontology bases

References