Natural Language Processing in the Medical and Biological Domains : a Parallel Perspective

Pierre Zweigenbaum

LIMSI, CNRS, Orsay, France

ERTIM, INALCO, Paris, France

SMBM 2008, September 3

Natural Language Processing in the Medical and Biological Domains : *Why are they different ?*

Pierre Zweigenbaum

LIMSI, CNRS, Orsay, France

ERTIM, INALCO, Paris, France

SMBM 2008, September 3

Introduction

Natural language processing in two related domains

- Are they more or less the same ?
- Do the observed differences come from the two domain topics ?
- Must other origins be called for (tasks, etc.) ?

Natural Language Processing in the *Medical* Domain

Natural Language Non-artificial language used by humans

Natural Language Processing (NLP) Computer processing (analysis, generation, etc.) of natural language utterances

Medical Domain That of *Medical Informatics*:

- Health care (treat patients)
- Associated information and knowledge management (medico-economic goals, best practice)
- Acquiring new knowledge (medical research)

Natural Language Processing in the *Biomedical* Domain

Natural Language Processing As in the medical domain

Note : Text Mining

- Data mining from text [Hearst, 2003]
- Beyond simple information extraction : synthesis, hidden links, new knowledge
- Generally relies on Natural Language Processing

Biomedical Domain That of *Biomedical Informatics* :

- Molecular Biology
- Genomics
- *omics

1. Observations
    - MEDLINE View
    - Outline of Medical NLP Work
    - Outline of Biomedical NLP Work
2. Two Domains
    - Clinical Sublanguage
    - Biomolecular Sublanguage
3. User Needs and Tasks
    - Medical NLP : Diverse User Needs
    - Biomedical NLP : More Focused User Needs
4. Impact of Text Genres
    - Open Access
    - Language
5. Attractivity Factors
6. Wrap-up

Medical and Biomedical NLP

According to MEDLINE

1. “Natural Language Processing” [Main Heading]
   - MeSH record : *91(87) ; was see under ARTIFICIAL INTELLIGENCE 1987-90*
2. Genome biology : union of numerous descriptors
   - Used simple approximating expression of [Demner-Fushman et al., 2007] :

```
("genes"[TIAB] NOT Medline[SB])
 OR "genes"[MeSH Terms] OR gene[Text Word]
 OR "genetics"[Subheading]
```

Approximation of BioNLP : 1 & 2

- Manual check
- Also examine “text mining”[all fields]

Medical and Biomedical NLP in MEDLINE

Discussion

- See also [Rebholz-Schuhman et al., 2007]
  - 1990-1999 *vs* 2000-2005 in Medical Informatics and Biomedical Informatics
  - Frequent bigrams reveal common apparition of *ontology*, *text mining*, *SVM*
  - Here, focus on NLP + much simpler study
- Search biases :
  - MEDLINE does not contain all Bio/Medical/NLP publications
  - Search expressions roughly approximate actual goal

Medical and Biomedical NLP in MEDLINE

Boundaries and quantities

prehistory	NLP[mh]	NLP & <gene>	"text mining"
< 1983	1983–2008.08	1999–2008.08	1999–2008.08
0	1263	265	244

Medical and Biomedical NLP in MEDLINE

Boundaries and quantities

prehistory	NLP[mh]	NLP & <gene>	“text mining”
< 1983	1983–2008.08	1999–2008.08	1999–2008.08
0	1263	265	244

- Earliest NLP in MEDLINE :
  - Chi EC, Sager N, Tick LJ, Lyman MS (1983) Relational data base modelling of free-text medical narrative, *Med Inform (London)*
  - Gabrieli ER, Speth DJ (1986) Automated analysis of the discharge summary, *J Clin Comput*

Medical and Biomedical NLP in MEDLINE

Boundaries and quantities

prehistory	NLP[mh]	NLP & <gene>	“text mining”
< 1983	1983–2008.08	1999–2008.08	1999–2008.08
0	1263	265	244

- Earliest NLP & <gene> in MEDLINE :
  - Rindflesch TC, Hunter L, Aronson AR (1999) Mining molecular binding terminology from biomedical text, *Proc AMIA Symp*
  - Rzhetsky A, Koike T, Kalachikov S et al. (2000) A knowledge model for analysis and simulation of regulatory networks, *Bioinformatics*

Medical and Biomedical NLP in MEDLINE

Boundaries and quantities

prehistory	NLP[mh]	NLP & <gene>	“text mining”
< 1983	1983–2008.08	1999–2008.08	1999–2008.08
0	1263	265	244

- Early Text mining in MEDLINE :
  - Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN (1999) MedMiner : an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques

Medical and Biomedical NLP in MEDLINE

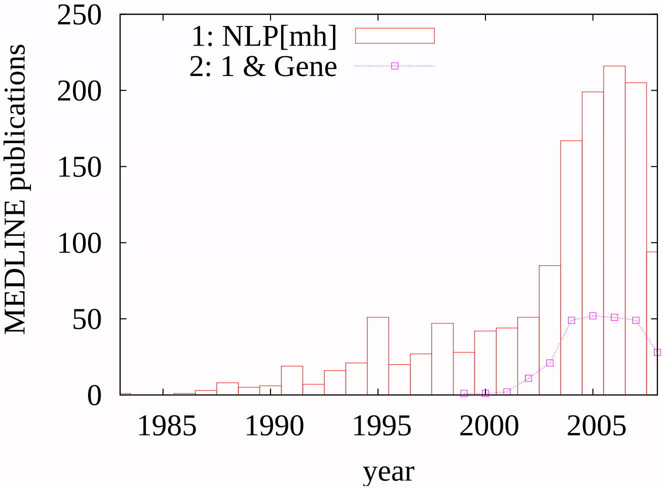
Boundaries and quantities

prehistory	NLP[mh]	NLP & <gene>	“text mining”
< 1983	1983–2008.08	1999–2008.08	1999–2008.08
0	1263	265	244

- Who uses “text mining” ($n = 244$) ?
  - Text mining & NLP = 89/244 (36%) :
    - Text mining does not imply NLP ?
  - Text mining & <gene> = 139/244 (57%)
  - Text mining & biomedical (manual) = 196/244 (80%) :
    - Text mining ⇒ BioNLP

Medical and Biomedical NLP in MEDLINE

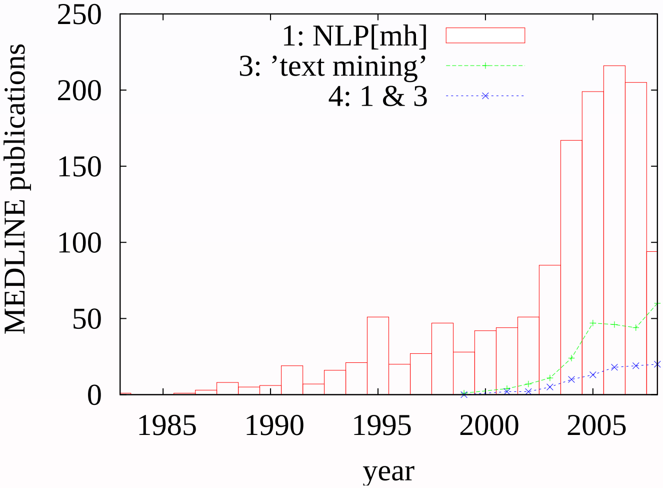
Along the years : NLP[mh] & <gene>



Hand-check : in 2007, 82 papers/185 are biomedical NLP

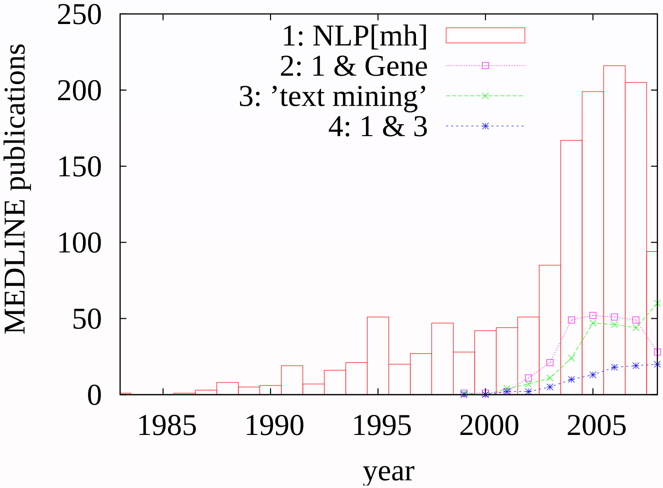
Medical and Biomedical NLP in MEDLINE

Along the years : NLP[mh] & “text mining”



Medical and Biomedical NLP in MEDLINE

Along the years



Outline of Medical NLP Work

A constellation of activities

Medical NLP : Community

IMIA WG 6 1981- Coding, Medical concept representation, NLP

- Triennial workshop (1981-)

AMIA NLP SIG/WG 19 ? ? then 2000- Natural language processing

- Started the first ACL workshop
  "Natural Language Processing in the Biomedical Domain" (2002)
- A growing proportion of sessions at the AMIA Symposium :
  - NLP, Text mining, links to Terminology and Ontology

EFMI WG NLU 2002 ?- Natural language understanding

Medical NLP : Lexical Resources

- UMLS Specialist Lexicon
- UMLS Specialist Lexical tools
- + Scattered resources in various teams

Strong link with terminology development and maintenance (coding)

Medical Ontologies

Principled ontology design ; issues important in medical domain (mereology, non-existence, uncertainty. . .)

- GALEN
- SNOMED CT
- Foundational Model of Anatomy

Role of IMIA WG 6

Medical NLP : Named Entity Recognition

- Mostly viewed through coding and indexing
- More Automated Term Recognition than really NER
    - Includes term/concept normalization
- MetaMap [Aronson, 2001]

Medical NLP : Indexing and IR

- MeSH, MEDLINE : SAPHIRE [Hersh], MTI [Aronson et al., 2000]
- SNOMED, ICD : numerous works
- UMLS : MetaMap
- Concept-based indexing

Medical NLP : Semantic Analysers

- LSP-MLP : clinical, literature [Sager et al., 1987]
- MedLEE : clinical [Friedman et al., 1994]
- SemRep : literature (mostly) [Rindflesch]
- ...

Medical Text Mining

Literature-based discovery

A large proportion of works in Literature-based discovery target medical relations

- Swanson & Smalheiser : Arrowsmith (disease, substance) `http://arrowsmith.psych.uic.edu/cgibin/arrowsmith_uic/start.cgi`
- Hristovski et al. : BITOLA (disease, substance, drug) — maybe_treats `http://www.mf.uni-lj.si/bitola/`
- etc.

Outline of Biomedical NLP Work

A culture of “community collaboration”

Biomedical NLP : Community

- ISMB BioLINK SIG on Text Data Mining
- Mailing lists : BioNLP (2001), ISMB BioLINK
- Workshops
  - BioLINK (2001-), BioNLP, SMBM, LBM. . .
- ACL SIG BioMed (2008)

Biomedical NLP : Lexical Resources

- Issue of recognition of genes/proteins, etc., and their variants
- Large lists of names extracted from reference databases
- Unification initiatives (BioLexicon [Sasaki et al. SMBM 2008])

Biomedical NLP : Ontologies

Active development

- Gene Ontology (GO)
- Gene Regulation Ontology
- Repository, unification : Open Biomedical Ontologies (OBO)

Kim/Rebholz-Schuhmann, SMBM Tutorial, 1/9/2008

Biomedical NLP : Annotated corpora

`http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml`

`http://mars.cs.utu.fi/PPICorpora/`

Corpus	Sentences
LLL (train)	77
HPRD50	145
PDG/PICorpus	283
IEPA	486
BioCreative-PPI (BC I)	1000
BioInfer	1100
AIMed	1955
GENIA	18546/9372 (term/event)
Genetag	20000
ITI TXM (PPI)	75000 (upcoming)

Biomedical NLP : Shared Tools

Example : repositories of UIMA Components

- JCoRe (JULIE Lab, U Jena)
- Tsujii Lab UIMA repository (U Tokyo)
- ClearTK (U Colorado)
- Mayo Clinic (upcoming)
- BioNLP-UIMA Component Repository
- U-Compare (Upcoming ; Tsujii lab, U Colorado, NaCTeM)

Tomanek/buyko/Goetz, SMBM Tutorial, 1/9/2008

Biomedical NLP : Challenges

Organization of shared tasks

- KDD Cup 2002
- TREC Genomics 2003–2007
- JNLPBA 2004
- LLL 2005
- AIMed 2005
- BioCreative I (2004) & II (2006)

Biomedical Text Mining

Literature-based discovery

- Strong interest for Biomedical LBD

Biomedical NLP : Wider Attraction of External Researchers

External to Bioinformatics

- Data mining
  - KDD Cup
- Computational linguistics
  - Workshops at ACL conferences
- Machine learning
  - LLL Challenge

1. Observations
   - MEDLINE View
   - Outline of Medical NLP Work
   - Outline of Biomedical NLP Work

2. Two Domains
   - Clinical Sublanguage
   - Biomolecular Sublanguage

3. User Needs and Tasks
   - Medical NLP : Diverse User Needs
   - Biomedical NLP : More Focused User Needs

4. Impact of Text Genres
   - Open Access
   - Language

5. Attractivity Factors

6. Wrap-up

Two sublanguages

Sublanguage Subset of language within a specialized domain that exhibits specialized constraints due to limitations of the words and relations of the subject matter
(Z. Harris, cited in Friedman et al. [2002])

- Particular word classes
- Particular statement types

A comparison of features of two sublanguages

[Friedman et al., 2002]

- Clinical domain
- Biomolecular domain

Two sublanguages

Sublanguage Subset of language within a specialized domain that exhibits specialized constraints due to limitations of the words and relations of the subject matter
(Z. Harris, cited in Friedman et al. [2002])

- Particular word classes
- Particular statement types

A comparison of features of two sublanguages

[Friedman et al., 2002]

- Clinical domain
- Biomolecular domain

Clinical Sublanguage : Entities

- Descriptions of *entities* and *events* associated with the *patient state*
- Primary concepts
  - disease, procedure, medication, vital sign, symptom, body location
  - mostly nouns
  - modifiers are generally adjectives or nouns

Clinical Sublanguage : Relations

- Simple relations
  - single finding + modifiers
  - verbs are frequently omitted : [*patient had*]
  - *fever and headache* ; *heart was enlarged* ; *pulse measured 70 bpm*
- Complex relations
  - connect findings to (findings | procedures | treatments)
  - with conjunctions (*and*, *with*),
  - prepositions and verbs associated to causality (*due to*, *led to*), etc.

Biomolecular Sublanguage : Entities

- Descriptions of *events* associated with *biomolecular substances* and their *interactions*
- Primary concepts
  - gene, protein, aminoacid... cell, structure, tissue, species
    - creative names*
  - descriptions of biomolecular pathways
    - process, pathway, disease
    - complex interactions and other relations
    - activate, inactivate, attach... signal, substitute, transcribe

Biomolecular Sublanguage : Relations

- Primary relations
  - expressed using verbs of interaction (*p53 binds to il2*)
  - frequently, nominalisations (*activation*) to allow for nesting
- Sequences of interactions
  - highly nested relations
  - *Inhibition of 4 e-bp1 phosphorylation enhanced 4 e-bp1 binding to eif-4e*

```
[action,promote,
    [action,inactivate,x,
        [action,phosphorylate,x,[protein,4 e-bp1]]],
    [action,attach,[protein,4 e-bp1],[protein,eif-4e]]]
```

Two Sublanguages : Summary

Clinical	Biomolecular
Patient reports	Scientific literature
Descriptive	Complex relations between biomolecular substances
Nouns and adjectives	Relations based on verbs

Some overlap

- Tissues, cells, molecular components (markers in pathological reports)
- Diseases (in association with biomolecular interactions)

Two Sublanguages : Summary

Clinical	Biomolecular
Patient reports	Scientific literature
Descriptive	Complex relations between biomolecular substances
Nouns and adjectives	Relations based on verbs

Some overlap

- Tissues, cells, molecular components (markers in pathological reports)
- Diseases (in association with biomolecular interactions)

1. Observations
   - MEDLINE View
   - Outline of Medical NLP Work
   - Outline of Biomedical NLP Work
2. Two Domains
   - Clinical Sublanguage
   - Biomolecular Sublanguage
3. User Needs and Tasks
   - Medical NLP : Diverse User Needs
   - Biomedical NLP : More Focused User Needs
4. Impact of Text Genres
   - Open Access
   - Language
5. Attractivity Factors
6. Wrap-up

Medical NLP User Needs

Those addressed by Medical NLP researchers

- Hospital
  - Patient records : coding and information extraction
  - Decision support :
    - [cf Patient records]
    - knowledge extraction (e.g. from guidelines)
    - literature search (e.g. InfoButtons)
  - Terminology management
- Research
  - Literature search : indexing, information retrieval etc.
  - Hypothesis testing : literature-based discovery

Biomedical NLP User Needs

Those addressed by Biomedical NLP researchers

- Literature search : indexing, information retrieval and co.
- Curation (building databases) : coding and information extraction
- Hypothesis testing : literature-based discovery

Facilitating Factors

More focused user needs

Less diverse tasks

Shared tasks

Facilitating Factors

More focused user needs

Less diverse tasks

Shared tasks

Facilitating Factors

More focused user needs

Less diverse tasks

Shared tasks

1. Observations
   - MEDLINE View
   - Outline of Medical NLP Work
   - Outline of Biomedical NLP Work
2. Two Domains
   - Clinical Sublanguage
   - Biomolecular Sublanguage
3. User Needs and Tasks
   - Medical NLP : Diverse User Needs
   - Biomedical NLP : More Focused User Needs
4. Impact of Text Genres
   - Open Access
   - Language
5. Attractivity Factors
6. Wrap-up

Genres of Input Texts

	Medical	Biomedical
Clinical reports	√	
Terms	√	√
Guidelines	√	
Outbreak reports	√	
Scientific literature	√	√

Different genres of texts induce different constraints

Clinical Reports : Privacy

Requirement : Protection of privacy

Solution : De-identification

- Additional, necessary effort to enable shared research
- Strong limitations on corpus sharing
- CMC ICD9-CM coding challenge
- i2b2 challenges
- Upcoming effort of AMIA NLP WG

The requirement for privacy protection

imposes a burden on Medical NLP research

Clinical Reports : Privacy

Requirement : Protection of privacy

Solution : De-identification

- Additional, necessary effort to enable shared research
- Strong limitations on corpus sharing
- CMC ICD9-CM coding challenge
- i2b2 challenges
- Upcoming effort of AMIA NLP WG

The requirement for privacy protection

imposes a burden on Medical NLP research

Scientific Literature : Open Access

- Catalogue
  - 1997 : MEDLINE access becomes free
- Journals
  - 1997 : PNAS Online
  - 1997 : Mol Biol Cell Online
  - 1998 : BioMed Central
  - 2000 : PLoS (2003 : PLoS Biology)
- Repositories
  - 2000 : Pubmed Central
- Access is still limited for some full-text articles

Open access to the scientific literature

fosters Biomedical NLP research [Bourne et al., 2008]

Scientific Literature : Open Access

- Catalogue
  - 1997 : MEDLINE access becomes free
- Journals
  - 1997 : PNAS Online
  - 1997 : Mol Biol Cell Online
  - 1998 : BioMed Central
  - 2000 : PLoS (2003 : PLoS Biology)
- Repositories
  - 2000 : Pubmed Central
- Access is still limited for some full-text articles

Open access to the scientific literature

fosters Biomedical NLP research [Bourne et al., 2008]

Clinical Reports : Local Languages

Requirement for localisation

Clinical reports must be written in the language of the user

- Doctors write/dictate/read in their own language
- Patients must be able to understand the contents of their files

Clinical Reports : Local Languages

Necessary effort to develop resources for each natural language

- Lexicon
- Morphology
- Terminology
- [Ontology]
- . . .

- POS-tagger
- NER
- Parser
- Relation patterns
- Coreference

Clinical Reports : Local Languages

- A large number of projects have made parallel efforts in different languages
  - German, French, Dutch. . .
- Few coordinated efforts to organize this diversity
  - FP3 MENELAS (1992–1995)
    Analysis of discharge summaries in French, English, Dutch
  - NoE Semantic Interoperability and Data Mining in Biomedicine : WP20, Multilingual medical dictionary
- English is an exception : NLM–UMLS

The multiplicity of local languages

leads to a dispersion of concrete efforts in Medical NLP

Clinical Reports : Local Languages

- A large number of projects have made parallel efforts in different languages
  - German, French, Dutch. . .
- Few coordinated efforts to organize this diversity
  - FP3 MENELAS (1992–1995)
    Analysis of discharge summaries in French, English, Dutch
  - NoE Semantic Interoperability and Data Mining in Biomedicine : WP20, Multilingual medical dictionary
- English is an exception : NLM–UMLS

The multiplicity of local languages

leads to a dispersion of concrete efforts in Medical NLP

Scientific Literature : One Language

One biomedical language

- Most international scientific literature is written in (scientific) English
- Language of experimental science

Scientific Literature : One Language

The unicity of language

simplifies sharing of resources in Biomedical NLP research

[See above]

- Lexicon
- Morphology
- Terminology
- [Ontology]
- . . .

- POS-tagger
- NER
- Parser
- Relation patterns
- Coreference

Scientific Literature : One Language

The unicity of language

simplifies sharing of resources in Biomedical NLP research

[See above]

- Lexicon
- Morphology
- Terminology
- [Ontology]
- . . .

- POS-tagger
- NER
- Parser
- Relation patterns
- Coreference

Scientific Literature : One Text Genre

- Scientific article in experimental science
- Actually, differentiate
  - Abstract
    - Structured abstract
  - Full-text paper
    - Various structures

The unicity of text genre

simplifies the construction of text corpora

Scientific Literature : One Text Genre

- Scientific article in experimental science
- Actually, differentiate
  - Abstract
    - Structured abstract
  - Full-text paper
    - Various structures

The unicity of text genre

simplifies the construction of text corpora

Scientific Literature : One Text Genre

- Scientific article in experimental science
- Actually, differentiate
  - Abstract
    - Structured abstract
  - Full-text paper
    - Various structures

The unicity of text genre

simplifies the construction of text corpora

A Conjunction of Enabling Factors

Single type of text

Public access to input texts

Single language

Shared corpora

A Conjunction of Enabling Factors

Single type of text

Public access to input texts

Single language

Shared corpora

A Conjunction of Enabling Factors

Single type of text

Public access to input texts

Single language

Shared corpora

A Conjunction of Enabling Factors

Single type of text

Public access to input texts

Single language

Shared corpora

1. Observations
   - MEDLINE View
   - Outline of Medical NLP Work
   - Outline of Biomedical NLP Work
2. Two Domains
   - Clinical Sublanguage
   - Biomolecular Sublanguage
3. User Needs and Tasks
   - Medical NLP : Diverse User Needs
   - Biomedical NLP : More Focused User Needs
4. Impact of Text Genres
   - Open Access
   - Language
5. **Attractivity Factors**
6. Wrap-up

Intrinsic Attractivity of Genomics

- Help health information processing
  - Health-related issues : a good deed
  - Much medico-economic motivation though
- Help biomedical research
  - Scientifically appealing
  - Promise of more fundamental outcomes
  - Scientific discoveries

Funding

Funding for research in medical information processing

- Fluctuations over the years

Funding for biomedical research

- Sustained level of funding since genome sequencing

Resources

A variety of shared resources

- Input text collections
- Lexical, terminological, ontological resources
- NLP/IE tools

facilitates entry of new players

- Bioinformaticians
- [General] computational linguistics researchers
- Machine learning researchers
- Industry

Annotated Corpora

Most crucial to progress in the field are annotated corpora

- Analysis
- Evaluation
- Training
  - Enables the use of machine learning methods

This is possible thanks to

- Open access
- One language
- Defined, common tasks

Annotated Corpora

Most crucial to progress in the field are annotated corpora

- Analysis
- Evaluation
- Training
    - Enables the use of machine learning methods

This is possible thanks to

- Open access
- One language
- Defined, common tasks

Challenges

A driving force in all domains—when possible

- Focus efforts
- Enable comparison of [methods and] systems
- Co-operative definition of tasks
- Comparative evaluation
- Clearly defined framework

Depend on tasks and corpora

1. Observations
   - MEDLINE View
   - Outline of Medical NLP Work
   - Outline of Biomedical NLP Work
2. Two Domains
   - Clinical Sublanguage
   - Biomolecular Sublanguage
3. User Needs and Tasks
   - Medical NLP : Diverse User Needs
   - Biomedical NLP : More Focused User Needs
4. Impact of Text Genres
   - Open Access
   - Language
5. Attractivity Factors
6. Wrap-up

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

Wrap-up : Biomedical *vs* Medical NLP

Less diverse tasks

More focused user needs

Single type of text

Public access to input texts

Single language

Shared tasks

Shared corpora

Machine learning

Challenges

References I

A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus : The MetaMap program. *J Am Med Inform Assoc*, 8 :17–21, 2001.

A. R. Aronson, O. Bodenreider, F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch, and J. Wilbur. The NLM indexing initiative. *J Am Med Inform Assoc*, 7 :17–21, 2000.

P. E. Bourne, J. L. Fink, and M. Gerstein. Open access : Taking full advantage of the content. *PLoS Computational Biology*, 4(3) :e1000037, Mar. 2008. doi : 10.1371/journal.pcbi.1000037.

D. Demner-Fushman, S. M. Humphrey, N. C. Ide, R. F. Loane, J. G. Mork, M. E. Ruiz, P. Ruch, L. H. Smith, W. J. Wilbur, and A. R. Aronson. Combining resources to find answers to biomedical questions. In *Sixteenth Text Retrieval Conference TREC-2007*, pages 205–215, Gaithersburg, MD, 2007.

C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2) :161–174, 1994.

C. Friedman, P. Kra, and A. Rzhetsky. Two biomedical sublanguages : a description based on the theories of Zellig Harris. *J Biomed Inform*, 35(4) :222–235, 2002.

M. A. Hearst. What is text mining ? `http://www.ischool.berkeley.edu/~hearst/text-mining.html`, 2003.

D. Rebholz-Schuhman, G. Cameron, D. Clark, E. van Mulligen, J. Coatrieux, E. Del Hoyo Barbolla, F. Martin-Sanchez, L. Milanesi, I. Porro, F. Beltrame, I. Tollis, and J. Van der Lei. SYMBIOmatics : synergies in Medical Informatics and Bioinformatics–exploring current scientific literature for emerging topics. *BMC Bioinformatics*, 8(8 Suppl 1) :S18, Mar. 2007. doi : 10.1186/1471-2105-8-S1-S18.

N. Sager, C. Friedman, and M. S. Lyman, editors. *Medical Language Processing : Computer Management of Narrative Data*. Addison Wesley, Reading, MA, 1987.