BOOTStrep

Bootstrapping Of Ontologies and Terminologies STrategic REsearch Project

The BOOTStrep BioLexicon: a Lexical Resource for Biomedical Text Mining

Simonetta Montemagni

ILC-CNR

simonetta.montemagni@ilc.cnr.it

Outline

- The BioLexicon
  - Who
  - Why
  - How
  - What
  - Where from
    - Verbs
  - Representation
  - How many entries
  - Evaluation
  - Distribution
  - Conclusions

"I like to verb words …"

The BioLexicon: **who**

Joint and collaborative work of the following teams:

- **D. Rebholz-Schuhmann, P. Pezik, V. Lee, J.J. Kim**
  European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD

EMBL-EBI

European Bioinformatics Institute
is an Outstation of the
European Molecular Biology Laboratory.

- **N. Calzolari, M. Monachini, S. Montemagni, R. del Gratta, S. Marchi, V. Quochi, G. Venturi**
  ILC-CNR, Area della Ricerca del CNR, Via Giuseppe Moruzzi N° 1, 56124 Pisa, Italy

- **S. Ananiadou, J. McNaught, Y. Sasaki, Paul Thompson**
  School of Computer Science, The University of Manchester, 131 Princess Street, M1 7DN, UK

MANCHESTER
1824
The University of Manchester

The BioLexicon: **why**

- Text Mining needs information about words
  - the lexical component still remains a major bottleneck
- TM systems in the biomedical domain must be provided with a substantial lexicon covering a realistic vocabulary and providing the kinds of linguistic information appropriate to grasp the knowledge embedded in texts
  - Biomedical term variants (orthographic, semantic, geographical, …)
    - better information retrieval
  - Terminological verbs and their combinatorial properties (subcategorization frames and predicate-argument structure)
    - better information extraction and question answering
  - Word derivations
    - to reach similar meaning expressed in different ways (e.g. *activation* vs *activate*)

The BioLexicon: **how**

- **General Requirements**

  Modularity, extensibility, conformity to standards, reusability

- **Biomedical Domain Specific Requirements**

  Gene names, protein names, bio-events and participants, …

- **Linguistic/Terminological Requirements**

  term variants, source identifiers, acronyms, syntactic and semantic properties of terms, …

- **Text Mining / Machine Learning Requirements**

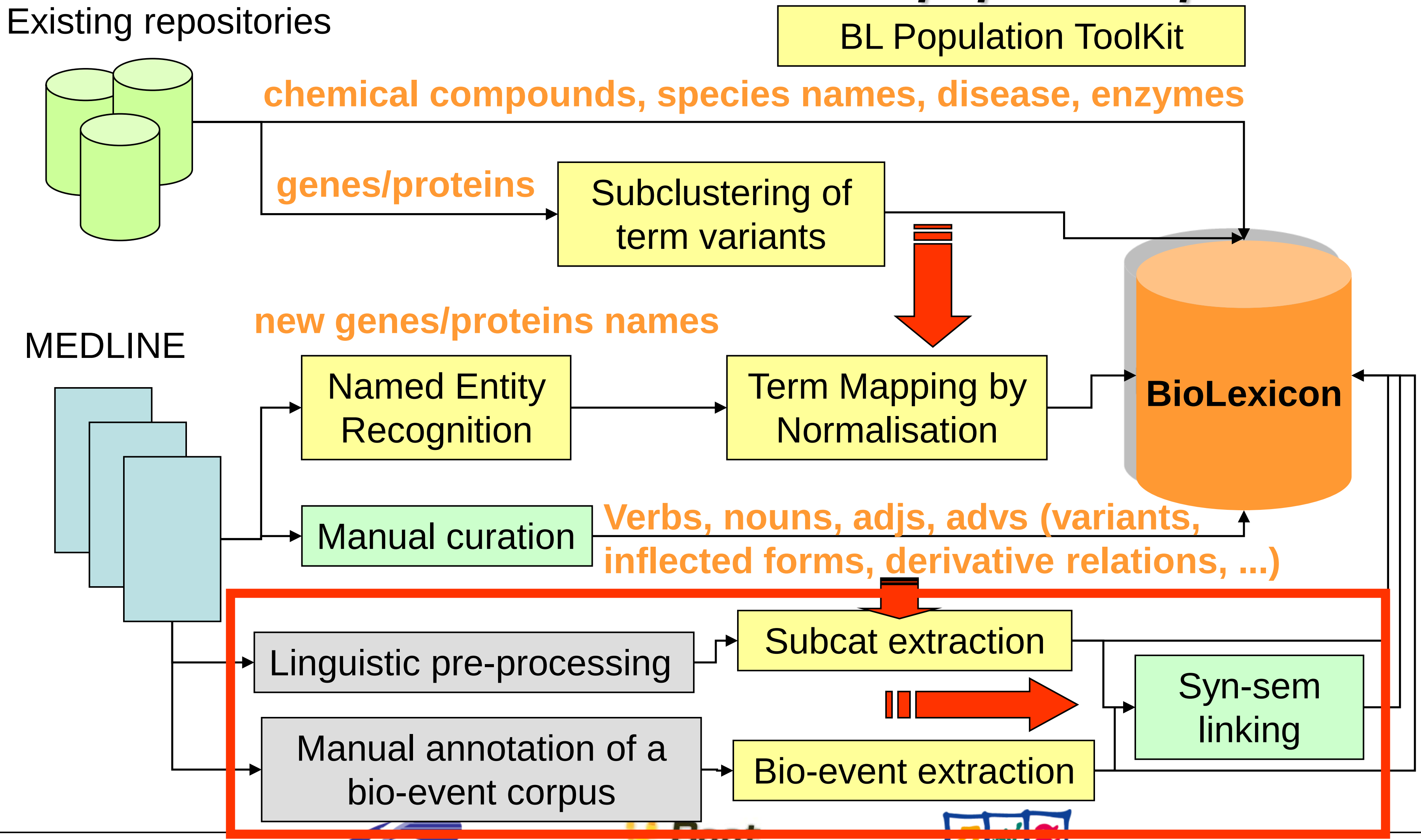
  Confidence scores for automatically extracted info (e.g. variants, subcusterizations, subcat frames, …)

The BioLexicon: **what**

- integrated lexical-terminological resource of ~2.2M lexical entries for bio-text mining with information about
  - nouns, verbs, adjectives, adverbs
  - both domain-specific and general language words
- populated with terms gathered from
  - available biomedical sources
  - texts (biomedical literature)
- including rich linguistic information ranging over different linguistic descriptions levels
  - e.g. derivational morphology, subcategorization patterns, predicate argument structure, syntax-semantics linking
- combining features of both terminologies and open-domain computational lexicons
- conforming to international lexical representation standards (the ISO/DIS 24613 “Lexical Mark-up Framework”)
- providing links to the Gene Regulation Ontology

The BioLexicon: **where from**

Incremental population process



Existing repositories

BL Population ToolKit

chemical compounds, species names, disease, enzymes

genes/proteins

Subclustering of term variants

MEDLINE

new genes/proteins names

Named Entity Recognition

Term Mapping by Normalisation

BioLexicon

Manual curation

Verbs, nouns, adjs, advs (variants, inflected forms, derivative relations, ...)

Linguistic pre-processing

Subcat extraction

Syn-sem linking

Manual annotation of a bio-event corpus

Bio-event extraction

The BioLexicon: **verbs**

- Accurate bio-TM applications focused on event extraction require lexical resources providing an exhaustive account of the **semantic and syntactic combinatorial properties of lexical units** conveying **event information**
  - use of different predicates to describe events
    - E.g. *methylate, phosphorylate*
  - general language predicates may have different properties
    - E.g. *the patient **presented** with influenza to the doctor* vs *the patient **presented** the doctor with influenza*
- Current resources
  - BioFrameNet and PASBio: corpus-based but small-scale resource
  - SPECIALIST lexicon: wide coverage but not corpus-driven
- Need for **large-scale domain-specific** lexical resource providing predicate-argument information
  - based on domain-specific corpora
  - containing both syntactic and semantic information

The BioLexicon: **verb information**

- bootstrapped from biomedical corpora
  - the most relevant verbs are included in the lexicon
  - their encoded behaviour is domain-specific
- containing *both* syntactic and semantic information
  - syntactic subcategorization (e.g. *act* ARG1#PP-*as*#)
  - semantic event frame information (e.g. *bind* AGENT#THEME#LOC#)
  - explicit link between the two (e.g. *express* AGENT>ARG1#THEME>ARG2#LOC>PP-*in*#)
- acquired semi-automatically
  - **syntactic frames**: extracted through unsupervised learning on dependency-annotated corpus (Enju parser) of approximately 6 million tokens (MEDLINE abstracts on E.Coli as well as full papers)
  - **semantic event frames**: based on a manually annotated corpus of gene regulation bio-events (677 abstracts)
  - **syntax-semantics linking**: added manually for those verbs showing both info types (168 verbs)

The BioLexicon: **verb subcat** (1)

- particular requirements for Subcategorization Frames (SCFs) in biomedical language
  - average number of arguments in SCFs higher than general language
- "tabula rasa" approach to SCF extraction: no a priori knowledge about the set of possible SCFs
  - no distinction between argument/modifier
  - 92 different induced SCFs
    - SPECIALIST Lexicon: very limited number of complementation patterns
- SCFs complemented with information about individual dependencies of verbs
  - peculiar status of prepositional phrases in bio-texts
  - many of the strongly selected modifiers spread over different SCFs
    - radically underestimated role
  - typical verbal dependencies, corresponding to either arguments or strongly selected modifiers, detected through the ll association score

The BioLexicon: **verb subcat** (2)

Verb	SCF	p(subcat|v)	% of passive usages
activate	ARG1#ARG2#	0.59	0.20
	ARG1#ARG2#PP-by#	0.05	0.28
	ARG1#	0.28	0
	ARG1#ARG2#PP-in#	0.08	0.44

Full parsing

Verb	*SLOT*	*ll score*	*% of passive usages*
activate	ARG2	4566.23	0.27
	PP-in	124.59	0.46
	PP-by	452.15	0.35

Preposition-based parsing

The BioLexicon: **bio-event frames** (1)

- Extracted from a corpus of 677 MEDLINE abstracts manually annotated by biologists
- Annotation consisted of:
  - Identifying relevant *gene regulation* events centred on verbs and nominalised verbs (e.g. *expression*)
  - Finding all semantic arguments in same sentence
    - Syntactic representation used to constrain chosen spans
  - Assigning a semantic role to each argument
    - A set of 13 event-independent roles were defined for the task
  - Assigning named entity types to semantic arguments (where appropriate)
    - A hierarchy of NEs, specially tuned to *gene regulation,* was created
    - Organised into five entity-specific super-classes

The BioLexicon: **bio-event frames** (2)

verb	Bio-event frames
activate	Agent#Theme#
	Agent#Theme#Condition#
	Agent#Theme#Location#
	Agent#Theme#Manner#
	Agent#Theme#Source#
	Theme#
	Theme#Condition#

verb	Bio-event frames with NE types
activate	Agent-DNA#Theme-DNA#
	Agent-Organisms#Theme-Protein#
	Agent-Protein#Theme-DNA#

The BioLexicon:
syntax-semantics linking

different types of mapping were performed:

activate	Theme	ARG2	0	ARG1			RED
	Agent	ARG1	Theme	ARG2	Condition	PP-in	ISO
	Agent	ARG1	Theme	ARG2	Manner	PP-by	ISO
	Agent	ARG1	Theme	ARG2	Location	PP-in	ISO
	Agent	ARG1	Theme	ARG2	Source	0	AUG
	Theme	ARG2	Condition	PP-in	0	ARG1	RED
	Agent	ARG1	Theme	ARG2			ISO
	Theme	ARG1					ISO
	Agent	ARG1	Theme	ARG2	Manner	PP-in	ISO

Useful information for **mixed syntax-semantics approaches**

The BioLexicon: **representation model**

- The BL model is conformant to ISO-LMF (ISO 24613:2008)
  - **high-level objects**: the meta-model, i.e. a set of independent lexical objects with relations among them
  - **low-level objects**: a set of Data Categories, i.e. linguistic *constants* in the form of attribute-value pairs (either drawn from the ISO-12620 or defined for the special domain)
- XML DTD for the entire lexicon
- The implementation consists of a flexible, extensible relational MySQL database
- Automatic population procedures relying on a dedicated input data structure, the BioLexicon XML Interchange Format (XIF)
- An XML LMF conformant export function is available

The BioLexicon: **the starting point**

Semantic type	Resources
Cell	Cell ontology
Cell Component	Gene Ontology GO:0005575 cellular component
Chemical	CHEBI, IMR:0000947 chemical
Disease	OMIM
Enzyme	Enzyme commission
Gene	BioThesaurus
Ligand	IMR - INOH Protein name/family name ontology
Nuclear Receptor	GO:0004879 ligand-dependent nuclear receptor activity

Semantic type	Resources
NucleicAcid Region	Sequence Ontology :Region
Operon	RegulonDB, ODB (Operon DataBase)
Organism	NCBI Species
Transcription Factor-BindingSite	Sequence Ontology
Protein	BioThesaurus
Protein Complex	Corum database
Protein Domain	InterPro
Transcription Regulator	RegulonDB, TransFac, Gene Ontology Annotation

The BioLexicon **by numbers**

Entries and variants by semantic type

Sem. Type	# Entries	# Variants
Gene/Prot	1640608	1408312
Gene/Prot (synsets)	358335	936126
Organisms	482992	182610
Enzymes	4016	4164
Protein Domains	16940	15412
Protein Compl.	2104	418
Chemicals	19637	77475
Diseases	19457	11314
Molecular Roles	8850	29831
Cell	842	512
Trans. Factors	160	129
Operons	2672	368
Sequences	1431	741

Entries by part of speech

POS	# entries
Nouns	2231574
Adjectives	3428
Verbs	1154
Adverbs	550

Verbs

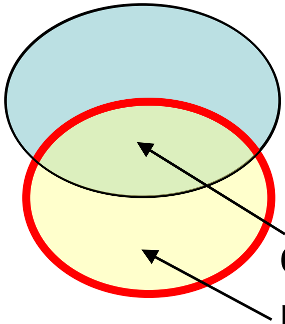
	domain-specific	general
	658	496
inflected forms	6261	
related entries (e.g. absorb -> absorption/N, absorber/N, absorbing/J, absorbable/J, absorbent/J, absorbently/R)	2763	-
verb-SCF associations	1404	-
verb-SLOT associations	1710	-
bio-event frames	856	-
syntax-semantics mappings (concerned with 168 verbs)	668	-

The BioLexicon: **intrinsic evaluation**

- Comparison with two existing large-scale dictionaries
  - WordNet: General English Thesaurus
  - NLM Specialist Lexicon: Biomedical Lexicon
- Coverage evaluation

WordNet

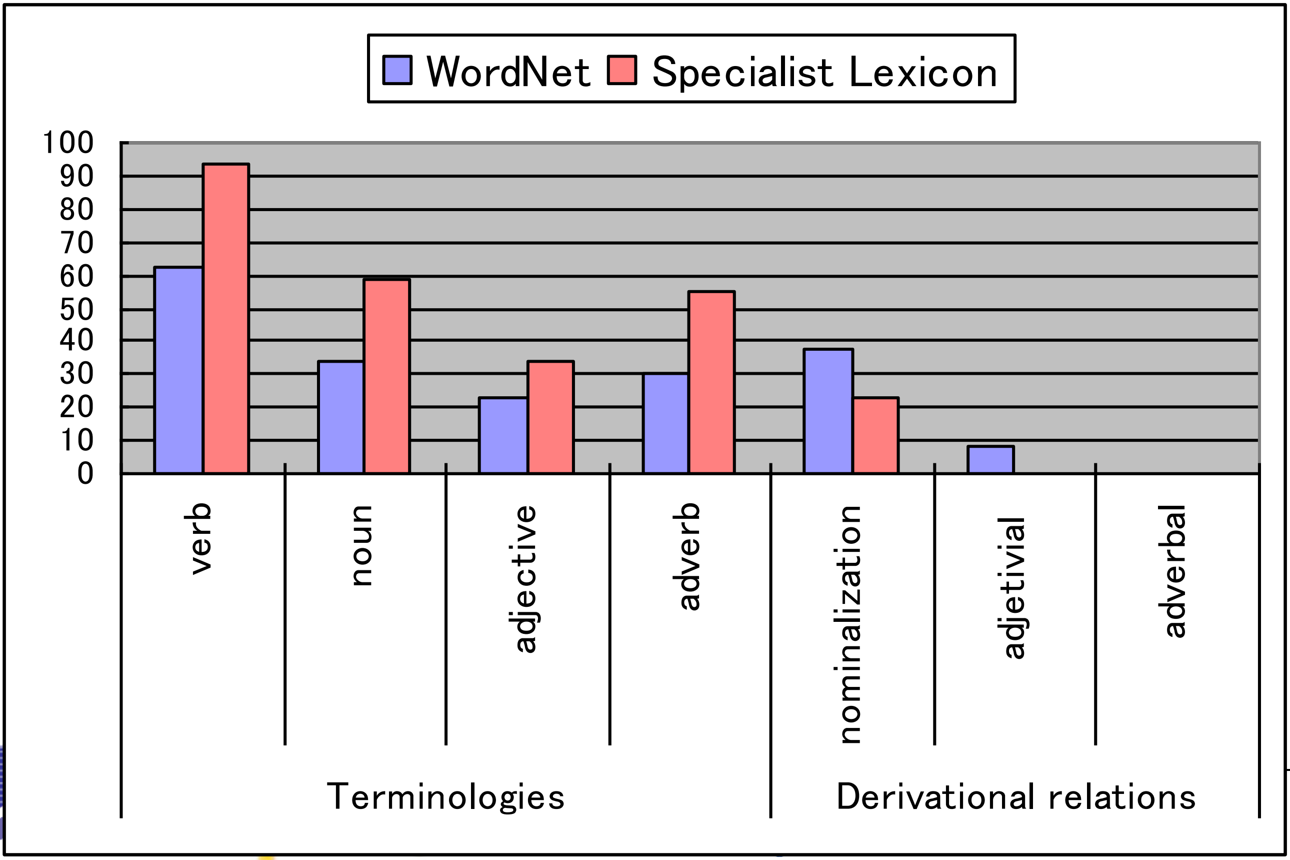
Specialist Lexicon



covered

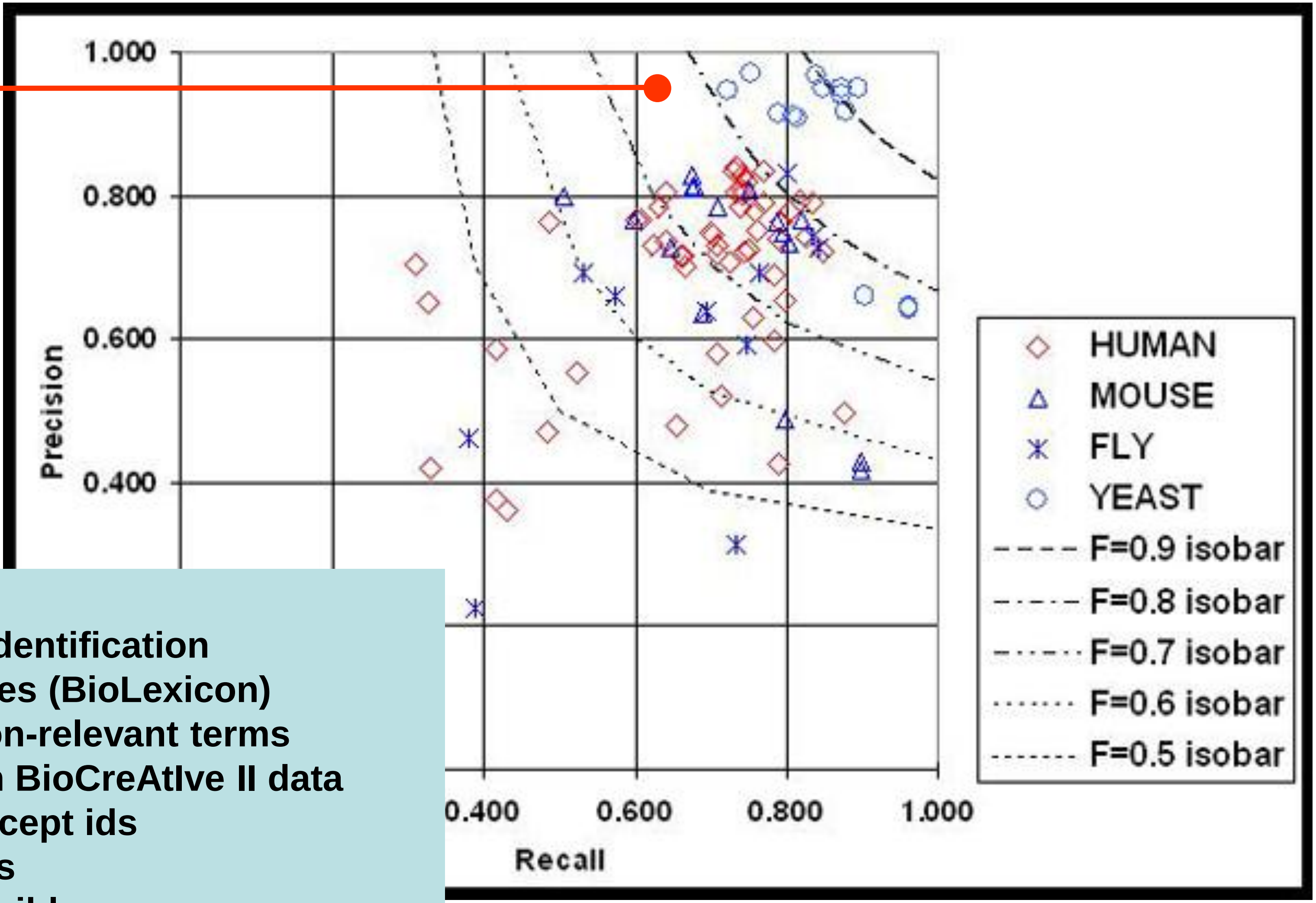
not covered

BioLexicon



BioLexicon in BioCreAtlve II, GN task

Biolexicon, human



Applied methods:

- Abner for gene identification

- Statistical features (BioLexicon) for filtering of non-relevant terms

- Classification on BioCreAtlve II data

- Only human concept ids

⇒ Baseline results

⇒ Highly reproducible

⇒ Available as Whatizit module (BioLexHuman)

The BioLexicon: **extrinsic evaluation**

Task-based evaluation (still ongoing)

Task	Data	Tool
IR	•TREC Genomics Track 2007	•BLTagger •NeMine (NER)
IE	UoM Gene Regulation Corpus	•BLTagger •NeMine (NER) •Enju with the BL

NeMine (http://text0.mib.man.ac.uk/~sasaki/bootstrep/nemine.html)

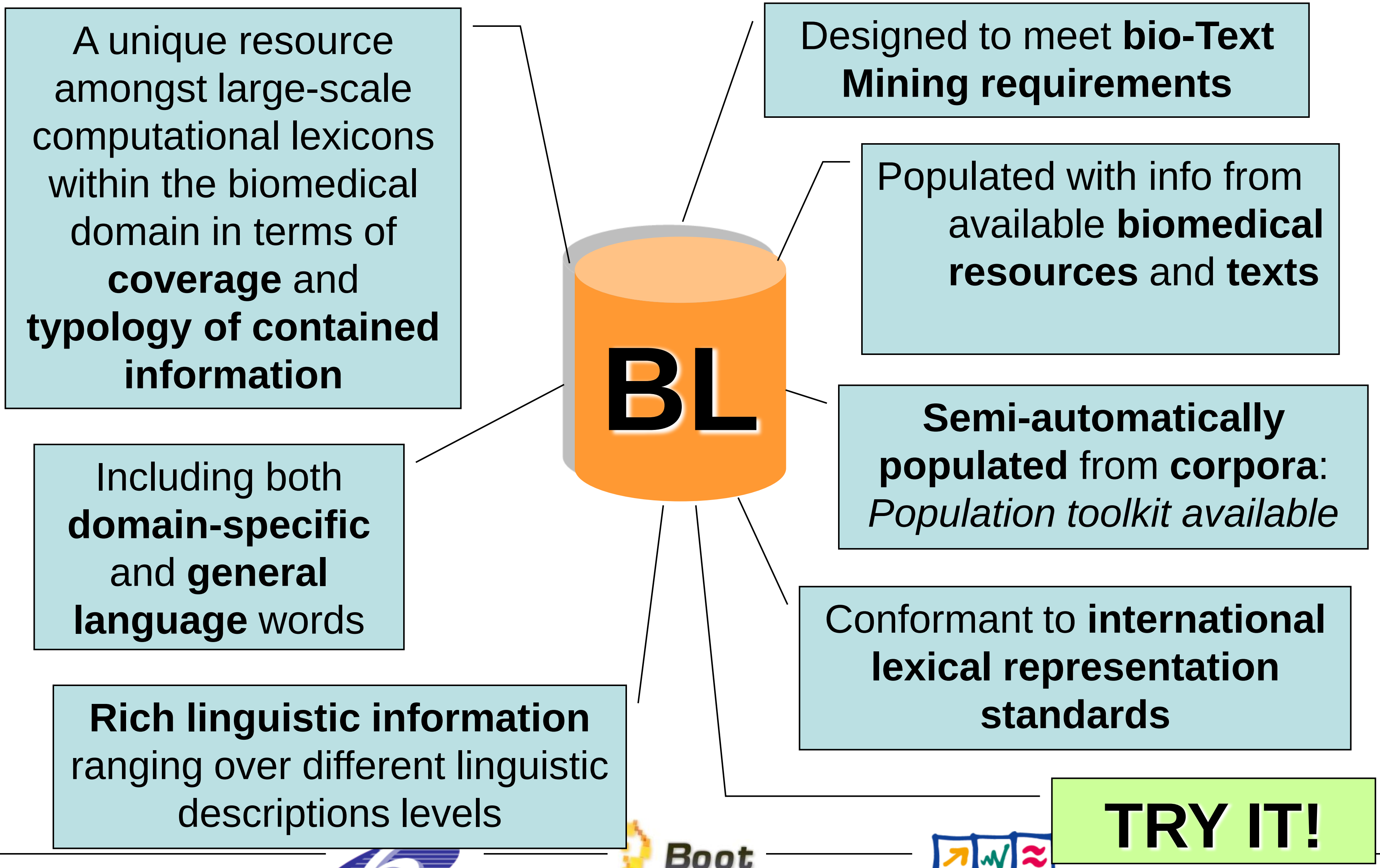
BioLexicon **distribution**

- The BioLexicon (MySQL version) is distributed through the ***European Language Resources Association*** (**ELRA**)
  - http://www.elra.info or http://www.elda.org

- Benefits
  - Servicing of bug reporting through ELRA
  - Organisational embedding into other lexical resources
  - Long-term availability
  - Support to European language infrastructures

- Different licence types for
  - Commercial use
  - Research use by commercial organisations
  - Research use by academic organisations

Conclusions



A unique resource amongst large-scale computational lexicons within the biomedical domain in terms of **coverage** and **typology of contained information**

Including both **domain-specific** and **general language** words

Rich linguistic information ranging over different linguistic descriptions levels

Designed to meet **bio-Text Mining requirements**

Populated with info from available **biomedical resources** and **texts**

Semi-automatically populated from **corpora**: *Population toolkit available*

Conformant to **international lexical representation standards**

TRY IT!

THANK YOU