

Static Dictionary Features for Term Polysemy Identification

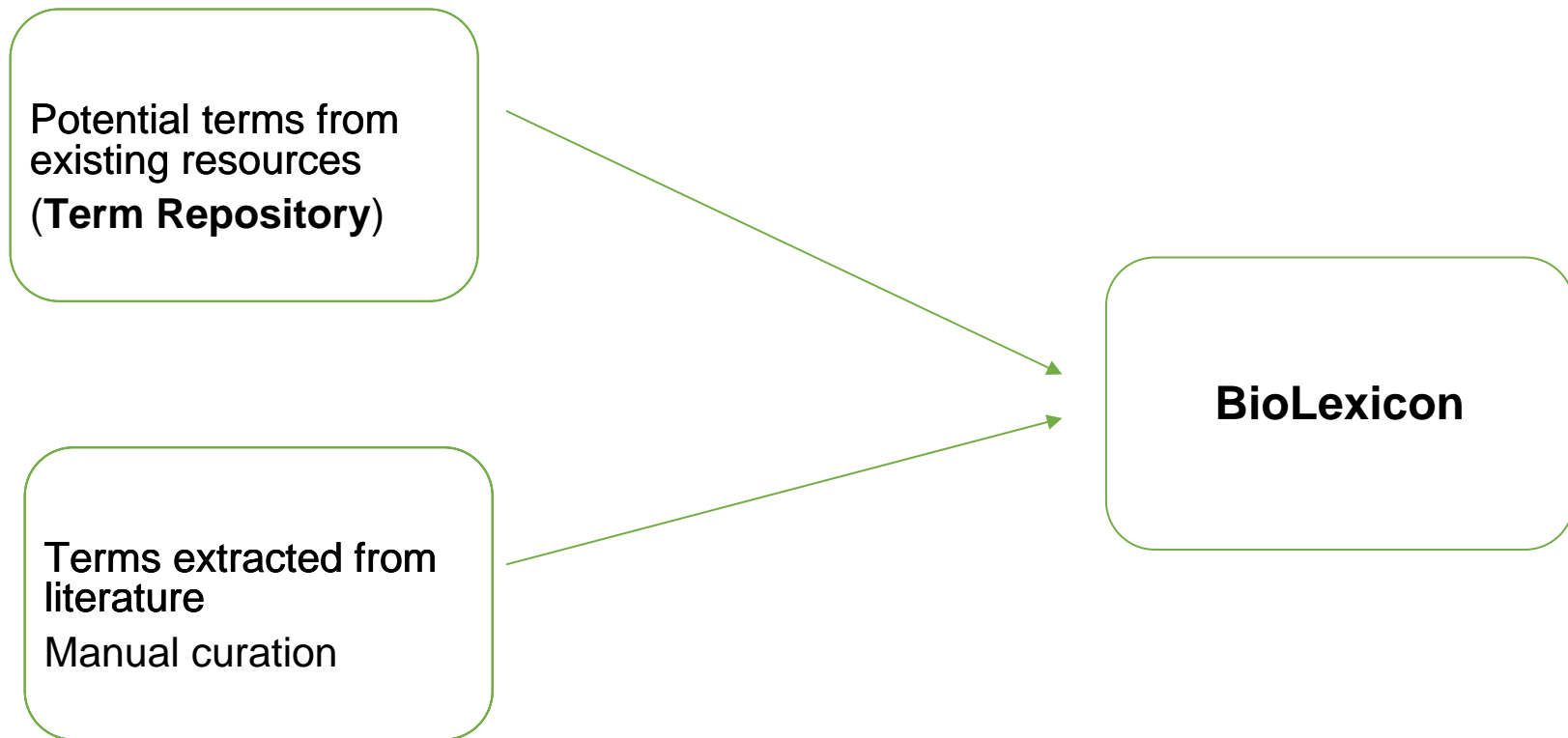
P. Pezik, A. Jimeno, V. Lee, D. Rebholz-Schuhmann

EMBL-EBI



Term Repository and BioLexicon

- A large lexical resource compiled as part of the BootStrep project



Term Repository I

Semantic Type	Synsets	Variants
Chemical entities	13,473	57,581
Enzyme names	4,016	7,658
PGNs	232,258	1,931,786
Species names	367,565	441,993

Terms organized into sets of synonymous variants

Term repository II

Semantic type	Resources
Cell	Cell ontology
CellComponent	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
NuclearReceptor	GO:0004879 ligand-dependent nuclear receptor activity
NucleicAcidRegion	Sequence Ontology :Region
Operon	RegulonDB, ODB (Operon DataBase)
Organism	NCBI Species
TranscriptionFactorBindingSite	Sequence Ontology
Protein	BioThesaurus
ProteinComplex	Corum database
ProteinDomain	InterPro
TranscriptionRegulator	RegulonDB, TransFac, Gene Ontology Annotation

Manually curated term sets (e.g. biologically relevant verbs)

Identifying term polysemy

- With more than 16 semantic types some internal term ambiguity can be captured by checking the number of synsets the term belongs to (*chicken* as a (pseudo-) protein name and as synonym of *Gallus gallus*).
- Because of the focus of the repository, most terms are domain-specific. Some cases of polysemy could never be indicated by the resource (e.g. *WHO* as a protein name).

Why indicate term ambiguity?

- Provided that indicators of ambiguity are available for a given term (per unique string) BioLexicon could be more easily applied to
 - IR (e.g. Query expansion). Conservative query expansion minimizing the risk of query drift.
 - IE (NER). Enriching and standardizing access to the feature set used for NER.
- Such indicators are static in that they are independent of the context in which a given term is used

Term	Ambiguity
chicken	high
Chicken tolloid-like protein 1	low

Identifying static polysemy indicators

- Identify the types of polysemy for the most numerous semantic type in Term Repository – protein and gene names
- Design a set of features directly indicating one or more polysemy types
- Provide an annotated corpus for E. coli names
- Evaluate the contribution of static polysemy indicators to the performance of a NER solution (PGN normalization). Static dictionary features are evaluated separately from context-dependent ones.
- Analogy: POS-tagging for English has been claimed to be 90% accurate with only the following two rules:
 - Use the more probable POS (static dictionary probability)
 - Annotate anything unknown as a proper noun

Major types of PGN polysemy

1. A PGN has a common English word homograph. We call this a case of domain-independent polysemy, e.g. (but, WHO). Sometimes this type of polysemy is introduced by pseudo terms by resulting from the poor quality of a lexical resource, e.g. Biothesaurus contains partial PGN terms such as human or, due to the fact that they were gathered from less trustworthy database description fields.
2. A PGN has a number of hyponyms and it is sometimes used synonymously with them. Examples of this type of polysemy include generic enzyme names, such as *oxidoreductase*). Sometimes a more specified case of holonymy triggers similar ambiguity, e.g. an operon name can be interpreted to denote any of the genes it contains. We call this a case of vertical polysemy (c.f. Fellbaum 1998).
3. A PGN is used for a number of orthologous or otherwise homologous genes. Thus the ambiguity in the gene name results from the fact that the same name is used for structurally identical genes found in different species.
4. A PGN has a biomedical homograph, e.g. retinoblastoma. We refer to this as a case of domain-specific polysemy (Jimeno et al. 2008).
5. Last but not least the very use of the umbrella term PGN suggests another type of polysemy, where the same name is used to denote a gene and its product. Generally, however, gene names are not distinguished from protein names.

#	Feature	Polysemy type
1	BNC frequency	1
2	Number of synsets	2,3
3	NCBI taxonomy ids	3
4	Generic enzyme	2
5	Medline frequency	4,1
6	MESH nodes	4

Training and test corpora

- BioCreAtivE human gene normalization
- E. coli PGN corpus (109 Medline abstracts annotated at exact mention level, 96 used at the time of writing the paper). Annotator agreement still to be completed.

- <MedlineCitations>

- <MedlineCitation Owner="NLM" Status="MEDLINE">

<PMID>6997263</PMID>

<ArticleTitle>Regulation of Escherichia coli K-12 hexuronate system genes: exu regulon.</ArticleTitle>

- <AbstractText>

Two types of Escherichia coli K-12 regulatory mutants, partially or totally negative for the induction of

<zuniprot uniprotid="P0A8G3">uronic isomerase</zuniprot>

,

<zuniprot uniprotid="P0A8G3">uxaC</zuniprot>

; altronate oxidized nicotinamide adenine dinucleotide:

<zuniprot uniprotid="P0A6L7">uxaB</zuniprot>

;

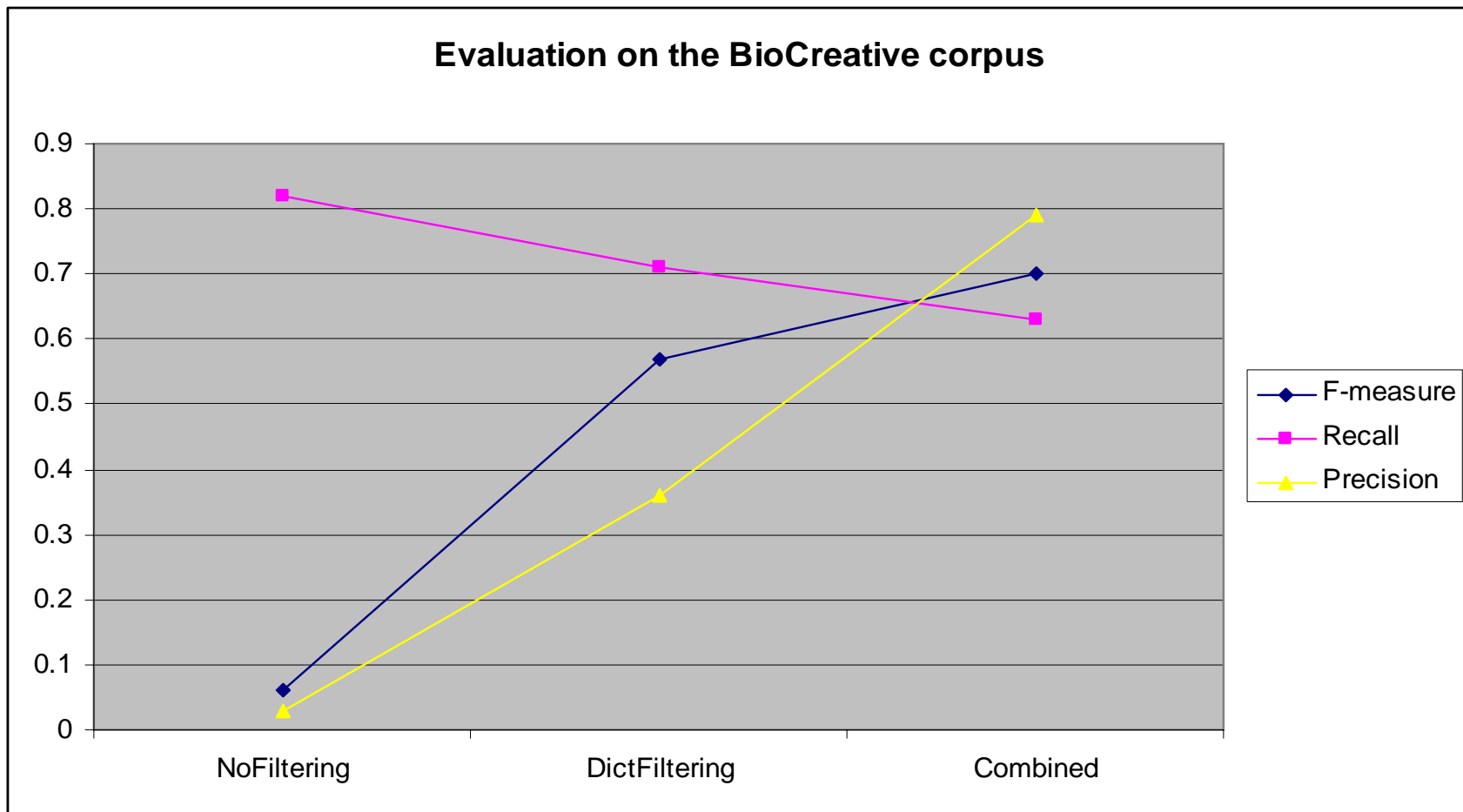
<zuniprot uniprotid="P24215">mannonate hydrolyase</zuniprot>

,

Training

- C4.5 decision tree trained on the corpora
- Performance of NER based on static dictionary features measured first
- Contribution of context-driven features measured separately

PGN normalization - BioCreAtivE



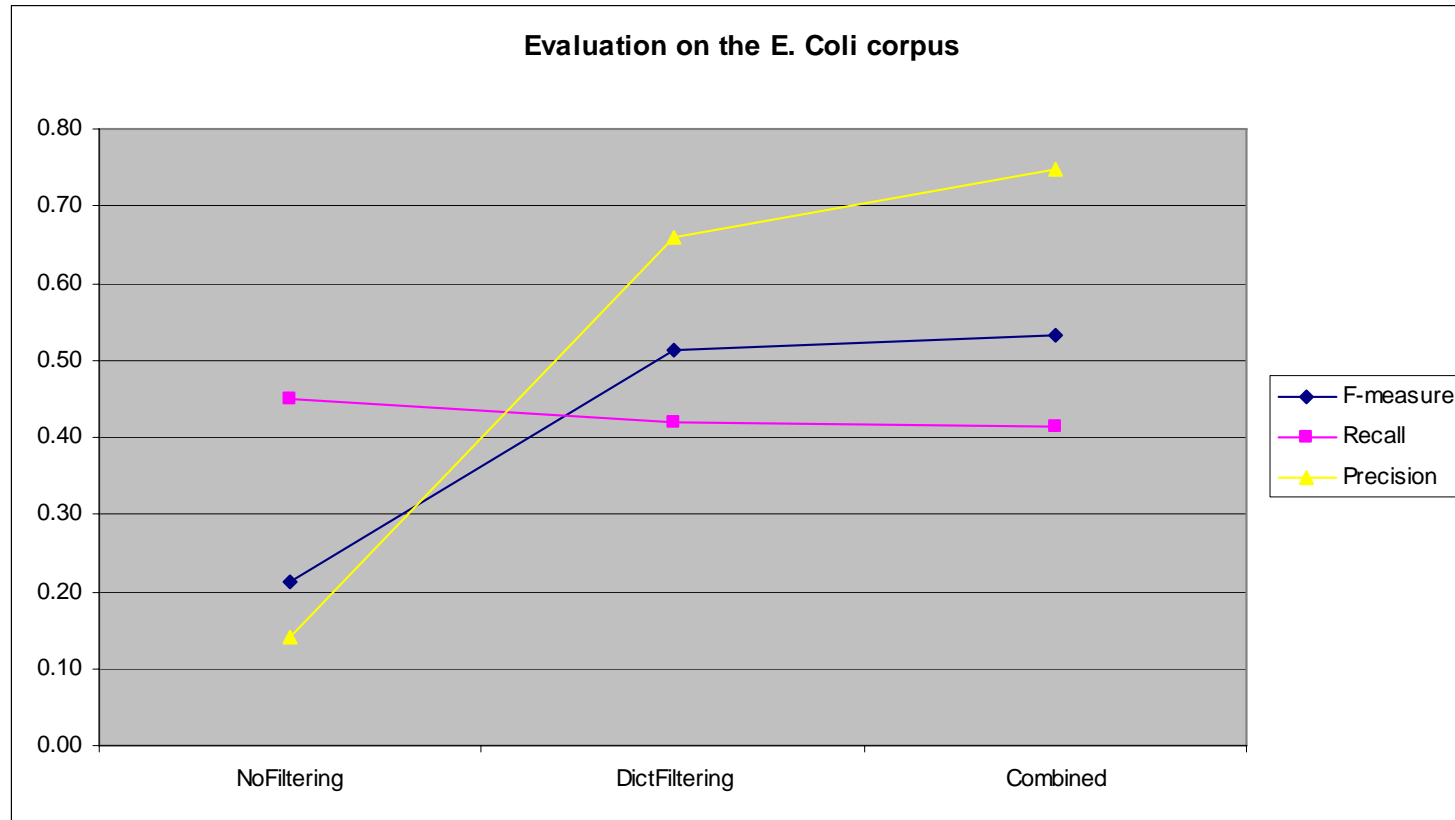
Major decision space splits for human PGNs

1. BNC frequency
2. Medline frequency
3. Number of synsets where a term occurs
4. Number of distinct species taxonomy identifiers

```
bnc_freq > 15
|  medlineFrequency <= 8472
|  |  nbofSynsets <= 21
|  |  |  medlineFrequency <= 787: false (3.0)
|  |  |  medlineFrequency > 787
|  |  |  |  bnc_freq <= 472: true (11.0/2.0)
|  |  |  |  bnc_freq > 472: false (2.0)
|  |  |  nbofSynsets > 21: false (3.0)
|  |  medlineFrequency > 8472: false (104.0)
```

Term	Polysemy type
chicken	1
alternative	1
tissue	1,4
translocation	4
p63	3
polymerase	2

PGN Normalization – E. Coli PGNs



Comparing results

- To what extent does the different annotation of PGNs in the *E. coli* corpus account for the differences in the results obtained? (Alex, 2006; Shipra et al. 2004)
- The initial recall of *E. coli* PGNs is only 0.45, which may be partly due to the occurrences of mutant genes that have not been recorded in existing PGN resources used.
- Another major reason for the initially low recall is the occurrences of operon names, which we annotate with several identifiers matching all the genes on a given operon. As an example, we have assigned as many as 9 matching identifiers to the *TOR* (*trimethylamine N-oxide reductase*) operon in the *E. coli* corpus. Not all of these gene names are associated with this operon in the lexical resources we have used.
- Yet another reason for the relatively low recall is the variability of operon names (e.g. *cyoABCDE* may stand for *cyoA*, *cyoB*, etc.), which occur in the corpus relatively frequently because of its gene-regulation focus. The drop in the recall as we apply the dictionary-filtering rules is rather insignificant (0.45 to 0.42) compared with the gain in the precision (from 0.15 to 0.66).

Conclusions

- Demonstrated how a set of features that provide indications of different polysemy types can be assigned statically to entries in a lexical resource
- In principle, the features can be applied to any other semantic type in Term Repository (currently carrying out a similar experiment with chemical names based on a gold standard corpus provided by the EPO)
- Although disambiguation based on static dictionary features does not outperform fully-fledged NER, it does effectively filter out highly polysemous terms and contributes to the performance of a NER system => it's worth including such information in a terminological resource
- Once computed and assigned to terms in the lexical resource, static polysemy indicators could be used for more conservative query expansion or relevance feedback, independently of the context in which they occur. Still needs to be evaluated.

Availability of the E.coli corpus

- <ftp://ftp.ebi.ac.uk/pub/software/textmining/bootstrep/ebicoli/>



3rd International Symposium on Semantic Mining in Biomedicine



Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)

D. Rebholz-Schuhmann
EMBL-EBI, U.K.

T. Salakoski
TUCS, Turku, FI

The SMBM'08 electronic submission system is now open. To submit your manuscript, please follow the instructions on the [Instructions for authors](#) page.

September 1-3 2008, Turku, Finland

The third International Symposium on Semantic Mining in Biomedicine (SMBM 2008), hosted by Turku Centre for Computer Science (TUCS), aims to bring together different communities: researchers from text and data mining in biomedicine, medical-, bio- and cheminformaticians, and researchers from biomedical ontology design and engineering. This is the follow-up event of SMBM 2005 (EMBL-EBI, U.K.) and SMBM 2006 (Friedrich-Schiller University of Jena, Germany).

SMBM 2008 is a three-day event with a tutorials day held on September 1st and the main conference on September 2-3.

Invited speakers

- ◆ Alfonso Valencia ([CNIO](#), Madrid, Spain)
- ◆ Pierre Zweigenbaum ([LIMSI-CNRS](#), Orsay, France)

Organizers



ECCB 2008 Workshop: Annotation, interpretation and management of Mutations (AIMM)

Mutations play a key role in the understanding of genetic mechanisms and complex diseases. This workshop will focus on solutions in text mining, data warehousing and machine learning that allow better integration of mutation relevant information into a bioinformatics infrastructure (e.g. workflows, databases, machine learning techniques). Altogether, the meeting participants will discuss the methods for the prediction of phenotypic effects induced by mutations (e.g. by text mining), the support to clinical decision processes involving mutations and the means that allow access and management of mutations with annotations from different data resources. Synergistic use of these technologies should facilitate inferences of knowledge from sequence to structure to function and to phenotypes. The workshop brings together members of different disciplines to improve know-how and technology transfer as well as better hypothesis generation for yet un-annotated mutations.

Organizers

Christopher J. O. Baker , PhD
Principal Investigator
Data Mining Department
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
Tel: +65 6874 3495
Email: cbaker@i2r.a-star.edu.sg

Dietrich Rebholz-Schuhmann, MD, PhD
Research Group Leader
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD,
United Kingdom
Tel: +44 (0)1223 492 594
Email: Rebholz@ebi.ac.uk