LexMapr: a rule-based text mining tool for ontology-driven harmonization of short biomedical specimen descriptions

Gurinder Gosal[1], Emma Griffiths[1], Damion Dooley[1], Ivan Gill[1], Dan Fornika[2], Heather Tate[3], Maria Sanchez[3], Ruth Timme[3], William Hsiao[1, 2]

[1]University of British Columbia, Canada, [2]BC Centre for Disease Control, Canada, [3]US Food & Drug Administration, USA

Introduction

- **LexMapr** is an **open-source**, **ontology-driven, rule-based, text-mining** system developed to harmonize pathogen sample metadata, often encoded as inconsistent free text short phrases.
- LexMapr combines basic lexicographic transformation with light Natural Language Processing and other functionality to standardized text to ontology terms.
- LexMapr subsequently performs ontology-driven classification of specimen data using third party classifications, initially with extended Interagency Food Safety Analytics Collaboration (**IFSAC+**) food categorization schema.
- LexMapr addresses many challenges in processing short textual biomedical data, such as specimen phrases (**Figure 1)**.

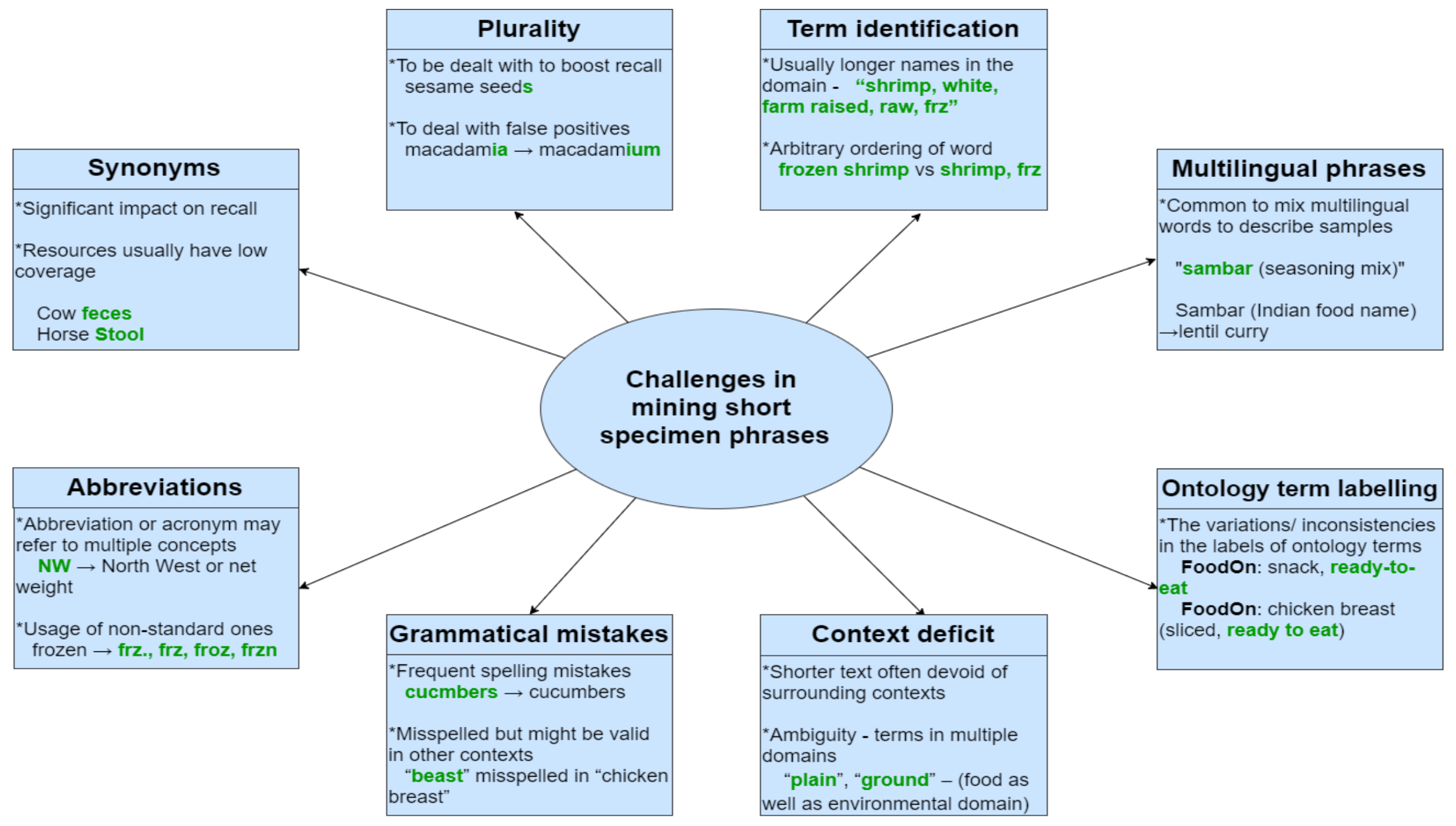
Challenges in mining short specimen phrases
Plurality
*To be dealt with to boost recall
sesame seeds
*To deal with false positives
macadamia → macadamium
Term identification
*Usually longer names in the domain - "shrimp, white, farm raised, raw, frz"
*Arbitrary ordering of word
frozen shrimp vs shrimp, frz
Synonyms
*Significant impact on recall
*Resources usually have low coverage
Cow feces
Horse Stool
Multilingual phrases
*Common to mix multilingual words to describe samples
"sambar (seasoning mix)"
Sambar (Indian food name) →lentil curry
Abbreviations
*Abbreviation or acronym may refer to multiple concepts
NW → North West or net weight
*Usage of non-standard ones
frozen → frz., frz, froz, frzn
Ontology term labelling
*The variations/ inconsistencies in the labels of ontology terms
FoodOn: snack, ready-to-eat
FoodOn: chicken breast (sliced, ready to eat)
Grammatical mistakes
*Frequent spelling mistakes
cucmbers → cucumbers
*Misspelled but might be valid in other contexts
"beast" misspelled in "chicken breast"
Context deficit
*Shorter text often devoid of surrounding contexts
*Ambiguity - terms in multiple domains
"plain", "ground" – (food as well as environmental domain)

Figure 1: Challenges in mining of short biosample phrases

LexMapr

- LexMapr has been developed as a rule-based system by using different manually developed rules along with a number of lexical resources in the form of locally created lookup tables. LexMapr architecture and pipeline are shown in **Figures 2 and 3**.

Resources used:

- **Ontologies**: FoodOn, GenEpiO, NCBITaxon, Uberon, CHEBI, Unit ontology –have been used for the underlying specimen domain (*users can dynamically select ontologies for other domains).
- **Lookup tables**: For Abbreviations (AbbLex), Spelling corrections (ScorLex), Non-English food names (NefLex) and Additional food synonyms (SynLex).
- **Dataset for rule development:** EnteroBase dataset that describes foodborne pathogen isolate source descriptions in short textual form (3391 unique sample descriptions extracted from >50000 samples) has been used for mining rules.

LexMapr availability

- LexMapr is an open source tool and the source code (written in Python language) has been made available at *https://github.com/Public-Health-Bioinformatics/LexMapr*
- LexMapr is currently available as a locally installable command-line tool and it has been released on bioconda with the latest version is available at: *https://anaconda.org/bioconda/lexmapr*
- For new users to LexMapr, a tutorial is available at: *https://docs.google.com/presentation/d/1RI1JIqjp8VcFbssd3OyAg3OxbCriKTrg6qxYpu9nYy4/edit#slide=id.p1*
- A user-friendly GUI is under development and will be made available very soon.

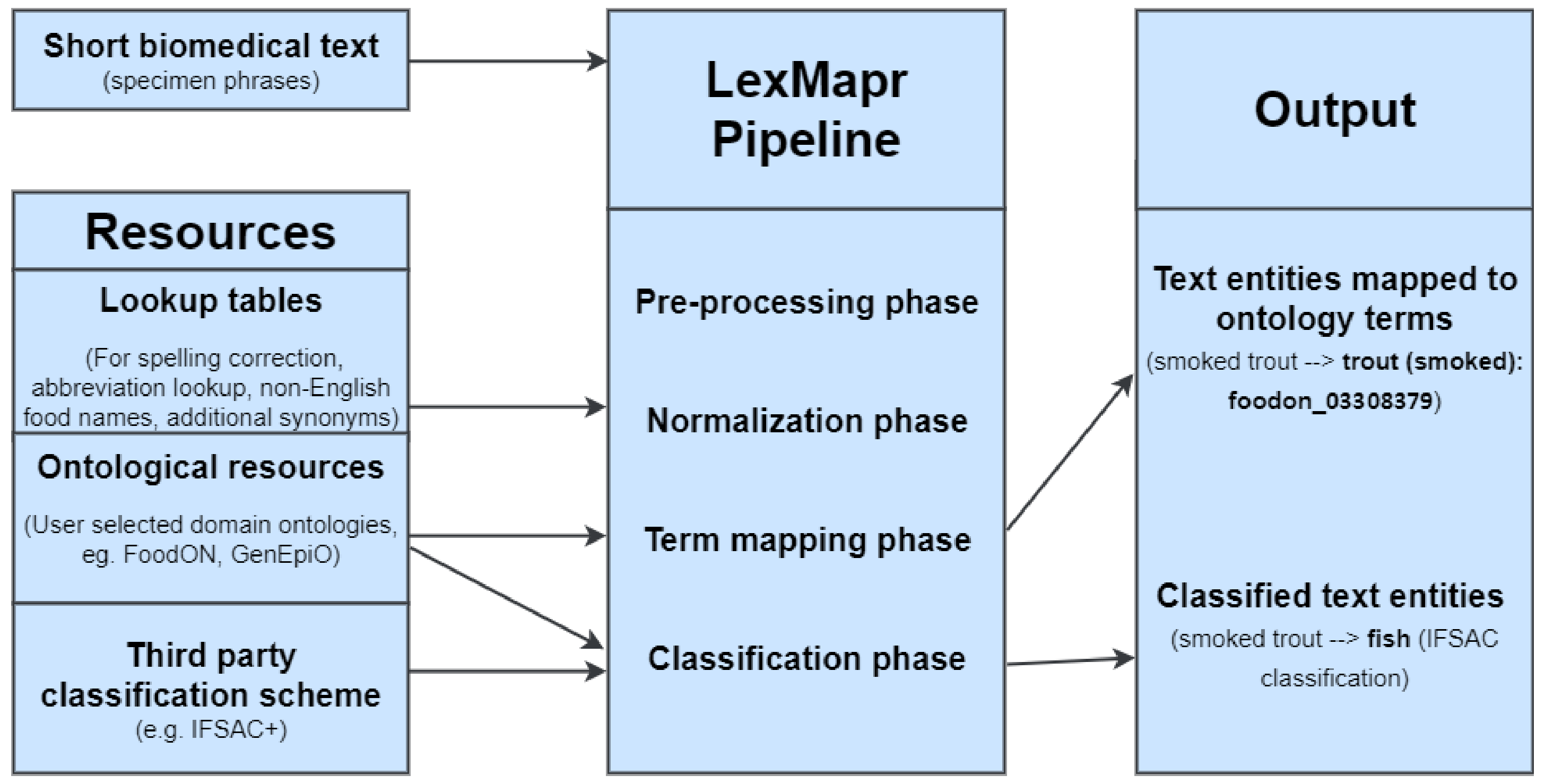
Short biomedical text
(specimen phrases)
Resources
Lookup tables
(For spelling correction, abbreviation lookup, non-English food names, additional synonyms)
Ontological resources
(User selected domain ontologies, eg. FoodON, GenEpiO)
Third party classification scheme
(e.g. IFSAC+)
LexMapr Pipeline
Pre-processing phase
Normalization phase
Term mapping phase
Classification phase
Output
Text entities mapped to ontology terms
(smoked trout --> trout (smoked): foodon_03308379)
Classified text entities
(smoked trout --> fish (IFSAC classification)

Figure 2: LexMapr architecture

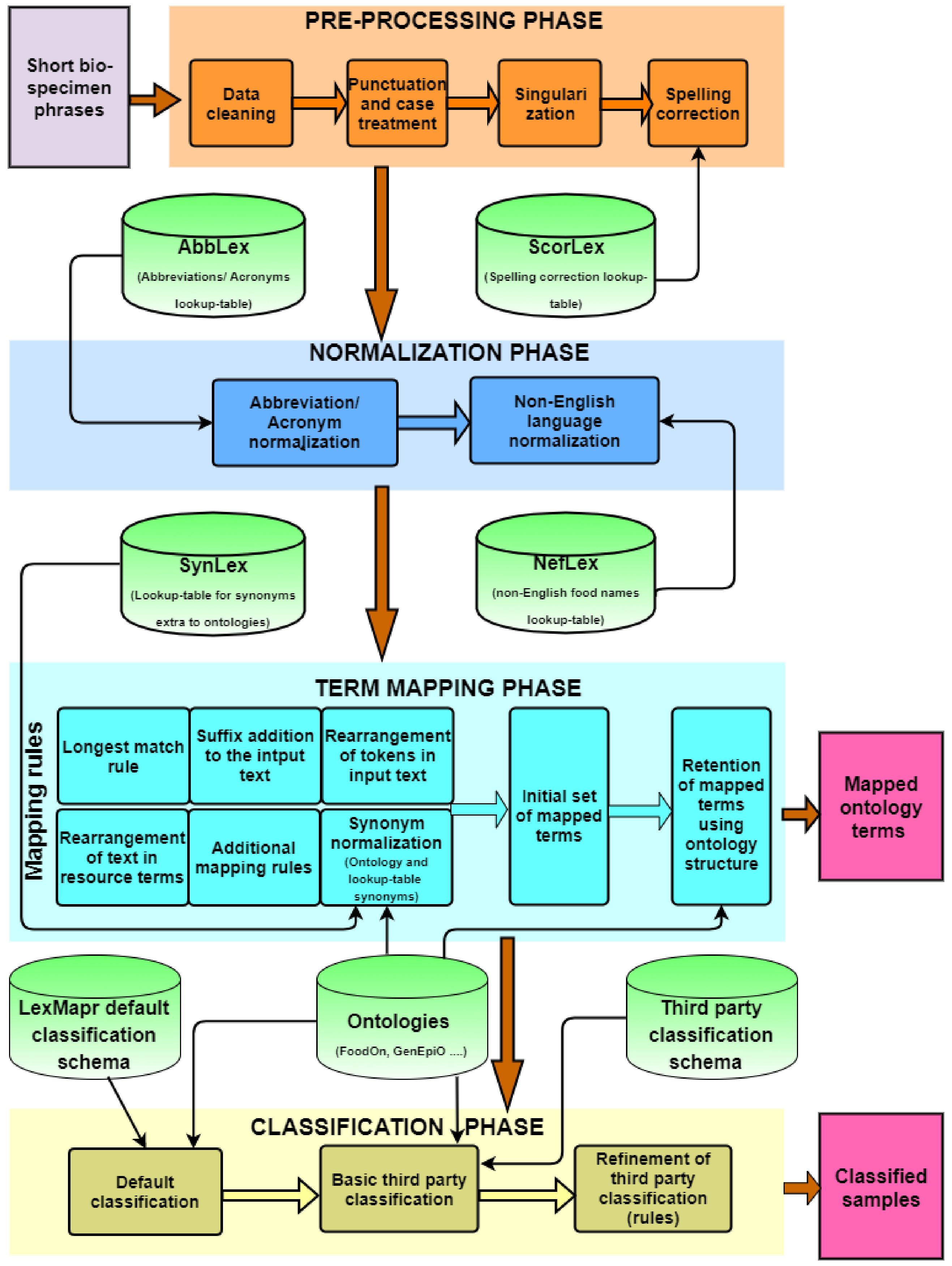
Short bio-specimen phrases
PRE-PROCESSING PHASE
Data cleaning
Punctuation and case treatment
Singulari zation
Spelling correction
AbbLex
(Abbreviations/ Acronyms lookup-table)
ScorLex
(Spelling correction lookup-table)
NORMALIZATION PHASE
Abbreviation/ Acronym normalization
Non-English language normalization
SynLex
(Lookup-table for synonyms extra to ontologies)
NefLex
(non-English food names lookup-table)
TERM MAPPING PHASE
Mapping rules
Longest match rule
Suffix addition to the intput text
Rearrangement of tokens in input text
Rearrangement of text in resource terms
Additional mapping rules
Synonym normalization
(Ontology and lookup-table synonyms)
Initial set of mapped terms
Retention of mapped terms using ontology structure
Mapped ontology terms
LexMapr default classification schema
Ontologies
(FoodOn, GenEpiO ....)
Third party classification schema
CLASSIFICATION PHASE
Default classification
Basic third party classification
Refinement of third party classification (rules)
Classified samples

Figure 3: LexMapr pipeline

Results

Table 1: A snapshot of term mapping and classification results

Sample description	Mapped ontology terms	Rule	IFSAC+ Classification
feces-bovine	bovine:foodon_03414374, feces:uberon_0001988	Punctuation Treatment	cow, clinical/research
beef, ground	ground beef food product:foodon_00001282		beef
Rodent (colon)	colon:uberon_0001155, rodentia:ncbitaxon_9989	Synonym Usage	clinical/research
fresh cheese curd	cheese curd:foodon_03310352, food (fresh):foodon_00002457		dairy
Walnuts	walnut:foodon_03316466	Inflection (Plural) Treatment	nuts
sesame seeds	sesame seed:foodon_03310306		seeds
tumeric powder	turmeric food product:foodon_00002323	Spelling Correction Treatment	herbs
cantelope	cantaloupe fruit food product:foodon_00001288		melon fruit
Frz Catfish	catfish:foodon_03412620, frozen:pato_0001985	Abbreviation normalization	fish
Snow Crab, froz	frozen:pato_0001985, snow crab:foodon_03411497		crustaceans
smoked trout	trout (smoked):foodon_03308379	Permutation of Tokens	fish
haldi	turmeric food product:foodon_00002323	Non English Usage Treatment	herbs
ground beef	ground beef food product:foodon_00001282	Suffix Addition	beef

Evaluation

LexMapr performance has been tested on foodborne pathogen sample data from two different surveillance systems - The **US FDA's GenomeTrakr** system and The **US National Antimicrobial Resistance Monitoring System (NARMS)**'s Resistome Tracker platform.

Term mapping evaluation

- **Evaluation dataset for term mapping**- 710 testing samples from **GenomeTrakr** that are completely independent of the previous training or testing samples.

*(Previously term coverage assessed at 89% (accuracy 95%) based on strict criteria for >2000 unique samples from Enterobase, GenomeTrakr and BC Public Health Laboratory data).

Table 2: LexMapr term mapping evaluation results based on strict criteria (does not count partial matches)

Measure criteria	No. of specimens	Correct match	Missing match	Pipeline recall	Spurious match	Accuracy (pipeline)	F-Measure (Fl)
Strict	710	632	45	93.35	33	95.04	94.18

Classification evaluation

- **Evaluation dataset for classification-** 500 unique sample descriptions from NARMS's Resistome Tracker for IFSAC+ classification

Table 3: LexMapr classification evaluation results

No. of specimens	Correct match	Missing match	Pipeline recall	Spurious match	Accuracy (pipeline)	F-Measure (Fl)
500	453	13	97.21	34	93.02	95.07

Acknowledgements

LexMapr project has been supported by funding from **USDA** (Agreement Number: 58-8040-8-014-F), **CIHR** (Award reference: PJT-159456), **Genome British Columbia / Genome Canada** (286GET).

- **Contact:** gurinder.gosal@bccdc.ca, william.hsiao@bccdc.ca