Lexical Substitution for the Medical Domain

Martin Riedl, Michael R. Glass and Alfio Gliozzo

Lexical Substitution

Given:
Sentence and target word
Goal:
Get substitutes and rank substitutions for target word that fit into the context
Evaluation:
Compare substitutions against gold standard (several measures exist)

Target Term

I play my ***bugle*** every day.

Instrument:1, flugelhorn:1, ajuga:0

#Annotators who marked as valid substitute

	SemEval Lexical Substitution	Medical Lexical Substitution
Multi-Word Expressions	Some	Many
Dataset	201 terms with 10 sentences for each	456 terms and 701 sentences total
Annotators	7	2

IBM T.J. Watson Research Center

Delexicalized Regression Approach

Dataset is positive/negative substitutions in context

in meningococcemia may result from acquired ***defects*** in the protein C pathway
[abnormality:1, derangement:1, tetralogy:0, body dysmorphic disorder:0, …]

Abs (absorption) should always be followed to confirm a positive ***RPR***.
[rapid plasma reagin:1, vdrl:0, serology:0, tpha:0, serologic test:0, …]

Binary Classification

Sentence: $w_1\ w_2\ \ldots\ t\ \ldots\ w_n$

target

Substitute candidates:
$s_1, s_2, \ldots, s_n$

Binary decision for each tuple:
<sentence, target, substitute>

Generate Features

Use only delexicalized features to get high generalization for unseen words
Features describe the relationship between the target and substitute or the substitute and the context
e.g. for abnormality:
Target word and substitute have the same POS tag:
SAME_POS:1.0
Do they share the same entry in UMLS:
UMLS_SAME: 1.0

Train a Logistic Regression Model

If any annotator marked a substitute as valid it is a positive example

Others are negative examples

Context Independent Features

Distributional Thesauri Features

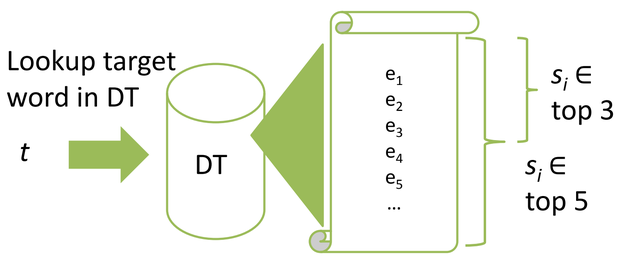
Built two thesauri from Medical corpus based on similarity of context distributions

Dependency Parse Context
tandem mass spectrometry
ms
mass spectrometer
hplc
mass spectrometric

Mass Spectrometry

Neighboring Word Context
spectrometry
technique
ultrasonography
Angiography
polymerase chain reaction

Feature: Is substitute contained in the top N entries of the target word DT (N=1,3,5,10,20,50)



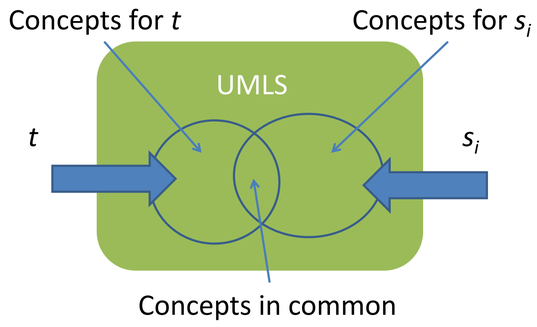
Medical Lexicon (UMLS)

Unified Medical Language System (UMLS) is a lexical resource like WordNet

Terms can be mapped to Concepts, like synsets

Features for number of concepts for each, number of shared concepts and binary empty-intersection feature

Map both target and substitute into UMLS



Context Dependent Features

N-Gram Features

Using Google Web 1T

freq(Ngram(substitute)) / freq(Ngram(target))
E.g.
Meningococcemia may result from acquired ***defects*** in the protein C pathway

s_i = abnormality

Ngram	Feature	Value
@	Ngram_0_0	freq(abnormality) / freq(defects)
acquired @	Ngram_1_0	freq(acquired abnormality) / freq(acquired defects)
…	…	…
from acquired @ in the	Ngram_2_2	freq(from acquired abnormality in the) / freq(from acquired defects in the)

Distributional Thesauri Features

Distributional thesauri contain associations between terms and contexts

Check overlap of the context in the sentence for both target and substitute:
E.g. using the ngram based Medline thesaurus:
abnormality acquired_@_in
defects acquired_@_in
If both exist in the database add a binary feature:
Medline_context_match: 1.0

Part-of-Speech

Characterize Context

POS tag of target word and substitute word

POS Tag Ngram (without POS from target word)
e.g.: DT NP VBZ JJ TO VB .

Results

Rank substitute candidates by system score.

Precision at *n* is percent correct in the first *n*.

Report on Mean Average Precision and Precision at 1

Compared to a baseline using only the distributional thesaurus context independent

Significant improvement ($p < 0.01$)

Ablation study shows strong impact for Distributional Thesaurus and UMLS

System	MAP	P@1
Baseline	0.6408	0.5365
All Features	**0.7048**	**0.6366**
w/o DT	0.5798	0.4835
w/o UMLS	0.6618	0.5651
w/o Ngrams	0.7009	0.6252
w/o POS	0.7027	0.6323

Error Analysis

The most common cause of thrombocytopenia during pregnancy is gestational thrombocytopenia, which is a mild thrombocytopenia with platelet levels remaining greater than 70,000/mL.

t = mild thrombocytopenia
s = severe thrombocytopenia

Antonym is most obvious error class