The SPECIALIST NLP Tools

Dr. Chris J. Lu

The Lexical Systems Group

NLM. LHNCBC. CGSB

June, 2010
Table of Contents

• Introduction
• Lexical Tools
  ▪ Lvg
  ▪ Norm
• Text Categorization Tools
• Questions
Introduction
Introduction
Introduction - LB

LexBuild

Text Processing

Lexical Tools

SPECIALIST LEXICON
Introduction - Lexicon
Introduction - LC

Text Processing

Lexical Tools

SPECIALIST LEXICON
Introduction - LA
Introduction - Numbers
Introduction - SCRT
Introduction – Lexical Tools
Introduction – Text Tools
Introduction - GSpell
Introduction - dTagger
Introduction - VTT
Introduction - TC
Introduction – NLP Tools

The SPECIALIST LEXICON

Lexical Tools

Text Tools

NLP Projects
Introduction – NLP Tools

The SPECIALIST LEXICON

Lexical Tools

Text Tools

LexBuild

Ten 3

Numbers

LexCheck

LexAccess

Lexical Tool

Text Tools

NLP Projects

Software Development
Introduction – NLP Tools

The SPECIALIST LEXICON

Lexical Tools

Text Tools

NLP Projects

Software Development
Lexical Tools
Lexical Tools

- A suite of text utilities
Lexical Tools

- A suite of text utilities take the given input
Lexical Tools

- A suite of text utilities that generate, mutate, and filter out lexical variants from the given input.
Four Tools

Input

Lvg
Norm
LuiNorm
WordIndex

Output...
Output.3
Output.2
Output.1
Tool Types

• Command line tools
  – lvg (Lexical Variants Generation)
  – norm
  – luiNorm
  – wordInd

• Lexical Gui Tool (lgt)

• Web Tools

• Java API’s
Functions

- Used in nature language processing for
  - aggressive text pattern matching
  - creating normalized and expanded terms
  - making word, term, phrase indexes
  - matching queries with indexed entries
  - increasing recall and/or precision
Facts

• Release annually
• Free distributed with open source code
• 100% Java (since 2002)
• Run on different platforms
• One complete package
• Documents & supports
LVG - 2010

- 62 flow components
- 37 options
  - input filter options (3)
  - global behavior options (13)
  - flow specific options (2)
  - output filter options (19)
Flow Components

leave : inflect : leaves
leave : leaving : left
<table>
<thead>
<tr>
<th>Command Line Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&gt; lvg -f:i</code></td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
<tr>
<td>leave</td>
</tr>
</tbody>
</table>
Fielded Output

> lvg -f:i
leave

```
> lvg -f:i
leave

![](diagram)
```
A Serial Flow

- Flow components can be arranged so that the output of one is the input to another.
A Serial Flow - Example

>` lvg -f:l:q:g:t:p:w

The Gougerot-Sjögren's Syndrome
The Gougerot-Sjögren's Syndrome

  gougerotsjogren syndrome|2047|

  16777215|l+q+g+t+p+w|1|
Parallel Flows

- Multiple flows can be defined

Input term

noOperation

Output term

Uninflect

Output terms

synonyms
### Parallel Flows - Example

```
> lvg -f:n -f:B:y

<table>
<thead>
<tr>
<th>ear</th>
<th>ear</th>
<th>ear</th>
<th>2047</th>
<th>1048575</th>
<th>n</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ear</td>
<td>aural</td>
<td>1</td>
<td>1</td>
<td>B+y</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ear</td>
<td>auricularis</td>
<td>1</td>
<td>1</td>
<td>B+y</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ear</td>
<td>otic</td>
<td>1</td>
<td>1</td>
<td>B+y</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>ear</td>
<td>otor</td>
<td>1</td>
<td>1</td>
<td>B+y</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
```
Input Filter Options

Input term \[ \text{lvg } -f:u -t:7 -F:8:6 \]

Output terms

Take field 7 from the input

C0035440|ENG|S|L0035434|VW|S0003894|

Rheumatic carditis, acute

acute Rheumatic carditis|S0003894
Global Behavior Options

Input term

Output terms
Output terms

> lvg -f:L -f:E -s:”\”

otitis

otitis\otitis\128\513\L\1

otitis\E0044452\128\513\E\2

Change separator to “\”
Output Filter Options

Input term \[ \text{lvg -f:L} \]

Output terms

hot

hot|hot|<adj+verb>|<base+positive+infinitive+pres1p23p>|L|1|

Show the category and inflection names
Norm

• Composed of 11 Lvg flow components to abstract away from:
  – case
  – punctuation
  – possessive forms
  – inflections
  – spelling variants
  – stop words
  – Diacritics, ligatures & symbols (Unicode to ASCII)
  – word order
| q0: map Unicode symbols to ASCII |
| g: remove genitives |
| rs: remove parenthetic plural forms |
| o: replace punctuation with spaces |
| t: strip stop words |
| l: lowercase |
| B: uninflect each words in a term |
| Ct: retrieve citations |
| q7: Unicode core Norm |
| q8: strip or map non-ASCII char |
| w: sort words by order |
Hodgkin's Diseases, NOS

<table>
<thead>
<tr>
<th>q0: map Unicode symbols to ASCII</th>
</tr>
</thead>
<tbody>
<tr>
<td>g: remove genitives</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
</tr>
<tr>
<td>t: strip stop words</td>
</tr>
<tr>
<td>l: lowercase</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
</tr>
<tr>
<td>q8: strip or map non-ASCII char</td>
</tr>
<tr>
<td>w: sort words by order</td>
</tr>
<tr>
<td><strong>q0</strong>: map Unicode symbols to ASCII</td>
</tr>
<tr>
<td>--------------------------------------</td>
</tr>
<tr>
<td><strong>g</strong>: remove genitives</td>
</tr>
<tr>
<td><strong>rs</strong>: remove parenthetic plural forms</td>
</tr>
<tr>
<td><strong>o</strong>: replace punctuation with spaces</td>
</tr>
<tr>
<td><strong>t</strong>: strip stop words</td>
</tr>
<tr>
<td><strong>l</strong>: lowercase</td>
</tr>
<tr>
<td><strong>B</strong>: uninflect each words in a term</td>
</tr>
<tr>
<td><strong>Ct</strong>: retrieve citations</td>
</tr>
<tr>
<td><strong>q7</strong>: Unicode core Norm</td>
</tr>
<tr>
<td><strong>q8</strong>: strip or map non-ASCII char</td>
</tr>
<tr>
<td><strong>w</strong>: sort words by order</td>
</tr>
</tbody>
</table>

**Norm**

Hodgkin's Diseases, NOS
<table>
<thead>
<tr>
<th>g: remove genitives</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs: remove parenthetic plural forms</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
</tr>
<tr>
<td>t: strip stop words</td>
</tr>
<tr>
<td>l: lowercase</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
</tr>
<tr>
<td>q8: strip or map non-ASCII char</td>
</tr>
<tr>
<td>w: sort words by order</td>
</tr>
<tr>
<td>Short Code</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>q0</td>
</tr>
<tr>
<td>g</td>
</tr>
<tr>
<td>rs</td>
</tr>
<tr>
<td>o</td>
</tr>
<tr>
<td>t</td>
</tr>
<tr>
<td>l</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>Ct</td>
</tr>
<tr>
<td>q7</td>
</tr>
<tr>
<td>q8</td>
</tr>
<tr>
<td>w</td>
</tr>
</tbody>
</table>

**Example:**

<table>
<thead>
<tr>
<th>Input</th>
<th>Processed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hodgkin's Diseases, NOS</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>Normalize (Norm)</td>
<td>Function</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td><strong>g</strong></td>
<td>remove genitives</td>
</tr>
<tr>
<td><strong>t</strong></td>
<td>strip stop words</td>
</tr>
<tr>
<td><strong>l</strong></td>
<td>lowercase</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td>uninflect each words in a term</td>
</tr>
<tr>
<td><strong>Ct</strong></td>
<td>retrieve citations</td>
</tr>
<tr>
<td><strong>q0</strong></td>
<td>map Unicode symbols to ASCII</td>
</tr>
<tr>
<td><strong>q7</strong></td>
<td>Unicode core Norm</td>
</tr>
<tr>
<td><strong>q8</strong></td>
<td>strip or map non-ASCII char</td>
</tr>
<tr>
<td><strong>w</strong></td>
<td>sort words by order</td>
</tr>
<tr>
<td>Norm</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>q0: map Unicode symbols to ASCII</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>g: remove genitives</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Hodgkin Diseases NOS</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>Hodgkin Diseases</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>B: uninflect each words in a term</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>q8: strip or map non-ASCII char</td>
<td>Hodgkin Diseases</td>
</tr>
<tr>
<td>w: sort words by order</td>
<td>Hodgkin Diseases</td>
</tr>
<tr>
<td>q0: map Unicode symbols to ASCII</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>g: remove genitives</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>rs: remove parenthetic plural forms</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>o: replace punctuation with spaces</td>
<td>Hodgkin Diseases NOS</td>
</tr>
<tr>
<td>t: strip stop words</td>
<td>Hodgkin Diseases</td>
</tr>
<tr>
<td>l: lowercase</td>
<td>hodgkin diseases</td>
</tr>
<tr>
<td>B: uninfl ect each words in a term</td>
<td></td>
</tr>
<tr>
<td>Ct: retrieve citations</td>
<td></td>
</tr>
<tr>
<td>q7: Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8: strip or map non-ASCII char</td>
<td></td>
</tr>
<tr>
<td>w: sort words by order</td>
<td></td>
</tr>
</tbody>
</table>
### Norm

<table>
<thead>
<tr>
<th>Operation</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>q0:</td>
<td>map Unicode symbols to ASCII</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>g:</td>
<td>remove genitives</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>rs:</td>
<td>remove parenthetic plural forms</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>o:</td>
<td>replace punctuation with spaces</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>t:</td>
<td>strip stop words</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>l:</td>
<td>lowercase</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>B:</td>
<td>uninflect each word in a term</td>
<td>hodgkin disease</td>
</tr>
<tr>
<td>Ct:</td>
<td>retrieve citations</td>
<td></td>
</tr>
<tr>
<td>q7:</td>
<td>Unicode core Norm</td>
<td></td>
</tr>
<tr>
<td>q8:</td>
<td>strip or map non-ASCII char</td>
<td></td>
</tr>
<tr>
<td>w:</td>
<td>sort words by order</td>
<td></td>
</tr>
<tr>
<td>Process</td>
<td>Description</td>
<td>Example</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>q0</td>
<td>map Unicode symbols to ASCII</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>g</td>
<td>remove genitives</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>rs</td>
<td>remove parenthetic plural forms</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>o</td>
<td>replace punctuation with spaces</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>t</td>
<td>strip stop words</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>l</td>
<td>lowercase</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>B</td>
<td>uninflect each words in a term</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>Ct</td>
<td>retrieve citations</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>q7</td>
<td>Unicode core Norm</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>q8</td>
<td>strip or map non-ASCII char</td>
<td>hodgkin disease</td>
</tr>
<tr>
<td>w</td>
<td>sort words by order</td>
<td>hodgkin disease</td>
</tr>
</tbody>
</table>
q0: map Unicode symbols to ASCII
Hodgkin's Diseases, NOS
Hodgkin Diseases, NOS
Hodgkin Diseases,
Hodgkin Diseases NOS
Hodgkin Diseases
hodgkin diseases
hodgkin disease
hodgkin disease

q7: Unicode core Norm

q8: strip or map non-ASCII char

w: sort words by order
q0: map Unicode symbols to ASCII

q7: Unicode core Norm

q8: strip or map non-ASCII char

Hodgkin Diseases, NOS

Hodgkin's Diseases, NOS

Hodgkin Diseases, NOS

Hodgkin Diseases NOS

Hodgkin Diseases

Hodgkin diseases

hodgkin disease

hodgkin disease

hodgkin disease

hodgkin disease

hodgkin disease
<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>q0</td>
<td>map Unicode symbols to ASCII</td>
<td>Hodgkin's Diseases, NOS</td>
</tr>
<tr>
<td>q7</td>
<td>Unicode core Norm</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>g</td>
<td>remove genitives</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>rs</td>
<td>remove parenthetic plural forms</td>
<td>Hodgkin Diseases, NOS</td>
</tr>
<tr>
<td>o</td>
<td>replace punctuation with spaces</td>
<td>Hodgkin Diseases NOS</td>
</tr>
<tr>
<td>t</td>
<td>strip stop words</td>
<td>Hodgkin Diseases</td>
</tr>
<tr>
<td>l</td>
<td>lowercase</td>
<td>Hodgkin diseases</td>
</tr>
<tr>
<td>B</td>
<td>uninflect each words in a term</td>
<td>Hodgkin disease</td>
</tr>
<tr>
<td>Ct</td>
<td>retrieve citations</td>
<td>Hodgkin disease</td>
</tr>
<tr>
<td>w</td>
<td>sort words by order</td>
<td>disease hodgkin</td>
</tr>
</tbody>
</table>
Norm: Example

- Hodgkin Disease
- HODGKINS DISEASE
- Hodgkin's Disease
- Disease, Hodgkin's
- HODGKIN'S DISEASE
- Hodgkin's disease
- Hodgkins Disease
- Hodgkin's disease NOS
- Hodgkin's disease, NOS
- Disease, Hodgkins
- Diseases, Hodgkins
- Hodgkins Diseases
- Hodgkins disease
- hodgkin's disease
- Disease;Hodgkins
- Disease, Hodgkin
- ...

Disease hodgkin
Example - Norm

Hodgkin’s Disease

Query
Example - Norm

Hodgkin’s Disease

Query → norm → Normed term → disease hodgkin
Example - Norm

Hodgkin’s Disease

Query → norm → Normed term → disease hodgkin

Indexed Database

Normalized String

SQL
Example - Norm

Hodgkin’s Disease

Query → **norm** → Normed term → disease hodgkin

Indexed Database Normalized String

Results that matches the normalized query
Example 2 – UMLS Metathesaurus

Metathesaurus
English Strings

→ norm

Normalized string index
MRXNS.ENG

→ WordInd

Normalized word index
MRXNW.ENG
Example 2

Query → norm → Normalized string index

→ Normalized term → Normalized word index

→ Metathesaurus Concepts

SUIS
Example 2 – String Name

Normed term

MRXNS.ENG

SUIS

ENG|disease hodgkin|C0019829|L0019829|S0006131|
ENG|disease hodgkin|C0019829|L0019829|S0033754|
ENG|disease hodgkin|C0019829|L0019829|S0048925|
ENG|disease hodgkin|C0019829|L0019829|S0048926|
ENG|disease hodgkin|C0019829|L0019829|S0220234|
ENG|disease hodgkin|C0019829|L0019829|S0376583|
ENG|disease hodgkin|C0019829|L0019829|S0378160|

……..
Example 2 – String Name
Example 2 – Concept Name

```
| ENG | disease hodgkin | C0019829 | L0019829 | S0006131 |
| ENG | disease hodgkin | C0019829 | L0019829 | S0033754 |
| ENG | disease hodgkin | C0019829 | L0019829 | S0048925 |
| ENG | disease hodgkin | C0019829 | L0019829 | S0048926 |
| ENG | disease hodgkin | C0019829 | L0019829 | S0220234 |
| ENG | disease hodgkin | C0019829 | L0019829 | S0376583 |
| ENG | disease hodgkin | C0019829 | L0019829 | S0378160 |
```

……..
Example 2

Metathesaurus concepts that match the normalized query
Text Categorization Tools
TC Tools

• Based on Journal Descriptor Indexing (JDI) methodology (by Susanne Humphrey)
• Uses a small set of high level descriptors:
  ▪ Journal Descriptors (JDs)
  ▪ Semantic Types (STs)
• Used for categorizing text, indexing contents, retrieving records, and word sense disambiguation (WSD)
Facts for TC Tools

• Release annually (since 2007)
• Free distributed with open source code
• 100% Java
• Run on different platforms
• One complete package
• Documents & supports
• Provides Java APIs, command line tools, GUI tools, and Web tools
TC Tools

• Two types of categorization:
  ▪ Journal Descriptor Indexing (JDI): categorizes text according to Journal Descriptors (JDs)
  ▪ Semantic Type Indexing (STI): categorizes text according to Semantic Types (STs)

• St WSD tool (2009)
Journal Descriptors (JDs)?

• Set of 122 MeSH descriptors representing high-level categories, mostly biomedical disciplines.

• Used for indexing journals *per se*

• Assigned by human indexer to the 4100 journals

• Source is from: List of Serials for Online Users file (lsi.xml)
Journal Descriptors

• Examples of JD from lsi.xml
  ▪ JID - 03132144
    TA - Transplantation (the journal *Transplantation*)
    JD - Transplantation
  
  ▪ JID - 9802574
    TA - Pediatr Transplant
      (the journal *Pediatric Transplantation*)
    JD - Pediatrics; Transplantation
  
  ▪ JID - 0052631
    TA - J Pediatr Surg (the *Journal of Pediatric Surgery*)
    JD - Pediatrics; Surgery
JDI Methodology

• Training set is about 3.4 million MEDLINE documents (3 years)

• JDI uses statistical associations between words in MEDLINE training set record TI/AB and the JD/s corresponding to the journal in the training set record

• But
  ▪ JDs are not in a MEDLINE record
  ▪ JDs are in the NLM serial record from lsi2007.xml
JDI – Link to JDs

• Example of link between MEDLINE records and JDs

  ▪ Training set MEDLINE record:
    PMID - 10919582
    TI - Combined liver and kidney transplantation in children.
    JID - 0132144

  ▪ Transplantation serial record:
    JID - 0132144
    JD - Transplantation
JDI – Link to JDs

• Example of Training set MEDLINE record with “imported” JD Transplantation:

  - PMID - 10919582
  - TI - Combined liver and kidney transplantation in children.
  - JD - Transplantation
JDI - JD Score (Word)

• JDI of the word “transplantation”

1|0.275691|Transplantation
2|0.070315|Hematology
3|0.044303|Nephrology
4|0.031517|Pulmonary Disease (Specialty)
5|0.029425|Gastroenterology

• Transplantation score

\[
\frac{\text{no. of docs in training set in which TI/AB word transplantation co-occurs with JD Transplantation}}{\text{no. of docs in training set in which the word transplantation occurs in TI/AB}} = 0.275691
\]
JDI - JD Score (Word)

• JDI of the word “kidney”

1|0.140088|Nephrology
2|0.080848|Transplantation
3|0.057162|Urology
4|0.032341|Toxicology
5|0.024398|Pharmacology

• Nephrology score

\[
\text{no. of docs in training set in which TI/AB word kidney co-occurs with JD Nephrology} = \frac{\text{no. of docs in training set in which the word kidney occurs in TI/AB}}{\text{no. of docs in training set in which the word kidney occurs in TI/AB}}
\]

= 0.140088
JDI - JD Score (Phrase)

• JDI of the phrase “kidney transplantation”

1|0.178269|Transplantation
2|0.092195|Nephrology
3|0.037875|Hematology
4|0.034381|Urology
5|0.017438|Gastroenterology

• Score for Transplantation is average of Transplantation score for word kidney and Transplantation score for word transplantation

• A JD score for a phrase is the average of that JD’s score across the words in the phrase
STI - Semantic Types

• What are Semantic Types (STs)?

• Set of 135 semantic types in the Semantic Network in NLM’s Unified Medical Language System (UMLS).

• Concepts in the UMLS Metathesaurus are assigned one or more STs which semantically characterize those concepts.

• For example, “aspirin” is assigned the STs Pharmacologic Substance (phsu) and Organic Chemical (orch).
Semantic Type Indexing (STI)

- JDI has word-JD vectors representing JD indexing of each of the 304,000 words in the training set.

- STI also has word-ST vectors representing ST indexing of each training set word.

- Thus, STI of text can be performed exactly the same way as JDI of text. An ST score for a text is the average of that ST’s score for words in the text. The scores for all the STs comprise the ST vector for the text.
TC – St WSD

- Words Senses disambiguation (WSD)

Free Text → MetaMap (MMTX) → Metathesaurus Concept
TC – St WSD

- Words Senses disambiguation (WSD)
TC – St WSD

- Words Senses disambiguation (WSD)

Free Text → MetaMap (MMTX) → Concept 1 → Concept 2 → Concept n → TC → Best Concept
Example – St WSD

• “transport” is ambiguous:
  ▪ Biological Transport (ST is Cell Function, cell)
  ▪ Patient Transport (ST is Health Care Activity, hlca)

• STI of text results in ranked list of STs.
  ▪ If cell ranks higher than hlca, then meaning is Biological Transport.
  ▪ If hlca ranks higher than cell, then meaning is Patient Transport.
STI of PMID 9674486 in WSD collection

Input: Preliminary results of bedside inferior vena cava filter placement: safe and cost-effective. The use of inferior vena cava filters (IVCFs) is increasing in patients at high risk for venous thromboembolism; however, there is considerable controversy related to their cost. We inserted eight percutaneous IVCFs at the bedside. The hospital charges for bedside IVCF insertion were substantially lower compared with those for IVCF insertion performed in the Radiology Department or operating room. There was one death (unrelated to the procedure) and one asymptomatic caval occlusion believed to be caused by thrombus trapping. Bedside IVCF insertion is safe and cost-effective in selected patients. This practice averts the potential complications associated with transporting critically ill patients.

--- ST scores and rank based on document count for word ---

<table>
<thead>
<tr>
<th>Score</th>
<th>Rank</th>
<th>STI</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>0.4897</td>
<td>hlca</td>
<td>Health Care Activity</td>
</tr>
<tr>
<td>46</td>
<td>0.4086</td>
<td>celf</td>
<td>Cell Function (Biological Transport)</td>
</tr>
</tbody>
</table>
TC – St WSD

- Words Senses disambiguation (WSD)

...... transport...

MetaMap (MMTX)

Patient Transport (ST: Health Care Activity)

Biological Transport (ST: Cell Function)

TC

Best Concept
• Three methods for contexts of the ambiguity:
  ● ambig-sentence - sentence with ambiguity
  ● ambig-sentences - all sentences with ambiguity
  ● doc - entire MEDLINE document

• Three score systems:
  ● DC: document count
  ● WC: word count
  ● CS: combines score
TC – St WSD

- Published research on STI as a tool for word sense disambiguation (WSD) in natural language processing (NLP) using UMLS Metathesaurus, disambiguating 45 ambiguous strings from NLM’s WSD collection.

- Best unsupervised WSD methods
  - 2007: 75.39%
  - 2008: 75.00%
  - 2009: 77.37%
  - 2010: 77.36%

- First release in 2009.
Questions