

Text Crunching for the real Web

OBJECTIVE: To assess the role of smoking on low birth weight (LBW). STUDY DESIGN: From Massachusetts for 1998, 79,904 birth certificates were reviewed. Birth weight, gestational age, plurality and maternal race were analyzed in relation to the mother's smoking status and pregnancy. The etiologic fraction (EF) was calculated for smoking and LBW for the group as a whole as well as for various subgroups. RESULTS: A total of 11.7% of women acknowledged smoking during pregnancy. The overall LBW rate was 6.83%. The relative risk (RR) of LBW among smokers was 1.58. For all birth outcomes, the overall LBW rate was 6.83% (95% CI: 5.4-7.3). For singleton pregnancies it was 10.9% (95% CI: 9.6-12.1) (14% for singleton whites and 10.2% for singleton blacks). At term, the EF of smoking on LBW was 13.4% (95% CI: 11.5-15.3), with an EF of 16.1% (95% CI: 14.5-18.7) for term singletons (21.4% among whites and 14.6% among blacks). Among very LBW infants, smoking accounted for 1.7% (95% CI: -0.5-3.8) of the outcome (5.8% among singletons). When stratifying for the effect of smoking, the rate of LBW was 6.38% among nonsmokers, 9.5% (RR 1.46, 1.36-1.61) among light smokers, 11.67% (RR 1.82, 1.63-2.05) among moderate smokers and 11.73% (RR 1.87, 1.32-2.57) among heavy smokers. Sixty percent of the overall population effect of smoking on LBW was in the category of light smokers. CONCLUSION: The amount of LBW attributable to smoking was 6.4% in this sample. Among those who smoked, LBW was 58% more among smokers, and 60% of the overall population effect of smoking on LBW was noted among li

Harald Kirsch

European Bioinformatics Institute
Hinxton, Cambridge (UK)

NOE Semantic Mining

Summer School, Tihany (Hungary)
June 2005



Semantic Web vs. “old style” Web

*The Semantic Web is an extension of the current web in which **information is given well-defined meaning**, ...¹*

¹Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*, Scientific American, May 2001

Semantic Web vs. “old style” Web

*The Semantic Web is an extension of the current web in which **information is given well-defined meaning**, ...¹*

- $> 8 \cdot 10^9$ pages indexed by Google
- $> 15 \cdot 10^6$ abstracts in MEDLINE
- $> 11 \cdot 10^3$ full papers in PUBMED CENTRAL

¹Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*, Scientific American, May 2001

Semantic Web vs. “old style” Web

*The Semantic Web is an extension of the current web in which **information is given well-defined meaning**, ...¹*

- $> 8 \cdot 10^9$ pages indexed by Google
- $> 15 \cdot 10^6$ abstracts in MEDLINE
- $> 11 \cdot 10^3$ full papers in PUBMED CENTRAL

How can we assign **well-defined meaning** to all that?

¹Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*, Scientific American, May 2001

Strategies / Tasks

How can we assign **well-defined meaning** to all that?

1. concentrate on biomedical publications

Strategies / Tasks

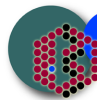
How can we assign **well-defined meaning** to all that?

1. concentrate on biomedical publications
2. automatically annotate available content for
 - curators, users
 - on-the-fly annotation upon request
 - *nasty* technical details of input formats

Strategies / Tasks

How can we assign **well-defined meaning** to all that?

1. concentrate on biomedical publications
2. automatically annotate available content for
 - curators, users
 - on-the-fly annotation upon request
 - *nasty* technical details of input formats
3. proactively annotate content
 - publishers, authors
 - integrated into publishing process
 - *nice* XML input format



Semantic What?

What is **well-defined meaning** (semantic) anyway?

²B. Mons: *Which gene did you mean?*, BMC Bioinformatics, 2005, 6:142

Semantic What?

What is **well-defined meaning** (semantic) anyway?

1. mentions of real world objects: **named entities**
 - detect** — find the terms
 - identify** — disambiguate and relate to an ontology
2. **relations** between those objects

²B. Mons: *Which gene did you mean?*, BMC Bioinformatics, 2005, 6:142

Semantic What?

What is **well-defined meaning** (semantic) anyway?

1. mentions of real world objects: **named entities**
 - detect** — find the terms
 - identify** — disambiguate and relate to an ontology
2. **relations** between those objects

To keep the challenge manageable, I will restrain [...] to term identification ...²

²B. Mons: *Which gene did you mean?*, BMC Bioinformatics, 2005, 6:142

Term Identification

Named entity recognition

1. trinitrophenylaminocaproyldipalmitoylphosphatidylethanolamine
2. *Lambertia Ericifolia*
3. Aterox P17

Term Identification

Named entity recognition

1. trinitrophenylaminoc
2. Lambertia Ericifolia
3. Aterox P17



phosphatidylethanolamine

Identification

Term Identification

Named entity recognition

1. trinitrophenylaminoc...phosphatidylethanolamine
2. *Lambertia Ericifolia*
3. Atperox P17



Identification

1. ?
2. TaxID 188621
3. UNIPROT Q9SJZ2



Term Identification — two methods

- “classic” Named Entity Recognition
 1. recognition
based on term and context features
 2. (assignment to a semantic identifier)

Term Identification — two methods

- “classic” Named Entity Recognition
 1. recognition
based on term and context features
 2. (assignment to a semantic identifier)
- lexicon based approach

Lexicon

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Term Identification — two methods

- “classic” Named Entity Recognition
 1. recognition
based on term and context features
 2. (assignment to a semantic identifier)
- lexicon based approach
 1. recognition implies assignment
lookup
 2. disambiguation

Lexicon

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Which Lexicon(s)?

Examples of publicly available sources:

source	terms
Entrez Taxonomy	210 000
Swissprot	200 000
Gene Ontology	20 000
drug names (MedlinePlus)	8 000
	438 000

ontology \neq lexicon \neq database

- lexicon
 - words, terms and their behaviour in text
 - denote object (database) or concept (ontology)
- database
 - objects and their features
 - “leaf nodes” of ontologies
- ontology
 - concepts and their relations

No comprehensive lexical resource available.
We take what we can get.

Spotting all these Terms

Two steps

1. tokenization
2. lookup

... mass of gymnin was ...

terms table

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Spotting all these Terms

Two steps

1. tokenization
2. lookup

... **mass** of gymnin was ...

terms table

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Spotting all these Terms

Two steps

1. tokenization
2. lookup

... mass of gymnin was ...

terms table

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Spotting all these Terms

Two steps

1. tokenization
2. lookup

... mass of gymnin was ...

terms table

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Spotting all these Terms

Two steps

1. tokenization
2. lookup

... mass of gymnin was ...

terms table

<i>term</i>	<i>ID</i>
...	
[Gg]ymnin	P84200
...	

Spotting all these Terms

Two steps

1. tokenization
2. lookup

... mass of <P acc="P84200">gymnin</P> was ...

terms table

<i>term</i>	<i>ID</i>
...	
[Gg]gymnin	P84200
...	

Terms vs. Words

Multi Word Terms

- Bile salt export pump (UniProt: O95342)
- *Buteo rufinus cirtensis* (TaxID: 115231)
- transcription export complex (GO:0000346)
- Theramycin Z (MedlinePlus: 202221)

Annotating Text

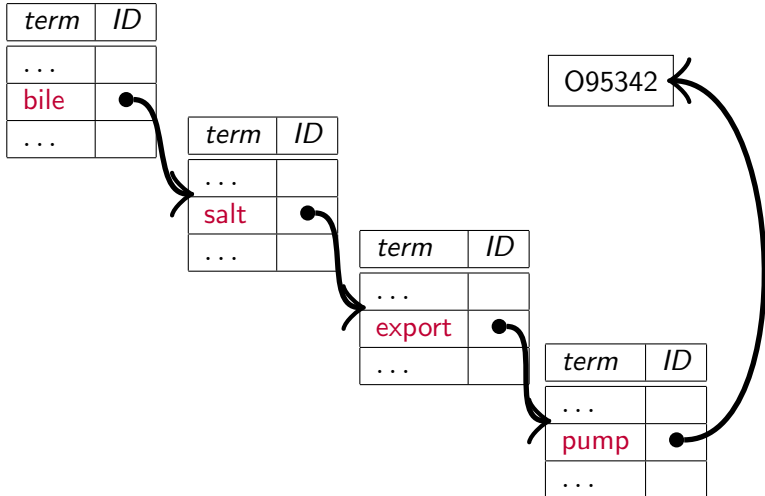
Mutations of the **bile** salt export pump ...

lookup won't work

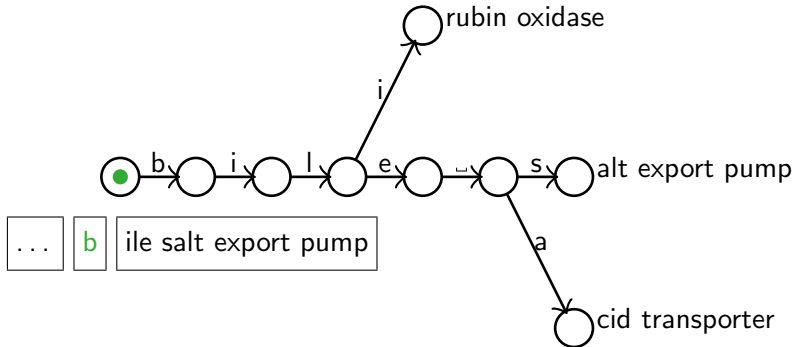
terms table

<i>term</i>	<i>ID</i>
...	
bile salt export pump	O95342
...	

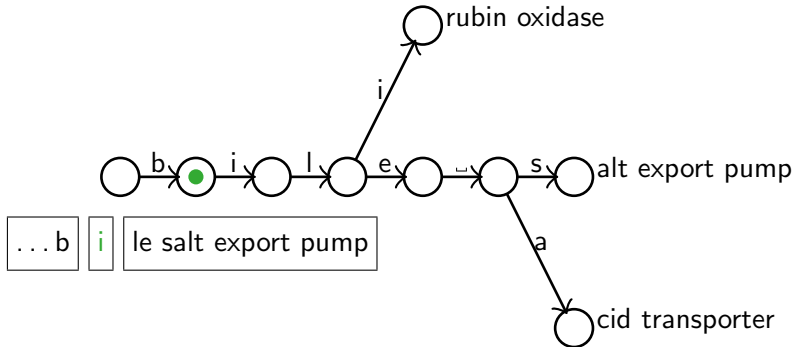
Cascaded Lookup



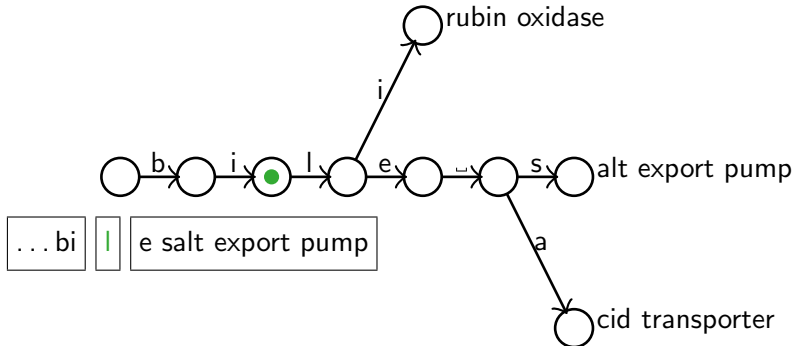
Cascaded Lookup, Character Level



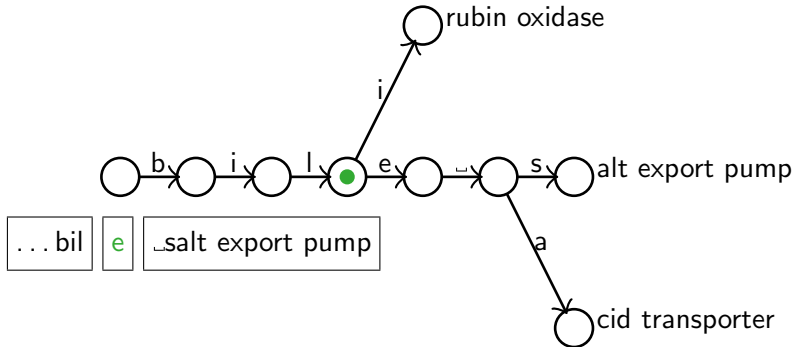
Cascaded Lookup, Character Level



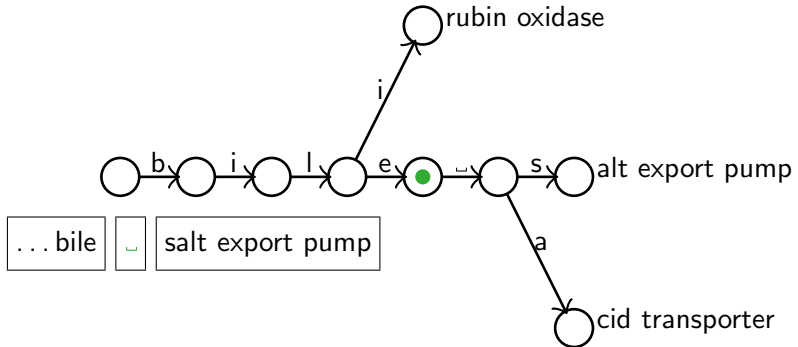
Cascaded Lookup, Character Level



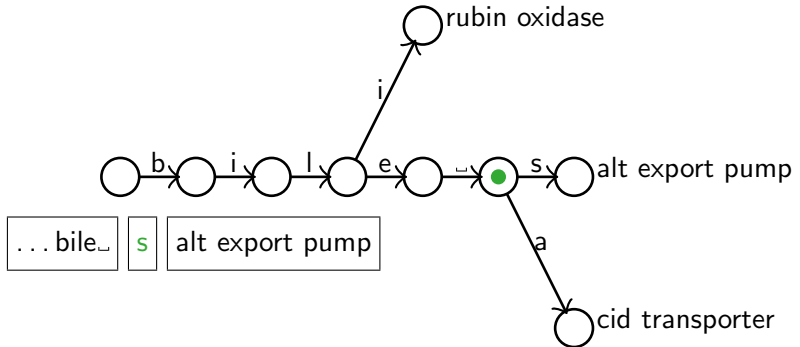
Cascaded Lookup, Character Level



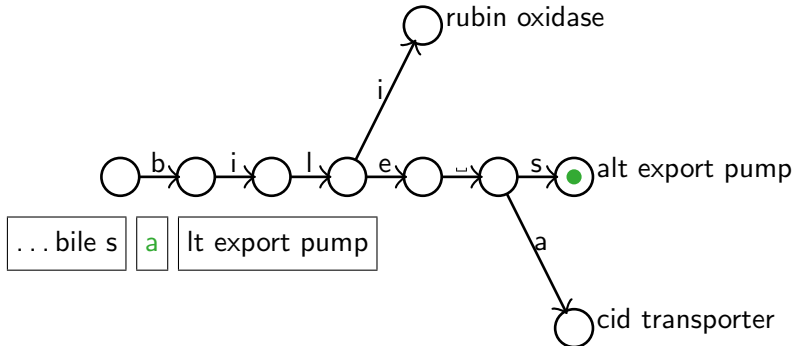
Cascaded Lookup, Character Level



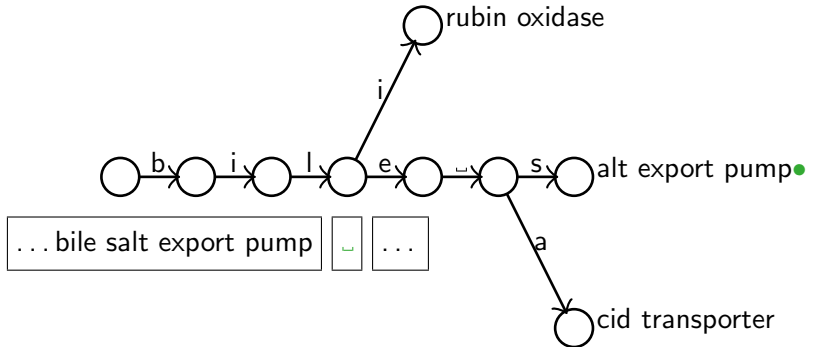
Cascaded Lookup, Character Level



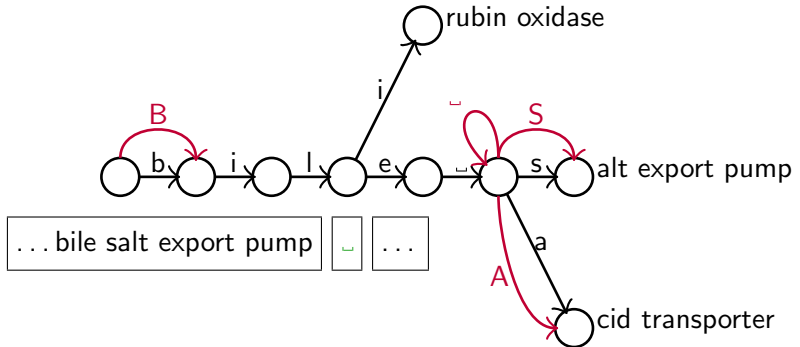
Cascaded Lookup, Character Level



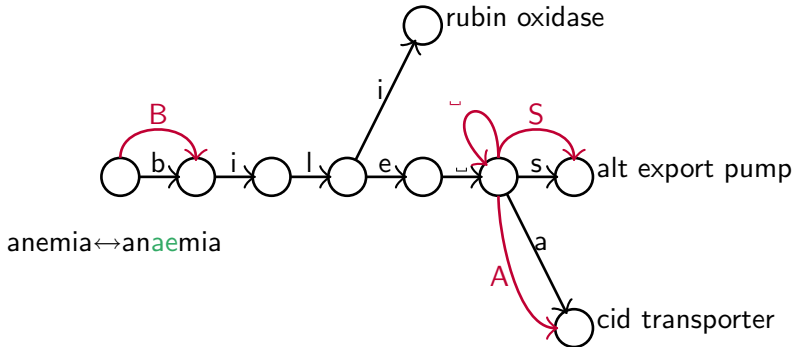
Cascaded Lookup, Character Level



Cascaded Lookup, Character Level



Cascaded Lookup, Character Level



You know what a **Finite Automaton** is!

Shameless Plug

Java Finite Automaton Library — monq.jfa

<http://www.ebi.ac.uk/Rehholz-srv/whatizit/software>

```
<mwt>
  <template><uniprot fb="%1" ids="% (2,0,,)">%0</uniprot></template>

  <t p1="0" p2="Q8WXI3">ASB-10</t>
  <t p1="0" p2="P38666">F8J2_190</t>
  <t p1="92" p2="043557">LIGHT</t>
  <t p1="0" p2="P53667" p3="P53668" p4="P53669">LIM domain kinase 1</t>
  <t p1="0" p2="032175" p3="032175">yusI</t>
</mwt>
```

... <uniprot fb="92" ids="043557">LIGHTs</uniprot> ...

multi word terms, orthographic variations, *general text fiddling*

Shameless Plug

Java Finite Automaton Library — monq.jfa

<http://www.ebi.ac.uk/Rehholz-srv/whatizit/software>

```
<mwt>
  <template><uniprot fb="%1" ids="% (2,0,,)">%0</uniprot></template>

  <t p1="0" p2="Q8WXI3">ASB-10</t>
  <t p1="0" p2="P38666">F8J2_190</t>
  <t p1="92" p2="043557">LIGHT</t>
  <t p1="0" p2="P53667" p3="P53668" p4="P53669">LIM domain kinase 1</t>
  <t p1="0" p2="032175" p3="032175">yusI</t>
</mwt>
```

... <uniprot fb="92" ids="043557">LIGHTs</uniprot> ...

multi word terms, orthographic variations, *general text fiddling*

Shameless Plug

Java Finite Automaton Library — monq.jfa

<http://www.ebi.ac.uk/Rehholz-srv/whatizit/software>

```
<mw>
  <template><uniprot fb="%1" ids="% (2,0,,)">%0</uniprot></template>

  <t p1="0" p2="Q8WXI3">ASB-10</t>
  <t p1="0" p2="P38666">P8J2_190</t>
  <t p1="92" p2="043557">LIGHT</t>
  <t p1="0" p2="P53667" p3="P53668" p4="P53669">LIM domain kinase 1</t>
  <t p1="0" p2="032175" p3="032175">yusI</t>
</mw>
```

... <uniprot fb="92" ids="043557">LIGHTs</uniprot> ...

multi word terms, orthographic variations, *general text fiddling*

Shameless Plug

Java Finite Automaton Library — monq.jfa

<http://www.ebi.ac.uk/Rehholz-srv/whatizit/software>

```
<mwt>
  <template><uniprot fb="%1" ids="% (2,0,,)">%0</uniprot></template>

  <t p1="0" p2="08WAI3">ASB-10</t>
  <t p1="0" p2="P38666">F8J2_190</t>
  <t p1="92" p2="043557">LIGHT</t>
  <t p1="0" p2="P53667" p3="P53668" p4="P53669">LIM domain kinase 1</t>
  <t p1="0" p2="032175" p3="032175">yusI</t>
</mwt>
```

... <uniprot fb="92" ids="043557">LIGHTs</uniprot> ...

multi word terms, orthographic variations, *general text fiddling*

Disambiguation

- frequent English words ↔ protein/gene names:
Had, who, how, Last, light
- terminology overlaps:
insulin — drug or protein?
- polysemic abbreviations:
MIC — *minimum inhibitory concentration*
Myosin IC heavy chain

Disambiguation of Polysemic Abbreviations

Solved for *frequent* abbreviations in MEDLINE (Sylvain Gaudan).

Observation

- many examples *can* solve the problem
- automatically generated training data

Lesson Learned

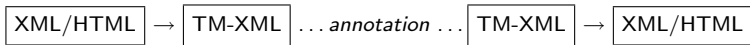
- stop annotating text
- start looking for automatically generated training data³

³even if this appears to be an oxymoron

Technical Issues — Document Format

- publishers have XML
- web sites provide HTML

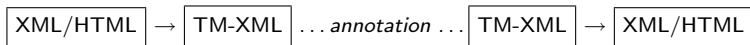
Do we need a **another** document format for text mining?



Technical Issues — Document Format

- publishers have XML
- web sites provide HTML

Do we need a **another** document format for text mining?



NO

⇒ symbiotic (parasitic?) tagging

Symbiotic Tagging

```
<p>...some text...</p>
```



```
<p><textmine>...some text...</textmine></p>
```

Symbiotic Tagging

```
<p>...some text...</p>
```



```
<p><textmine>...some text...</textmine></p>
```

- HTML is bad ⇒ don't be fuzzy about XML structure
- ignore everything outside `<textmine>`
- handle tags within `<textmine>` gracefully 😊
- cleanup after yourself

Summary

- tons of “old” documents need to be tagged
- databases / ontologies / lexicons ⇒ linkable terms
- combine tokenization+term lookup in finite automata
- disambiguation: look for automatic training data
- document format: symbiotic tagging

Thanks:

