HG8003 Technologically Speaking: The intersection of language and technology.

Structured Text and The Semantic Web

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 6
Location: LT8

HG8003 (2014)
<table>
<thead>
<tr>
<th>Lec.</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>01-16</td>
<td>Introduction, Organization: Overview of NLP; Main Issues</td>
</tr>
<tr>
<td>2</td>
<td>01-23</td>
<td>Representing Language</td>
</tr>
<tr>
<td>3</td>
<td>02-06</td>
<td>Representing Meaning</td>
</tr>
<tr>
<td>4</td>
<td>02-13</td>
<td>Words, Lexicons and Ontologies</td>
</tr>
<tr>
<td>5</td>
<td>02-20</td>
<td>Text Mining and Knowledge Acquisition</td>
</tr>
<tr>
<td>6</td>
<td>02-27</td>
<td>Structured Text and the Semantic Web</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quiz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Recess</td>
</tr>
<tr>
<td>7</td>
<td>03-13</td>
<td>Citation, Reputation and PageRank</td>
</tr>
<tr>
<td>8</td>
<td>03-20</td>
<td>Introduction to MT, Empirical NLP</td>
</tr>
<tr>
<td>9</td>
<td>03-27</td>
<td>Analysis, Tagging, Parsing and Generation</td>
</tr>
<tr>
<td>10</td>
<td>Video</td>
<td>Statistical and Example-based MT</td>
</tr>
<tr>
<td>11</td>
<td>04-03</td>
<td>Transfer and Word Sense Disambiguation</td>
</tr>
<tr>
<td>12</td>
<td>04-10</td>
<td>Review and Conclusions</td>
</tr>
<tr>
<td>Exam</td>
<td>05-06</td>
<td>17:00</td>
</tr>
</tbody>
</table>

> Video week 10
Introduction

➢ Review of Text Mining and Knowledge Acquisition

➢ Structured Text and Markup

➢ The Semantic Web
Too much information for people to handle: Information Overload

Text mining is:

The discovery by computer of new, previously unknown information, by automatically extracting information from a usually large amount of different unstructured textual resources.
LARGE amounts of data

➢ You can tolerate some noise
  ➢ conversion errors, spelling errors, etc.

➢ Shallow robust techniques are needed

➢ Typically only consider more things with more than \( n \) instances
  ➢ Hope that errors are infrequent
Template Filling

➤ Looking for known relations in text
  ➤ fill slots in a template

➤ Restricted search space gives high accuracy
Named Entity Recognition

➤ Identify interesting things
  (person, organization, location, date/time, ...)

➤ Typically done as a sequence labeling task
  ➤ Tag as Inside, Outside, Beginning, (IOB)

➤ Train a classifier with annotated text
  ➤ Features include: Words, Stems, Shape, POS, Chunks, Gazetteers
Relation Detection

> We can use patterns to find relation tuples

> < S, hypernym, A >

> S (such as | like | e.g.) A; A and other S; S (including | especially) A

> < A, synonym, B >

> both A and B; either A or B; neither A nor B

> Simple patterns are easy to find in vast data sources

> High frequency patterns can be quite reliable

> multiple patterns increase confidence
# Evaluation Measures

<table>
<thead>
<tr>
<th>System</th>
<th>Actual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>target</td>
<td>not target</td>
</tr>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>

Precision = \( \frac{tp}{tp+fp} \); Recall = \( \frac{tp}{tp+fn} \); \( F_1 = \frac{2PR}{P+R} \)

- **tp** True positives: system says Yes, target was Yes
- **fp** False positives: system says Yes, target was No
- **tn** True negatives: system says No, target was No
- **fn** False Negative: system says No, target was Yes
There is a lot of information out there

Much of it is unstructured text

Using NLP techniques we can extract this information

But we can’t trust it all

Well defined tasks on restricted domains work best
Structured Text
Structured Text and The Semantic Web

➤ The Internet

➤ The Structure of Markup

➤ The Structure of the Web

➤ The Future of the Web

➤ Linguistic features of the web
The Internet

➢ Global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP)

➢ Carries several services

➢ HTTP (Hyper Text Transfer Protocol) — The Web
➢ Email
➢ VoIP (Voice over IP) — Telephony/Skype
➢ FTP, . . . (File Transfer)
➢ Streaming Media — music, video
➢ Instant Messaging
Growth of the Internet

Internet users per 100 inhabitants 1997-2007 (Source: ITU)

http://upload.wikimedia.org/wikipedia/commons/a/a6/Internet_users_per_100_inhabitants_1997-2007_ITU.svg
Why Structured Text?

➤ Reduce Ambiguity

➢ Need to make meaning explicit

➢ Traditionally this is done by annotating text in some way
Annotation on how to print is called markup

- underlining to indicate boldface
- special symbols for passages to be omitted
- special symbols for printed in a particular font

This existed before computers

- Editors would markup hand-written manuscripts
- ...and pass them to type setters
- ...who would prepare the manuscript for printing
Printers’ Markup

Style of Type

- Wrong font (size or style of type)
- Lower case letter
- Set in LOWER CASE
- Capital letter
- SET IN capitols
- Set in lower case with INITIAL CAPITALS
- SET IN small capitals
- Set in small capitals with initial capitals

Equalize space between words

Insertion and Deletion

- Caret (insert a marginal addition)
- Delete (take it out)
- Correct letter or word marked
- Let it stand (all matter above dots)

Paragraphing

- Begin a paragraph
Early Computer Markup (troff)

Marked up with troff

Postscript and PDF (Portable Document Format) are similar
Visual Markup vs Logical Markup

➤ **Visual Markup** (Presentational)
➤ What you see is what you get (WYSIWYG)
➤ Equivalent of printers’ markup
➤ Shows what things look like

➤ **Logical Markup** (Structural)
➤ Shows the structure and meaning
➤ Can be mapped to visual markup
➤ Less flexible than visual markup
➤ More adaptable (and reusable)
Standard Generalized Markup Language: SGML

- ISO standard based on IBM’s GML
- Attempt to make markup independent of processor
  - Important for archiving information
- Emphasis on logical markup
- Popularized the use of `<tag>` </tag> notation
  - and entities &lt; &gt; when you need an <>
- Split the document into: Declaration, Prolog, Documentation
Hyper Text Markup Language: HTML

➤ Markup Language for web pages

➤ An extension of SGML

➤ Combines logical and visual markup

➤ Also allows hyperlinks (linking and anchoring)

➤ Created by Tim Berners-Lee at CERN (1989)
  ➤ to make physics papers and documentation more accessible
HTML example

> Logical

```html
<h1>Headline</h1>
<p>and some text</p>
```

> Visual

```html
<font size="3"><b>Headline</b></font>
<br>and some text
```
Logical allows various styles

```
<style>
H1 {
    font-size: 24px;
    color: blue;
    margin-top: 10px;
    margin-bottom: 15px;
}
</style>

➤ This can be done using CSS (Cascading Style Sheets)
➤ Separate Logical and Visual Structure
```
Benefits of Logical Tags

➢ Can transform things easily
  ➢ No bold for Japanese and Chinese (just use size)

➢ Logical form useful for other tasks
  ➢ Summarization
    * Just show <h1> ... <h3>
  ➢ Translation
    * Headers are noun phrases, not sentences

➢ Robustness: you can read the source directly
But still there is ambiguity!

- Tags on one site may not mean the same thing on another site

- Huge amount of information
  - Looking for Eric Miller may get the wrong one!
  - Looking for NTU gets
    * Nanyang Technological University
    * National Taxpayers Union
    * National Taiwan University

- What can we do?
Hypertext

➤ HTML crucially adds hyperlinks
   ➤ these extend text in a new way
   ➤ references that you can immediately access

➤ <href="http://somewhere.on.the.web">link me</a>

➤ <img src="http://somewhere.on.the.web/pic.jpg">

➤ Immediately accessible references are qualitatively different
<!doctype html>
<html>
  <head>
    <title>Hello HTML</title>
  </head>
  <body>
    <p>Hello World!</p>
    <p>Oh well, <span lang="fr">c’est la vie</span>, as they say in France.</p>
    <abbr id="anId" class="jargon" style="color:blue;" title="Hypertext Markup Language">HTML</abbr>
  </body>
</html>
The Structure of the Web

- 550 billion documents on the Web (2001)
  mostly in the invisible Web, or deep Web

- 11.5 billion indexable web pages (2005)

- 25.21 billion indexable web pages (2009)

- 109.5 million websites (2009)
Take out your clickers!
The Semantic Web
The Semantic Web

➢ What is it?

➢ How is it built?

➢ Why is it being built?

➢ Problems
The Semantic Web

➢ A vision of a more useful World Wide Web

➢ the meaning of information and services on the web is defined

➢ machines (and people) can understand the web

The Web as a universal medium for data, information, and knowledge exchange.
The Web as it is now

➤ Resources
  ➢ Identified by URLs (uniform resource locator)

➤ Links
  ➢ href, src, ...
  ➢ limited, non-descriptive

➤ Human User
  ➢ Exciting World - Semantics deduced from context

➤ Machine User
  ➢ Very little explicit information
The Current Web

Structured Text and The Semantic Web
The Web as it could be

➤ Resources
  ➢ Globally Identified by URIs (uniform resource identifier)
  ➢ Extensible, Relational

➤ Links
  ➢ Identified by URIs
  ➢ Extensible, Relational

➤ Human User
  ➢ Even more exciting world - Richer user experience

➤ Machine
  ➢ More processable information
The Semantic Web

Diagram showing the relationships between different entities:
- Software
- Document
- Library
- Image
- Topic
- Person
- Place

Relationships:
- hasManual
- requires
- inPartOf
- subject
- hasAuthor
- livesAt
Semantic Web Goals

➢ Web of Data
  ➢ provides common data representation framework
  ➢ makes possible integrating multiple sources
  ➢ so you can draw new conclusions

➢ Increase the utility of information by connecting it to definitions and context

➢ More efficient information access and analysis

  E.G. not just "color" but a concept denoted by a Web identifier:
  <http://en.wikipedia.org/wiki/Sapphire_(color)>
Identifiers and Characters

- Characters are always defined using Unicode

- Identifiers are Uniform Resource Identifiers
  
  - **Universal Resource Name**
    * e.g., urn:isbn:1575864606
    * uniquely identifies one edition of a book
  
  - **Universal Resource Locator**:
    * scheme:part?query#anchor
    * e.g., http://www3.ntu.edu.sg/home/fcbond/
    * http, skype, ssh, secondlife, ...

- Identifies or names a resource on the web
XML: eXtensible Markup Language

➤ XML is a set of rules for encoding documents electronically.

➤ Based on a simplified SGML

➤ XML’s design goals emphasize simplicity, generality, and usability.

➤ It is a textual data format

➤ It supports many encodings, with Unicode preferred

➤ It can represent arbitrary data structures, for example in web services.
In other words

- XML stands for EXtensible Markup Language
- is a markup language much like HTML
- was designed to carry data, not to display data
- tags are not predefined. You must define your own tags
- is designed to be self-descriptive
- XML is a W3C Recommendation
With XML You Invent Your Own Tags

<bitext>
<author xml:lang="eng">Francis Bond</author>
<author xml:lang="jpn">フランシス ボンド</author>
<email>fcbond@ntu.edu.sg</email>
<opendata>yes</opendata>
...
</bitext>

➢ The tags describe the content, but do not do anything.

➢ You have to define the meaning elsewhere
XML structure

➤ An **XML element** looks like this
   <tag attribute='value'>Content</tag>

➤ You can use **attributes** (easy to verify)
   <author xml:lang="eng">Francis Bond</author>

➤ or **nested elements** (flexible)
   <author>
     <lang>eng</lang>
     <name>Francis Bond</name>
   </author>

➤ If all elements are written correctly the document is **well formed**
This XML file does not appear to have any style information associated with it. The document tree is shown below.

```xml
<bitext>
  <author xml:lang="eng">Francis Bond</author>
  <author xml:lang="jpn">フラニス ボンド</author>
  <email>fcbond@ntu.edu.sg</email>
  <opendata>yes</opendata>
  <source>http://www.sil.org/iso639-3/iso-639-3_20090210.tab
  <source_language>eng</source_language>
  <target_language>jpn</target_language>
  <sentence sid="s1">
    <source>The Cathedral and the Bazaar</source>
    <target>大聖堂とバザール</target>
  </sentence>
  <sentence sid="s2">
    <source>Linux is subversive.</source>
    <target>リナックスは破壊的である。</target>
  </sentence>
  + <sentence sid="s3"></sentence>
  - <sentence sid="s4">
    <source>Certainly not I.</source>
    <target>ぼくもできなかった。</target>
  </sentence>
</bitext>
```
You can also define what tags are possible with

- DTD (Document Type Definition)
- XML Schema

```xml
<!DOCTYPE bitext [ 
  <!ELEMENT author (#PCDATA)>
  <!ATTLIST author xml:lang CDATA>
  <!ELEMENT sentence (source,target)>
  ...
```

This makes the structure explicit
Why Use a DTD?

➢ With a DTD, XML files can carry a description of their own format.

➢ With a DTD, independent groups of people can use a well defined, enforceable standard for interchanging data.

➢ Your application can use a standard DTD to verify that the data you receive from the outside world is valid.

➢ You can also use a DTD to verify your own data.
  ➢ You can catch format errors early

➢ BUT, XML only defines structure not content
Validation

- Validation is very important
- Ill-formed data makes parsing complex
- Early detection of errors is cost-effective
- Validated data is easy to maintain
Semantic Web Architecture

- User interface and applications
- Trust
- Proof
- Unifying logic
- Querying: SPARQL
- Ontologies: OWL
- Rules: RIF/SWRL
- Taxonomies: RDFS
- Data interchange: RDF
- Syntax: XML
- Identifiers: URI
- Character set: UNICODE
- Cryptography
We want to identify content

Annotate information with descriptions: triples of information

- subject
- predicate
- object

e.g. <dog, hyponym, animal>
e.g. <Petter Haugereid, teacher, HG8003>

The trick is that each element is a URI

You can say anything about anything
RDF graph describing Eric Miller

Relational Semantic Representation (again!)
RDFs are written using XML

```xml
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>
</rdf:RDF>
```

➢ Not designed for people to read (or write) directly
➢ Designed to be extensible and explicit
Why use URIs in RDFs with XML?

- A shared URI gives a shared definition
- You can add new URIs as necessary
- RDFs are complicated
  - you need to validate the syntax → XML
- RDFs assume the existence of the web
  - Nothing has meaning on its own
  - You shall know a URI by the company it keeps
OWL and Ontologies

- Allowing any URI makes information hard to combine

- Use Ontologies to link it together again
  - You can normally agree on a hypernym (supertype)

- Agreeing on an ontology is difficult
  - Many detailed ontologies
  - One common ontology
    - The Standard Upper Merged Ontology (SUMO)
    - Links many ontologies (including WordNet)
  - Much recent work on medical and library domains
Common Semantic Web Ontologies

➤ People + Organisations
  ➢ FOAF, HCard, Relationship, Resume

➤ Places
  ➢ Geonames, Geo

➤ Events
  ➢ RDFCalendar

➤ Social Media
  ➢ SIOC, Review
Topics + Tags
  - SKOS, MOAT, HolyGoat

eCommerce
  - GoodRelations, CC Licensing

More...
  - Scovo, DOAP, Recipes, Measurements, ...

General things
  - SUMO, WordNet
Semantic Web Architecture

User interface and applications

Trust

Proof

Unifying logic

Querying: SPARQL

Ontologies: OWL

Rules: RIF/SWRL

Taxonomies: RDFS

Data interchange: RDF

Syntax: XML

Identifiers: URI

Character set: UNICODE
Querying the (semantic) web is

- Slow (non local access → millions of times slower)
- Non-deterministic (the answers change)
- Vast (could be trillions of elements)

A whole new set of technical problems

Logic, Proof and Trust are still works in progress

- Reusing AI research from the 70s and 80s

But applications are going along without them

- Enhanced search using metadata
- Combinations of data
Semantic Web Lessons

➤ RDF as a general information model is applicable to many uses (many of which we never even thought about)

➤ Common data representation and architecture drives down costs (technical and social)

➤ Facilitates serendipitous interoperability
  - breaking down the barriers of domain knowledge

➤ When "Anyone can say anything about anything", who you trust is important (same as in text mining)

➤ Beneficial to solving interoperability in Open (rather than Closed) systems
The goal of the Semantic Web is to share knowledge

- Uses markup to give tractable annotation
  - Unicode
  - XML
  - URI
  - RDF
  - OWL
- Relies on web resources to make common assumptions explicit
Example of making information explicit

< Francis Bond, author, Translating the Untranslatable >

➢ Francis Bond: http://www3.ntu.edu.sg/home/fcbond/
  ➢ defined using Francis Bond’s homepage

➢ author: http://purl.org/dc/elements/1.1/creator
  ➢ defined using the Dublin Core Ontology

or http://nlpwww.nict.go.jp/wn-synset.cgi?synset=10794014-n
  ➢ defined using WordNet

➢ Translating the Untranslatable: urn:isbn:1575864606
  ➢ defined using the ISBN ontology
An Example of Data Integration

➤ Map the various data onto an abstract data representation
➤ make the data independent of its internal representation
➤ Merge the resulting representations
➤ Start making queries on the whole!
➤ queries that could not have been done on the individual data sets

Slides from “An introduction to the Semantic Web (Through an Example)”
By Ivan Herman (http://www.w3.org/People/Ivan/CorePresentations/IntroThroughExample/)
A simplified bookstore data (dataset “A”)

<table>
<thead>
<tr>
<th>ID</th>
<th>Author</th>
<th>Title</th>
<th>Publisher</th>
<th>Year</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Home Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>id_xyz</td>
<td>Ghosh, Amitav</td>
<td><a href="http://www.amitavghosh.com">http://www.amitavghosh.com</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Publ. Name</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>id_qpr</td>
<td>Harper Collins</td>
<td>London</td>
</tr>
</tbody>
</table>
Export your data as a set of relations

Diagram showing a graph with nodes labeled "The Glass Palace" and "2000" connected by edges labeled "a:title" and "a:year" respectively, and a node labeled "London" connected by an edge labeled "a:city" to a node labeled "Harper Collins" connected by an edge labeled "a:p_name" to a central node labeled "a:author" connected to a node labeled "Ghosh, Amitav" and a node labeled "http://www.amitavghosh.com" connected by an edge labeled "a:name" and "a:homepage" to a central node labeled "http://.../isbn/000651409X" connected by edges labeled "a:year" and "a:city".
Some notes on the exporting the data

➢ Relations form a graph

➢ the nodes refer to the “real” data or contain a string
➢ how the graph is represented in machine is immaterial for now
➢ Data export does not necessarily mean physical conversion of the data
  * relations can be generated on-the-fly at query time
    · via SQL “bridges”
    · scraping HTML pages
    · extracting data from Excel sheets
    · through text mining
    · etc.
  * One can export part of the data
Another Set of Data (set “F”)

http://.../isbn/000651409X

http://.../isbn/2020386682

Amitav Ghosh

Le palais des mirroirs

Christiane Besse
Export Set F
Merge Identical Resources
User of data “F” can now ask queries like:

➢ “give me the title of the original”
➢ well, “donnes-moi le titre de l’original”

➢ This information is not in the dataset “F”

➢ …but can be retrieved by merging with dataset “A”!
However, more can be achieved

➢ We “feel” that a:author and f:auteur should be the same

➢ But an automatic merge does not know that!

➢ Let us add some extra information to the merged data:

➢ a:author same as f:auteur

➢ both identify a “Person”

➢ a term that a community may have already defined:
  ▶ a “Person” is uniquely identified by his/her name and, say, homepage
  ▶ it can be used as a “category” for certain type of resources
Use more knowledge
Start making richer queries!

- User of dataset “F” can now query:
  - “donnes-moi la page d’accueil de l’auteur de l’originale”
  - well... “give me the home page of the original’s ‘auteur’”

- The information is not in datasets “F” or “A

- ... but was made available by:
  - merging datasets “A” and datasets “F”
  - adding three simple extra statements as an extra “glue”
Merge with Wikipedia Data
What did we do?

- We combined different datasets that
  - are somewhere on the web
  - are of different formats (mysql, excel sheet, XHTML, etc)
  - have different names for relations

- We could combine the data because some URI-s were identical (the ISBNs in this case)

- We could add some simple additional information, using common terminologies that a community has produced

- As a result, new relations could be found and retrieved
It could become even more powerful

➢ We could add extra knowledge to the merged datasets
  ➢ e.g., a full classification of various types of library data
  ➢ geographical information
  ➢ etc.

➢ This is where ontologies, extra rules, etc, come in
  ➢ ontologies/rule sets can be relatively simple and small, or huge, or anything in between...

➢ Even more powerful queries can be asked as a result!!!
Criticism of the Semantic Web

Doctorow’s seven insurmountable obstacles to reliable metadata are:

1. People lie
2. People are lazy
3. People are stupid
4. Mission Impossible: know thyself
5. Schemas aren’t neutral
6. Metrics influence results
7. There’s more than one way to describe something

http://www.well.com/~doctorow/metacrap.htm
Other Issues

Other reasons that result in metadata becoming obsolete are:

1. Data may become irrelevant in time

2. Data may not be updated with new insights

Reliable Metadata

➢ Information people use
  ➢ Number of links into a page
  ➢ Text on the page

➢ Even this gets gamed (link farms, spam pages, . . .)
The Semantic Web is about structuring data

Text Mining is about unstructured data

There is much more unstructured than structured data

- NLP can infer structure
- NLP makes the Semantic Web feasible
- the Semantic Web can be a resource for NLP
Readings

➤ The Semantic Web and Ontologies
http://www.obitko.com/tutorials/ontologies-semantic-web/

➤ Criticism of the Semantic Web
http://www.well.com/~doctorow/metacrap.htm


➤ Semantic Web in general: semanticweb.org

➤ Nice presentation: http://www.w3.org/2004/Talks/0120-semweb-umich/

Slides are taken from many of these.