Term Recognition
PSB Tutorial 2001

Sophia Ananiadou
Computer Science
University of Salford
Outline

• Scope and problems of term recognition
  – Term formation patterns
  – Ambiguity & variation

• Term recognition approaches
  – IR Automatic Indexing
  – Linguistics
  – Statistics
  – Hybrid

• Application to Molecular Biology
  – Statistics
  – HMM / Decision Trees
Scope of Automatic Term Recognition

• What do we expect from ATR?
• What are terms?
• Terms are the linguistic realisation of specialised concepts
  – *protein kinase C, basal cell carcinoma*
• TR is not a simple task

• ‘known’ words might be terms
• term recognition requires conscious thought
• no formal criteria to distinguish terms from non terms
• frequency is not always a good indicator
• unlike words chief motivation of term designation is *classification*
Term Recognition

• Terms characterise the content of a text (domain)

• ATR is important for robust processing
• handling unknown words;
• updating lexical and knowledge resources;
• maintaining consistency (avoiding terminological confusion)
Term Formation Patterns

- Words are different entities from terms.
- Term formation follows some pre-existing patterns
- Use **existing resources**, narrowing or widening the meaning of a wordform *mouse, table*
- Modification of existing resources, morphological and syntactic transformations
  - **affixation** *methyltransfer-ase, act-in*
  - **compounding** *protein kinase C*
  - **abbreviation** *EGFR (epidermal growth factor receptor)*
Term formation patterns

• Use of new resources (neologisms)
• Use of Roman / Greek letters / stems on their own or as part of compounds
  • protein kinase alpha
  • annexin II mRNA
• Use of numerical letters
  • NIH 3T3 fibroblasts
  • CYP1A1 promoter
  • Ca2+-calmodulin-dependent protein kinase IV
• Eponyms
  • Jurkat T cells
Problems with Terms

Ambiguity

• Several ways in which terms can be ambiguous

• **Syntactic** {noun/verb, adjective/noun}

• POS taggers are reasonably accurate (95%-99%) so correct syntactic identification not so problematic with appropriate training
Semantic

- Polysemy: when the same term form refers to many concepts related in meaning e.g. lens
  - a combination of lenses used in photography;
  - Anat. Crystalline lens;
  - Physics - a device for focusing or modifying the direction or movement of light, sound, electrons, etc.
- Homonymy: identical surface form, unrelated meaning
  - plant
  - cell
Problems with Terms

• Majority of terms are multi-word units (compounds)
  – the constituent bracketing problem.. Nested terms
  – [ Leukaemic[T [cell line]] Kit225]
  – problems in the interpretation of the compound

• Multidimensionality of terms
  – multiple inheritance within a hierarchy
  – pulmonary tuberculosis is both a respiratory disorder and an infectious disorder
  – UMLS ‘signs/symptoms’ ‘diseases/abnormalities’
Term Variation

• A concept can be represented by more than one term (synonyms or term variants)
• Term variation is a problem for IR, IE
• How do terms vary?
  ↑ morphologically enzyme activity/ies
  enzymatic activity
  ↑ syntactically enzyme amidolytic activity
  activity of enzyme
  enzyme and bactericidal activity
  ↑ semantically fermentation
Term Variation

↑ orthographically signalling pathway
  signaling pathway (US vs UK English)
  hyphens amino-acid amino acid

↑ compression or shortening of a term, identification of acronyms and other abbreviations
  • highly ambiguous as they may refer to different complex form
  • acronym ATR (Mountain Data Systems)
    Above Threshold Reporting / Air Turbo Rocket
    Acceptance Test Report / Automatic Target Recognition
    Aerial Target Review / Advanced Tactical Radar …etc…
A system for Term Variation

• FASTR system deals with term variation (French & English) on already extracted terms (Jacquemin)

• Term variants obtained through a metarule mechanism dynamically calculated

  \[ \text{serum albumin} \]

  rule 1: \( N_1 \not\in N_2 N_3 \)

  simple \( N + N \) rule
• To dynamically produce rules that handle variations of the simple $N + N$ construction
  
  Metarule $\text{Coord}(X_1 \rightarrow X_2 X_3) = X_1 \rightarrow X_2 \ C_4 \ X_5 \ X_3$
  
  $\text{coord}(N_1 \rightarrow N_2 N_3) = N_1 \rightarrow N_2 \ C_4 \ X_5 \ N_3$
  
  (instantiated rule)

  *serum albumin* $\rightarrow$ *serum and egg albumin*

  the new term found is *egg albumin*

• Insertion rule: *medullary carcinoma* $\rightarrow$ *medullary thyroid carcinoma*

• Permutation rule: *control centre* $\rightarrow$ *centre for disease control*
Methods of Automatic Term Recognition

• IR techniques (indexing) not NLP oriented

• Linguistic approaches

• Statistic approaches

• Hybrid
Automatic Indexing

• Many researchers naturally think that IR techniques can be relevant for ATR
• IR indexing techniques like ATR focus on identifying words.
• To what extent however are IR indexing techniques useful for ATR?
Basic assumptions of automatic indexing

• a word of a document is a likely index term for it
• a frequent word in a document is a likely index term
• a word appearing in a limited number of documents is a likely index term for those documents
• a word that is more frequent in one document compared with the entire collection is a likely index term for that document
• measures of the distribution of words in a collection rather than their frequency of occurrence can reveal useful index terms

(Kageura & Umino, 1996)
Index terms vs Technical terms

• The goal of indexing is to find index terms that allow you to discriminate one document from another.

• The goal of ATR is to find technical terms that indicate concepts of specialised languages.

• Index terms are not necessarily technical terms.
  – A set of index terms will typically include a mixture of general language words and technical terms.
  – Most indexing methods emphasise single word index terms.

• Technical terms are not necessarily index terms for some collection of documents.
Example of an indexing system used for ATR

- CLARIT (http://www.clarit.com)
- Advocates textual processing to detect complex terms
- Performs morphological analysis
- Detects NPs
- Constrains the compounds eventually proposed by using two heuristics:
  - if there is a close relationship between the elements of a compound and they behave like a single word (lexicalisation)
  - to what extent one can insert words between the elements of a compound
- Both heuristics are based on statistics
Linguistic Approach

• Based on linguistic pre-processing and annotations (result of taggers, shallow parsers)
• Detect recurrent syntactic term formation patterns
  – Noun + Noun
  – (Adj | Noun) + Noun
• Detect recurrent domain specific morphological term formation patterns
  – Noun + suffix (-ase, -in)
• Mostly concentrated on Noun Phrases
Linguistic Pre-processing

- **Tokenize text**
- As we tokenize we might want to label a token with some information e.g. `NF [orth=upper], Cys [orth=mixed]` etc
- Not an easy task as usually we have little linguistic information at this stage
- Deals with punctuation; commas; dashes; hyphens, whitespace
- **text-based medium** (one word or two?)
- compound nouns (data base or database or data-base)
- V-shaped perforation or ‘V’-shaped perforation
Morphological processing

• Morphological analysis is linguistics based; removes affixes, turns each affix into a linguistic feature bundle, produces word tree showing how morphemes combine to form a word
• Stemming (stripping endings, protein kinas)
• Often beneficial for statistically based term extraction because you use normalised forms (use stems instead of many variant forms)
• Typically very fast; quick and dirty; non linguistic oriented
• Stemming used in IR (depending on the stemmer, overstemming or understemming)
• Might lose information
• Porter’s algorithm for stemming; several languages

http://open.muscat.com/stemming/
Linguistic processing

- Tagging (rule based, statistically based taggers)
- Add POS labels to text words

polymorphs/NNP and/CC nuclear/JJ debris/NN are/VBP present/JJ in/IN the/DT *keratinous/NNP* cap/NN ./.

- Taggers don’t get it 100% right!
Systems based on linguistics

• Ananiadou, S. (1994) recognises single-word terms from domain of Immunology based on morphological analysis of term formation patterns (internal term make up)

• based on analysis of neoclassical and hybrid elements
  ‘alphafetoprotein’ ‘immunoosmoelectrophoresis’
  ‘radioimmunoassay’

• some elements are used mainly for creating terms and marked

  term ⊠ word + term_suffix
  term ⊠ term + word_suffix

• neoclassical combining forms (electro-),
• prefixes (auto-, hypo-) suffixes ( -osis, -itis)
periimmunoglobulinaemia

peri
[cat=prefix, type=term, prefixes=N/V/Adj, makes=N/V/Adj]

immuno
[cat=comb, type=term]

globulin
[cat=comb, type=term]

aemia
[cat=suffix, type=term, suffixes=comb, makes=N]
• **Justeson & Katz (1995, TERMS)** extract complex terms (TNP) based on two characteristics (which distinguishes them from non terms)
  • the syntactic patterns of TNPs are restricted
  • terms appear with the same form throughout the text, omissions of modifiers are avoided

• The text is tagged; a filter is applied to extract terms
  
  \[((A|N) + | ((A|N) * (NP)?) (A|N)*) N

• AN / NA / AAN / ANN / NAN / NNN / NPN

• Filtering based on simple POS pattern
  – A pattern must occur above a certain threshold to be considered a valid term pattern.

• Recall: 71%   Precision: 71% -- 96%
• **LEXTER** (Bourigault, 1994)
• Extracts French compound terms based on surface syntactic analysis and text heuristics
• Terms are identified according to certain syntactic patterns
• Uses a boundary method to identify the extent of terms
  – categories or sequences of categories that are not found in term patterns form the boundaries e.g. verbs, any preposition (except de and à) followed by a determiner. Non productive sequences become boundaries.
• Claims syntactic information alone is sufficient for ATR
• Domain independent
• A learning component detects productive nouns
• Precision: 95% although tests have shown that lots of noise is generated
Approaches using statistical information

Main measures used:

- Frequency of occurrence
- Mutual Information
- C/NC value
- Experiments also with loglike coefficient [Dunning, 1993]
Frequency of occurrence

- Simplest and most popular method for ATR
- Domain independent, requires no external resources
- Some filtering is used in form of syntactic patterns
- Drawback: low frequency terms might be valid terms
- Frequency works well with fixed phrases
- Systems using frequency of occurrence
  - Dagan & Church (TERMIGHT, 1994)
  - Enguehard & Pantera (1994)
  - Lauriston (TERMINO, 1996)
Mutual Information

• Used initially for the extraction of collocations (Fano, 1961, Church & Hanks, 1989)

• Collocations are expressions consisting of two or more words corresponding to a conventional way of saying things. Word co-occurrences
  ‘strong tea’ ‘weapons of mass destruction’ ‘broad daylight’
  ‘kick the bucket’ ‘powerful drug’

• Overlap between collocations / technical term
  – Collocations have limited compositionality
  – Technical terms uniquely identify a concept but accept a limited number of modifiers

• Firth (Contextual Theory of Meaning) ‘you shall know a word by the company it keeps’...
**Mutual Information**

- ‘The amount of information provided by the occurrence of the event represented by $y_i$ about the occurrence of the event represented by $x_k$ is defined as’

$$I(x_k, y_i) \equiv \log \frac{P(x_k, y_i)}{P(x_k) \cdot P(y_i)}$$

Fano (1961:27-28)

- This measure is about how much a word tells us about the other.
- Problems for MI come from data sparseness; overestimates bigrams composed of low-frequency words.
- Becomes complicated when used on strings longer than two words.
- Damerau (1993) and Daille (1994) used MI for the extraction of candidate terms (only for two-word candidate terms)
C/NC value (Frantzi & Ananiadou)

• **C/value**
  - total frequency of occurrence of string in corpus
  - frequency of string as part of longer candidate terms
  - number of these longer candidate terms
  - length of string (in number of words)

• **Extracts multi-word, nested terms**
  adenoid cystic basal cell carcinoma
  cystic basal cell carcinoma
  ulcerated basal cell carcinoma
  recurrent basal cell carcinoma
  basal cell carcinoma
C-value measure

\[
C\text{-value} = \begin{cases} 
\log_2 |a| \cdot f(a); & \text{a not nested} \\
\log_2 |a| \cdot \left( f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right); & \text{otherwise}
\end{cases}
\]
Adding contextual information: NC-value

- Assigns weights to context words (V/Adj/N)
- Number of terms a context word appears; its frequency as a context word; its total frequency in the corpus
- Weight measure assigns high values to words that tend to appear with terms:

\[
weight(w) = \frac{T(w)}{n}
\]
NC value

\[ NC-value(a) = 0.8 \times C-value(a) + 0.2 \times CF(a) \]

\( a \) is the candidate term,
\( C-value(a) \) is the \textit{C-value} for the candidate term \( a \),
\( CF(a) \) is the context factor for the candidate term \( a \)

- we obtain the CF by summing up the weights for its term context words, multiplied by their frequency appearing with this candidate term.
Sample NC value results from Nuclear Receptors corpus

<table>
<thead>
<tr>
<th>NC value</th>
<th>Candidate Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>127</td>
<td>Protein kinase</td>
</tr>
<tr>
<td>85</td>
<td>Growth factor</td>
</tr>
<tr>
<td>57</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>42</td>
<td>Protein kinase C</td>
</tr>
<tr>
<td>33</td>
<td>Amino acid</td>
</tr>
<tr>
<td>32</td>
<td>Cell line</td>
</tr>
<tr>
<td>31</td>
<td>Mitogen-activated protein kinase</td>
</tr>
<tr>
<td>30</td>
<td>PC12 cell</td>
</tr>
<tr>
<td>29</td>
<td>MAP kinase</td>
</tr>
<tr>
<td>21</td>
<td>14-3-3 protein</td>
</tr>
</tbody>
</table>

Precision: 99% --91%

Ananiadou et al. (2000)
Hybrid approaches

- Combination of linguistic information (filters), shallow parsing results and statistical measures
- Daille, B., Frantzi & Ananiadou

- Adding contextual information
- Intrinsic information about term (term formation patterns) and extrinsic (contexts)
Contextual Information and Term Extraction

• Types of contextual information
• syntactic (some syntactic categories are more important at term prediction)
• semantic (measure similarity based on thesaurus UMLS)
• terminological (context words which are terms are important)
• relevance of these different aspects of context encoded as weights and incorporated into:
  • Importance Weight (Maynard & Ananiadou)
How can we use terminological contextual information?

- Term disambiguation based on context clustering
- Information about unknown words
- Populating ontologies
- Improve term extraction
Automatic Term Extraction and Biology

- Terminological confusion
- Populating ontologies
- Discovery of relations between terms
- Discovery of domain tags for text annotation
- Important input to any Information Extraction system
Current research on Term Extraction in Biology

- Rule based (linguistics)
- Hybrid (statistics & linguistics)
- Machine Learning (Decision Trees, HMM)

Rule based

- Rule based (linguistic processing) Gaizauskas et al. http://www.dcs.shef.ac.uk/research/groups/nlp/pasta
- Use of (50) terminology lexicons derived from biology databases (SWISS-PROT) and annotated corpora
- Rule based terminology parser combines constituents of term classes resulting from morphological processing and lexical lookup
  - calmodulin N-methyltransferase
  - enzyme-modifier enzyme-head
- Derive rules semi-automatically from lexicon
Using a Hybrid approach

- **BioPath** project (LION / Valigen)
  - [http://www.cms.salford.ac.uk/NLP.html](http://www.cms.salford.ac.uk/NLP.html)
- Text Mining from Medline: 11 m abstracts, 70 - 100 m sentences, 2.5 m terms
- IE (pattern extraction), Information categorisation (clustering), ATR
- piSCOUT & termSCOUT deal with pattern definition and relation extraction; relation viewer tool.
- ATR is based on C/NC values augmented with semantic contextual information
- Precision: 97% - 99%
Methods based on Machine Learning

- Typical Methods: Decision Trees, HMM
- Both methods construct internal structures from training data, that is, decision trees and state transition networks.
- Since they need training data, they are called supervised learning methods.
- The performance of these methods are highly dependent on the set of features we choose in advance to characterize data.
Using decision trees (Nobata et al.)

http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/

Each node has one condition:

- Is the current word in the Protein term list?
  - Yes
  - No
- Does the previous word have figures?
  - Yes
  - No
  - Yes
- What is the next word’s POS?
  - Noun
  - Verb
  - …

Each leaf has one class:

- Unknown
- PROTEIN
- DNA
- RNA
- SOURCE

Sophia Ananiadou

ATR Tutorial
Generating a decision tree(1)

• C4.5 Programs for Machine Learning (Quinlan, R. J. 1993)
  – One of the modules to generate a decision tree from training data
  – Definition of features needs to be given

• Training data
  – created from manually tagged documents
Generating a decision tree(2)

• Features currently used

1. Part-of-speech information
   • from the MXPOST tagger (Ratnaparkhi, 1996)
     http://www.cis.upenn.edu/~adwait/statnlp.html
   • 45 categories based on those of Penn Treebank

2. Character type information (most useful)
   • 28 categories based on the shape of the word
     (ex. case, number, symbol)

3. Term lists
   • Lists for each class are prepared

• Combination of the above features can be used to identify & classify terms
Using HMM (Collier et al.)

- HMM are stochastic finite state machines adopted from speech recognition tasks.
- HMM looks for most likely sequence of classes that correspond to a word sequence e.g. interleukin-2 protein/DNA.
- To find similarities between known words (training set) and unknown words, character features are used.
  - These gave strong cues about the classes of words in molecular biology.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>GreekLetter</td>
<td>[alpha]protein</td>
</tr>
<tr>
<td>TwoCaps</td>
<td>[RelB]protein[TAR]RNA</td>
</tr>
<tr>
<td>CapsAndDigits</td>
<td>[2A]DNA</td>
</tr>
<tr>
<td>LowCaps</td>
<td>[kappaB]protein</td>
</tr>
<tr>
<td>InitCap</td>
<td>[Interleukin]protein</td>
</tr>
</tbody>
</table>
How can we use ATR in Molecular Biology?

• Populating & updating knowledge bases / ontologies
• Information Retrieval (terms as index terms, query terms)
• Information Extraction (identify salient concepts, template filling)
• Document categorisation (term clustering)
• Updating data models through text analysis (terms used as entities, relationships among terms)
• Visualisation of relations between concepts (context information)
• Semantic annotation of texts for biological functions etc...