Information Extraction Seminar

Current Trend in Information Extraction

Biological Named Entity Recognizer

Achmad Yani
Why NER?

1. Key part of Information Extraction system
2. Robust handling of proper names essential
3. Pre-processing for different classification levels
4. Information filtering
5. Information linking
NER Process

- Phase 1: Identification of proper names in texts
- Phase 2: Classification into Categories
- Phase 3: Other common tasks

Current Trend in Information Extraction

Computational Linguistics and Phonetics Department, Saarland University
The Exception

NE is not event recognition.

NE does not create templates, nor perform co-reference.

NE is not just matching text strings with pre-defined lists of names.

NE is not easy!
Category Problem: METONYMY

- Person vs. Artefact
- Organisation vs. Location
- Company vs. Artefact
- Location vs. Organisation

Examples:
- "The ham sandwich wants his bill." vs. "Bring me a ham sandwich."
- "England won the World Cup" vs. "The World Cup took place in England."
- "shares in MTV" vs. "watching MTV"
- "she met him at Heathrow" vs. "the Heathrow authorities"
Basic Problems in NE

- **Variation of NEs**
  - e.g. John Smith, Mr Smith, John.

- **Ambiguity of NE types**
  - John Smith (company vs. person)
  - Washington (person vs. location)

- **Ambiguity with common words,**
  - e.g. “may”
Current Trend in Information Extraction

NE System Architecture

1. Documents
2. Format Analysis
3. Tokeniser
4. Gazetter
5. NE Grammar
6. NE's
Information Extraction Seminar

BIOMEDICAL NER
Results: We have determined the crystal structure of a triacylglycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/beta-hydrolase fold and a catalytic triad comprising of residues Ser87, His286 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site. This is in contrast to the structures of PgL and Pet in which the active site is buried under a closed or partially opened 'lid', respectively.
Motivation

- Large volume of biological data in the literature.
- Continuously updated knowledge
Current Trend in Information Extraction

Criteria for Automated Intelligent Text Analysis

A System must identify biological entities
Eg. Gene, protein, chemical, cell and organism name.

Handling of unknown words
Special case: long compounding of word sequences

Criteria of Systems
System Approach

- Simbolic Approach
- Based on a set of manually developed rules.
- Not based on machine learning and statistical methods.
Current Trend in Information Extraction

Protein/Gen Parts
Correspond to protein/gene parts.
Examples: MH2 domain, and lysine residue.

Protein/Gen
Correspond to names of proteins or genes

Chemical
Examples: Indomethacin, N-methylformamide, and suberoylanilide hydroxamic acid.

Chemical Parts
Terms, like methyl groups, that correspond to parts of chemicals.

General
Terms that can be classified as belonging to more than one class but can not be classified as belonging to one of them.

Source
Terms that represent source terms including cells, cell parts and organisms.
Text Processing

- Text Source: 55 Medline abstracts
- Split into sentences
- Tokenized
- Part of speech tagged using Brill’s tagger.
Brill’s tagger Algorithm

- Known words (in vocabulary): assigning the most frequent tag associated to a form of the word

- Unknown words (out of vocabulary):
  - Proper noun if capitalised and simple noun else
  - Learning or guessing rules on the same basis as contextual rules.
Brill’s tagger Rules

- Lexical rules are used for the initialisation, and contextual rules are used to correct the tags.

- **Lexical rules**: \( \text{word} \rightarrow \text{tag} \) IF Condition  
  (example: identification of *suffixes* like "-tion")

- **Contextual rules**: \( \text{tag}_1 \rightarrow \text{tag}_2 \) IF Condition  
  (example: "preceding/following tag is \(X\)",  
  "preceding/following word is \(w\)"")
Design of the System

1. Identifying abbreviations
2. Core Terms and Functional Terms
   - C-Term Recognition
   - Extraction of F-Terms
3. Concatenation and Extension Rules
Identifying abbreviations

Pattern:

- *The abbreviation occurs in which the original term is followed by the abbreviation within the parenthesis.*
- E.g., *testosterone repressed prostrate message 2 (TRPM-2)*
C-Term Recognition

Types of C-Terms:

- **General c-terms**
  - Appropriate categorization is not possible.

- **Protein c-terms**
  - contain some information that can be associated with protein/gene names

- **Chemical c-terms**
  - contain some information that can be associated with chemical names

- Base on IUPAC conventions in naming chemicals.
- Exploit various morphological features and suffixes.
“We have developed a class of HDAC inhibitors, such as suberoylanilide hydroxamic acid (SAHA), that were initially identified based on their ability to induce differentiation of cultured murine erythroleukemia cells.”

- **Suberoylanilide hydroxamic acid (SAHA)** is a chemical, which inhibits HDAC an enzyme. The suffix –ic followed by acid helps in identifying the two words as chemical c-terms.
Example 2

“Polar organic solvents such as methanol or N-methylformamide inactivate lipases.”

- Methanol and N-methylformamide are both chemical names. These are first identified as chemical c-terms because they contain the chemical root forms methyl and meth.
### Word features to capture the c-terms

<table>
<thead>
<tr>
<th>S.No</th>
<th>Word Feature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Caps only</td>
<td>CBP</td>
</tr>
<tr>
<td>2</td>
<td>More than one Cap</td>
<td>hCG</td>
</tr>
<tr>
<td>3</td>
<td>letters and Digits</td>
<td>TRPM2, H4, p53</td>
</tr>
<tr>
<td>4</td>
<td>Single Cap</td>
<td>Aspirin, Asp1</td>
</tr>
<tr>
<td>5</td>
<td>Terms having special symbols (/, -, etc)</td>
<td>IL-7 etc</td>
</tr>
</tbody>
</table>
Extraction of Functional Terms

- Functional words are helpful for:
  - locating biological terms
  - categorizing the terms properly.
## Description of Functional Terms

<table>
<thead>
<tr>
<th>CLASS</th>
<th>FUNCTIONAL WORDS</th>
<th>DESCRIPTION</th>
<th>EXAMPLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein/Gene</td>
<td>Receptor, protein, factor, gene etc</td>
<td>Protein/Gene/RNA</td>
<td>CREB binding protein</td>
</tr>
<tr>
<td>Protein/Gene parts</td>
<td>Motif, domain, promoter, etc</td>
<td>Parts of Gene/Protein/RNA</td>
<td>MH2 domain</td>
</tr>
<tr>
<td>Chemicals</td>
<td>Steroid, drugs, etc.</td>
<td>Lipid, steroid, organic, inorganic compounds, and carbohydrates</td>
<td>Tertbutyldimethylsilyoxyan drost-4-ene steroid (9)</td>
</tr>
<tr>
<td>Chemical parts</td>
<td>Radical, ions, groups etc.</td>
<td>Chemical radicals, inorganic and organic ions etc</td>
<td>acetyl groups, methyl groups etc</td>
</tr>
<tr>
<td>Source terms</td>
<td>cells, cell lines, phage</td>
<td>Cell, Organisms, cell parts etc</td>
<td>MCF-10F cells</td>
</tr>
<tr>
<td>General terms</td>
<td>Mutants, molecules</td>
<td>Can belong to any of the above category,</td>
<td>asf1mutants</td>
</tr>
</tbody>
</table>
Current Trend in Information Extraction

Concatenation and Extension Rules

❖ First Rule:
  - Two f-terms that are next to each other can be combined into one.

❖ Second Rule:
  - If one of the f-terms is of general category and the other is not of general category then the latter’s category is adopted.

❖ Example:
  - *protein complex* identified as being of category protein while the f-term complex is of general category.
3rd Rule (Concatenation of 2 f-terms)
- when both f-terms being concatenated are not general then the category of the one on the right is adopted.
- Example: characterizes protein fragment as protein part since the two individual f-terms are of protein and protein part category respectively.

4th Rule (Concatenation of a c-term and f-term)
- When f-term is to the right of the c-term, it will determine the type.
- Example:
- In SIR3 Protein fragment, as argued above, protein fragment is a f-term phrase of type protein part. SIR3 is of course a general c-term. Together they form a protein part.
5th Rule (Concatenation of two c-terms)

- *when both c-terms being concatenated are not general then the category of the one on the right is adopted.*
- *H4 acetyltransferase* is examples of concatenation of c-terms. The right c-term provides the category information in the first case.
6th Rule (chemical c-term concatenation)

- The first involves a chemical c-term which combined with a term to its right is no longer of chemical category.

- *Xanthine* is a chemical core because of the chemical root *Xanth*. However, the concatenated phrase *Xanthine oxidase* is designated a protein category c-term phrase because of the nature of the c-term to its right.
Two non-adjacent terms as long as every word in between them is a noun is considered as an adjective or a numeral.

- Example: _CREB binding protein_ and _MOZ leukemia gene_ where the leftmost and rightmost words are already marked (as c-term and f-term) but the one in the middle is not.
Rule to Drop Annotated Phrases

- For example, a f-term that is not extended to left or right is dropped. Thereby the single word protein by itself is not considered a name.
55 Medline abstracts, hand annotated and associated the categories with each marked name.

For the pure name *detection* task (i.e., without considering the assignment of categories):

- The precision : 90.39%
- Recall          : 95.64%
- F-measure       : 92.94%
### Disambiguation – Preliminary version

<table>
<thead>
<tr>
<th>S. No</th>
<th>Total</th>
<th>Terms Disambiguated</th>
<th>Precision</th>
<th>Recall</th>
<th>F-mes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein/ Gene</td>
<td>291</td>
<td>104</td>
<td>93.69%</td>
<td>34.44%</td>
<td>50.37%</td>
</tr>
<tr>
<td>Protein/ Gene parts</td>
<td>95</td>
<td>43</td>
<td>95.56%</td>
<td>43.33%</td>
<td>59.62%</td>
</tr>
<tr>
<td>Chemical</td>
<td>154</td>
<td>116</td>
<td>92.06%</td>
<td>73.42%</td>
<td>81.69%</td>
</tr>
<tr>
<td>Chemical parts</td>
<td>13</td>
<td>8</td>
<td>100.00%</td>
<td>61.54%</td>
<td>76.19%</td>
</tr>
<tr>
<td>Source terms</td>
<td>36</td>
<td>29</td>
<td>96.66%</td>
<td>80.56%</td>
<td>87.88%</td>
</tr>
<tr>
<td>Protein + Protein parts</td>
<td>401</td>
<td>147</td>
<td>94.23%</td>
<td>40.05%</td>
<td>56.21%</td>
</tr>
</tbody>
</table>

*Computed values are based on specific metrics and calculations.*

**Source:** Computational Linguistics and Phonetics Department, Saarland University
PostProcessing Rules

- Come up with a list of help-words (h-terms) that are like f-terms in that they provide clue about the category but unlike f-terms are not considered part of the name.
### Disambiguation - after Post Processing

<table>
<thead>
<tr>
<th>S. No</th>
<th>Total</th>
<th>Terms Disambiguated</th>
<th>Precision</th>
<th>Recall</th>
<th>F-mes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein/Gene</td>
<td>291</td>
<td>189</td>
<td>96.43%</td>
<td>62.58%</td>
<td>75.86%</td>
</tr>
<tr>
<td>Protein/Gene parts</td>
<td>95</td>
<td>88</td>
<td>97.78%</td>
<td>88.89%</td>
<td>92.75%</td>
</tr>
<tr>
<td>Chemical</td>
<td>154</td>
<td>136</td>
<td>93.15%</td>
<td>86.08%</td>
<td>90.86%</td>
</tr>
<tr>
<td>Chemical parts</td>
<td>13</td>
<td>11</td>
<td>100.00%</td>
<td>84.62%</td>
<td>91.67%</td>
</tr>
<tr>
<td>Source terms</td>
<td>36</td>
<td>35</td>
<td>97.22%</td>
<td>97.22%</td>
<td>97.22%</td>
</tr>
<tr>
<td>Protein + Protein parts</td>
<td>401</td>
<td>277</td>
<td>96.85%</td>
<td>69.08%</td>
<td>80.64%</td>
</tr>
<tr>
<td>Term</td>
<td>Pre-Processing</td>
<td>Post-Processing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------</td>
<td>-----------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein</td>
<td>187</td>
<td>102</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein/Gene Parts</td>
<td>52</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>38</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical Parts</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source Terms</td>
<td>7</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>289</strong></td>
<td><strong>130</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Current Trend in Information Extraction

Final Performance

<table>
<thead>
<tr>
<th>Pre-Processing</th>
<th>Post-Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>78.61</td>
<td>96.26</td>
</tr>
<tr>
<td>86.54</td>
<td>91.9</td>
</tr>
</tbody>
</table>

Computational Linguistics and Phonetics Department, Saarland University
References


- “Introduction to Named Entity Recognition”, University of Sheffield (http://gate.ac.uk/talks/stupidpoint/diana-fb.ppt)


Thank You!