BIOADI - A Machine Learning Approach to Identify Abbreviations and Definitions in Biological Literature

Kuo, CJ[1] Ling, MHT[2,4] Lin, KT[1,3] Hsu, CN[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan

[2]School of Chemical and Life Sciences, Singapore Polytechnic, Republic of Singapore

[3]Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan

[4]Department of Zoology, The University of Melbourne, Parkville, Victoria, Australia

InCoB Singapore 2009

Background

- Abbreviation recognition
  - Find pairs of long forms (LF) and short forms (SF) of terms in text
- Gene/Protein names
  - Phosphoinositide 3'-kinase (PI3K)
  - Mitogen activated protein kinase (MAP kinase or MAPK)
- Serves as precursor to other applications
  - Extracting protein-protein interactions
  - Search engine development

Related Work

SYSTEM	YEAR	PRECISION	RECALL	F-SCORE
Stanford University Abbreviation Server	2002	95.00%	75.00%	83.82%
AbbRE	2002	96.00%	70.00%	80.96%
Schwartz and Hearst	2003	96.00%	82.00%	88.45%
SaRAD	2004	95.00%	85.00%	89.72%
Sohn et al.	2008	96.50%	83.20%	89.36%

- Testing was done on different modified corpus
- Difficult to compare between systems
- Can we do better?

This study

- Focus on 3 types of SF-LF pairs:
  - SF (LF) or SF [LF]
  - LF (SF) or LF [SF]
  - (SF, LF) or (LF, SF) or [SF, LF] or [LF, SF]

Feature extraction for model training

- Transform SF-LF pair data to a informative feature vector
- Select 4 sets of features
  - String morphological features
  - LF tokens
  - Numeric features between SF and LF
  - Contextual features of SF-LF pair

String morphological features

- Represent the literal information and character properties of SF or LF respectively

Example

Is the first letter of the string uppercase?

$$\text{Heat shock protein} \longrightarrow \text{TRUE}$$

Example

Does the string contain numbers?

$$\text{CA5} \longrightarrow \text{TRUE}$$

LF tokens

- Represent token information of LF
- Tokenize each LF with space and punctuations as delimiters
- Take each token as a binary feature
- Also apply token bi-grams as binary contextual features

Example

...rotease-binding property of alpha1-protease inhibitor (alpha1PI) was destroyed by acid...

- token features: {alpha1, protease, inhibitor}
- bi-grams features: {alpha1-protease, protease-inhibitor}

Numeric features between SF and LF

- Describe the mapping of SF letters and LF letters
- Calculate the character usage when a SF is abbreivated from a LF

Example

cyclic AMP (cAMP)

- Possible character mapping: {cyclic AMP, cyclic AMP}
- Character usage: $(4/4) * 100\% = 100\%$

Contextual features of SF-LF pair

- Represent contextual information of each potential pair
- 2 tokens precede the pair
- These tokens act as binary features respectively and together

Example

The role of protein kinase A (PKA) in the release of amylase from...

- token features: {role, of, in, the}
- bi-grams features: {role-of, in-the}

Model training and testing

- Implement 4 learning algorithms for this task
  - Support Vector Machine (LIBSVM)
  - Naive Bayes (MALLET)
  - Logistic Regression (MALLET)
  - Monte-Carlo Sampling Logistic Regression (Mallet)
- (Post-processing) Set a ruled-based filter to improve the precision
  - Remove false-positives which are usally in outputs

Testing BIOADI

- We annotate a corpus of 1200 abstracts from BioCreative II gene normalization dataset (BIOADI corpus)
- AB3P corpus (Sohn et al., 2008)
- Use one corpus for training, the other for testing
- Compare with
  1. Schwartz and Hearst (2004)
  2. Sohn et al. (2008)

Results (Precision and Recall)

- Training corpus: AB3P corpus/Testing corpus: BIOADI corpus

SYSTEM	YEAR	PRECISION	RECALL	F-SCORE
Schwartz and Hearst	2003	94.16%	77.66%	85.12%
Sohn et al.	2008	94.82%	78.32%	85.78%
This study (BIOADI)	2009	93.52%	79.95%	86.20%

- Training corpus: BIOADI corpus/Testing corpus: AB3P corpus

SYSTEM	YEAR	PRECISION	RECALL	F-SCORE
Schwartz and Hearst	2003	95.00%	78.83%	86.13%
Sohn et al.	2008	97.01%	83.56%	89.79%
This study (BIOADI)	2009	95.86%	84.64%	89.90%

Results (Computational Speed)

	TESTING SIZE OF ABSTRACTS			
SYSTEM	1200	1250	2450	5000
Schwartz and Hearst	0.873	0.897	1.598	3.138
Sohn et al.	159.292	135.254	292.343	630.917
This study (BIOADI)	13.355	13.316	25.059	45.506

All time in seconds

Contribution of this study

- An abbreviation recognizer with good performace (F-score), reasonable speed and cross-platform
  - Available for download
- A website for searching abbreviations
  - Linking to PubMed abstracts
  - http://bioagent.iis.sinica.edu.tw/BIOADI/
- Releasing soon...
  - Entire set of 1.687 million abbreivation pairs from 17.5 million abstracts
  - BIOADI corpus


BIOADI
Biomedical Abbreviation Definition Identifier
SF-LF Search
SF-LF Identification
User Manual
Blog
Download
Contacts
Category
Description
Download
Tool
Perl script for fetching PubMed abstracts
click here
Tool
Off-line Abbreviation Recognition Tool
click here
Data
All Identified Abbreviation Pairs. (will release after publication)
release soon
AIIA lab
2008-2009 c Institute of Information Science, Academia Sinica. 128, Sec. 2, Academia Rd., Nankang, Taipei, Taiwan.
CSS Template is retrieved from xhtmldev.net

Thanks for your attention!