Alignment-HMM-based Extraction of Abbreviations from Biomedical Text

Dana Movshovitz-Attias and William W. Cohen

Carnegie Mellon University
June 8, 2012

Abbreviations are Abundant in Bio-literature

- Commonly used for
  - Proteins/Genes/Molecules
  - Diseases
  - Experimental methods and other common terms

- Definitions change with context
  - APC matches over 100 unique abbreviations in MEDLINE

Two Main Uses of Abbreviations

- Common
  - < AIDS, acquired immunodeficiency syndrome >
  - < DNA, deoxyribonucleic acid >
  - Often not explicitly defined
  - Widely accepted as synonyms
  - More common in the abbreviated form

- Dynamic
  - Defined by the author
  - May be specific to one article
  - May overlap with other dynamic abbreviations
    - APC

Yu, H., Hripcsak, G., and Friedman, C. Mapping abbreviations to full forms in biomedical articles. JAMIA

Task

- Extract dynamic abbreviations explicitly defined in the text

> We earlier reported that when phenylalanine ammonialyase (PAL) activity in radish seedlings was inhibited by the competitive inhibitor 2-aminoindan-2-phosphonic acid (AIP), … The syringyl to guaiacyl (S/G) ratio in the lignin of AIP-grown plants, as determined by alkaline cupric oxidation and from Fourier-transform infrared (FT-IR) spectra, was higher in cotyledons, …

- Output
  - Abbreviation definition pair
    〈 *short form, long form* 〉
  - Alignment
  - Score

1. < PAL, phenylalanine ammonia-lyase >
2. < AIP, 2-aminoindan-2-phosphonic acid >
3. < S/G, syringyl to guaiacyl >
4. < FT-IR, Fourier-transform infrared>

Types of Abbreviations

- Standard acronyms
  < AMS, Associated Medical Services >
- Missing letters
  < EDI-2, Eating Disorders Inventory >
- Chemical formulas
  < MTIC, 5-(3-N-methyltriazen-1-yl)-imidazole-4-carboxamide >
- Substitutions: word ↔ symbol
  < NaB, sodium butyrate >
- Out-of-order
  < NTx, cross-linked N-telopeptides >
- Synonyms
  < anti-Tac, antibody to the alpha subunit of the IL-2 receptor >

1. Schwartz and Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. PSB.
2. Chang, Schutze, and Altman. 2002. Creating an online dictionary of abbreviations from medline. JAMIA.

Extraction Method

1. Parse text and extract *candidate* definitions
2. Align candidate definitions
3. Predict abbreviation

Extraction Method

1 Parse text and extract *candidate* definitions

anti-sperm antibodies were studied by indirect mixed anti-globulin reaction test (MAR)

⟨ MAR, by indirect mixed anti-globulin reaction test ⟩

- long form (short form)
- short form (long form)
- Patterns of multiple abbreviations
  - "anti-sperm (ASA), anti-phospholipid (APA), and antizonal (AZA) antibodies"

Extraction Method

1 Parse text and extract *candidate* definitions

anti-sperm antibodies were studied by indirect mixed anti-globulin reaction test (MAR)

⟨ MAR, by indirect mixed anti-globulin reaction test ⟩

- Length of long form is estimated

Schwartz and Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. PSB.

Extraction Method

2 Align candidate definitions

〈 MAR, by indirect mixed anti-globulin reaction test 〉

Alignment-HMM suited for abbreviation extraction

2 Alignment HMM

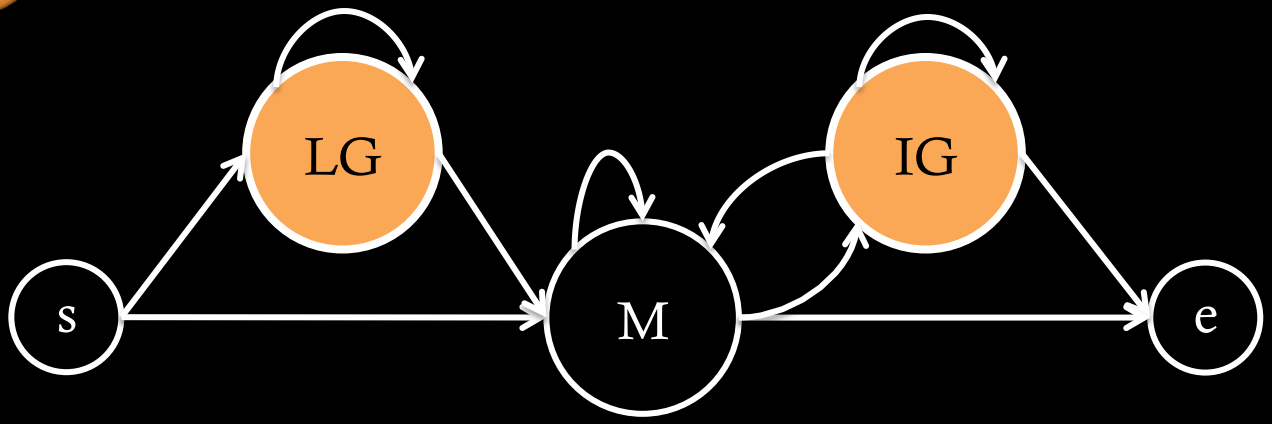
- Model an alignment of long and short form
- Series of edit operations
- Edit operations are emitted by an HMM

Operation	Short form	Long form
Deletion	ε	Alpha-numeric char
Match	character	(partial) word
Substitution	1	one
Substitution	Na	Sodium

- Previously used for string edit distance

Ristad, E.S. and Yianilos, P.N. Learning string-edit distance. Pattern Analysis and Machine Intelligence .
Bilenko, M. and Mooney, R.J. Adaptive duplicate detection using learnable string similarity measures. ACM.

2 Alignment HMM



- Affine gap cost model

```
cost(gap) = start + extend · length
```

- Leading (LG) and inner gaps (IG)
- Unsupervised: EM training on candidates
- We get P(align) with Viterbi

Extraction Method

2 Align candidate definitions

〈 MAR, by indirect mixed anti-globulin reaction test 〉

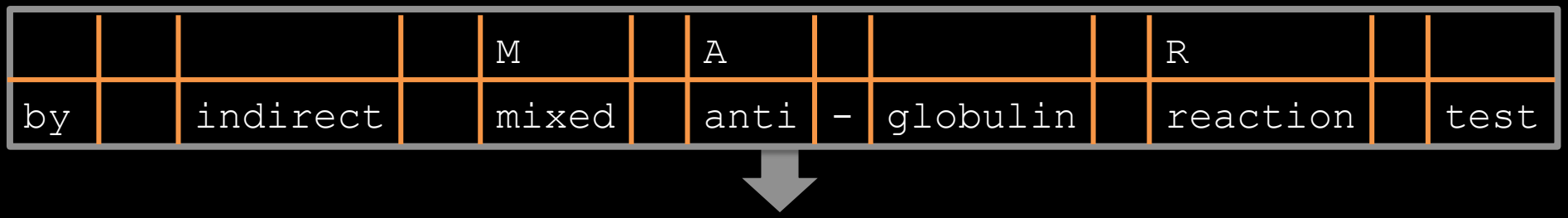
Artifact of extraction method

Top: Short form
Bottom: Long form

Quality of alignment

Extraction Method

3 Predict abbreviation



〈 MAR, mixed anti-globulin reaction test 〉

- Abbreviations are predicted only from valid alignments

Popular Extraction Algorithms

- SH (Schwartz and Hearst, 2002)
  - Widely used
  - Fast and simple rule-based algorithm
  - Hard to extend
  - Relatively Low recall
- Chang (Chang et al., 2002)
  - Alignment-based (Longest Common Subsequence)
  - Feature vector is extracted from the alignment
  - Used to train binary logistic regression
  - Processing of alignment leads to slow algorithm

1. Schwartz and Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. PSB.
2. Chang, Schutze, and Altman. 2002. Creating an online dictionary of abbreviations from medline. JAMIA.

Comparison with Popular Methods

Model	D (average %)			V (%)		
	P	R	F1	P	R	F1
Alignment HMM	98	**93**	**96**	95	91	**93**
SH	96	88	91	**97**	83	89
Chang $_{0.88}$	**99**	46	62	**97**	47	64
Chang $_{0.14}$	94	89	91	95	91	**93**
Chang $_{0.03}$	92	91	91	88	**93**	90
Chang $_{0}$	49	92	64	53	**93**	67

Metrics

$$P = \frac{\text{correct predicted abbreviations}}{\text{all predicted abbreviations}}$$

$$R = \frac{\text{correct predicted abbreviations}}{\text{all correct abbreviations}}$$

Four thresholds over regression score

MEDSTRACT (Development) 483 abbreviations

PubMed Sample (Validation) 76 abbreviations

Comparison with Popular Methods

Model	D (average %)			V (%)		
	P	R	F1	P	R	F1
Alignment HMM	98	**93**	**96**	95	91	**93**
SH	96	88	91	**97**	83	89
Chang $_{0.88}$	**99**	46	62	**97**	47	64
Chang $_{0.14}$	94	89	91	95	91	**93**
Chang $_{0.03}$	92	91	91	88	**93**	90
Chang $_{0}$	49	92	64	53	**93**	67

Metrics

$$P = \frac{\text{correct predicted abbreviations}}{\text{all predicted abbreviations}}$$

$$R = \frac{\text{correct predicted abbreviations}}{\text{all correct abbreviations}}$$

1. Highest F1 on both data sets
2. Comparable results with Chang 0.14
   - No need to select a threshold
   - Slow due to extra alignment processing
3. Recall is lower than precision – could improve using more edit operations

Main Advantages

1. High performance on standard dataset
2. Naturally generalizable to genres of abbreviations, using edit operations.
3. Associates probability with predicted definition
4. Unsupervised

dma@cs.cmu.edu
www.cs.cmu.edu/~dmovshov