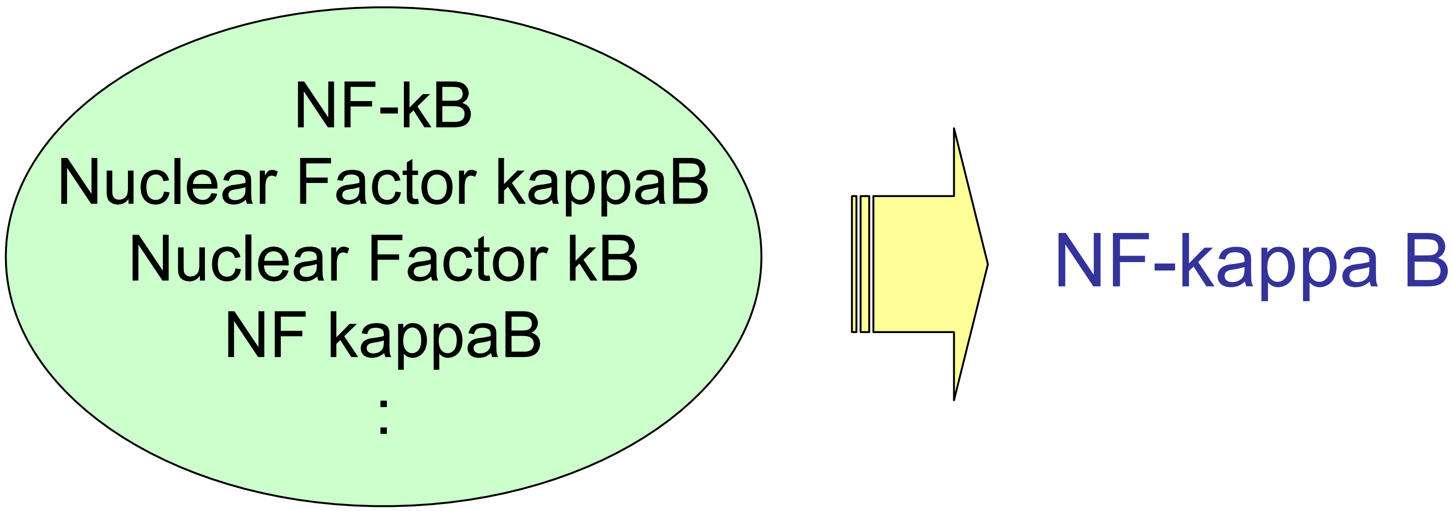
A Machine Learning Approach to Acronym Generation

Yoshimasa Tsuruoka[1], Sophia Ananiadou[4], and Jun'ichi Tsujii[123]

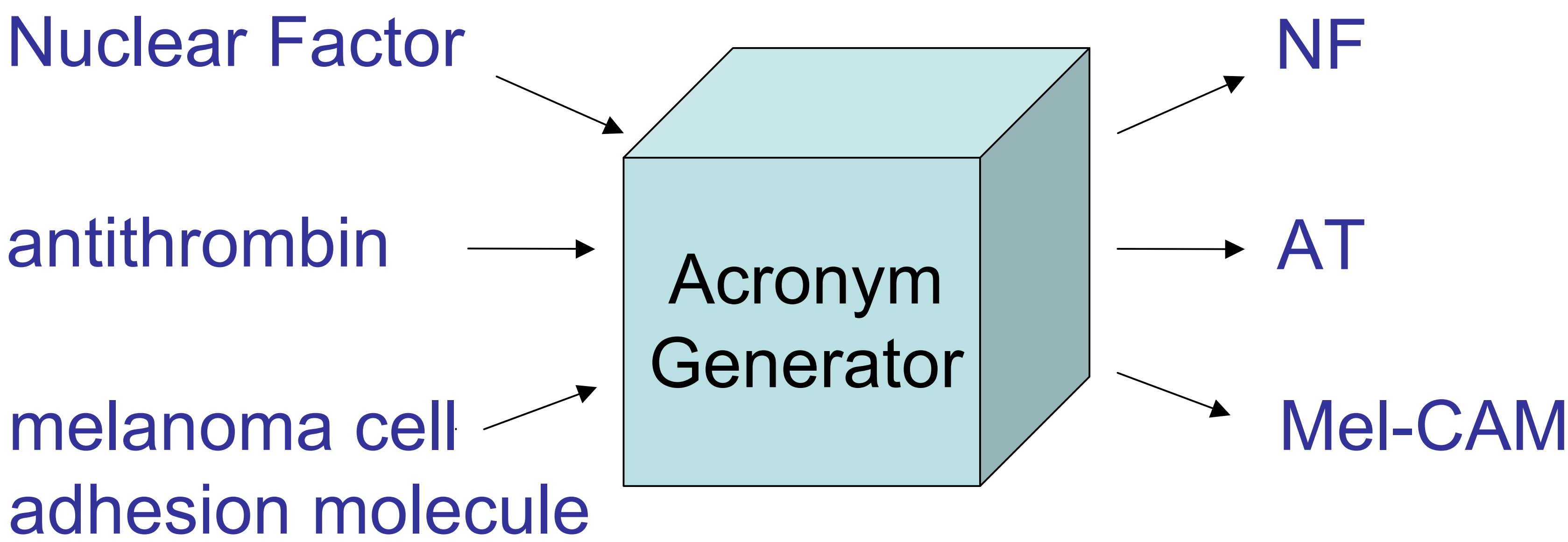
[1]Japan Science and Technology Agency
[2]The University of Tokyo
[3]The University of Manchester
[4]National Centre for Text Mining

Variation in Biomedical Terms



- Term variation is a big obstacle in knowledge integration.
  →Internal similarity of terms (edit-distance), spelling variation generator based on a probabilistic model, etc.
- Acronyms constitute a major source of difficulties

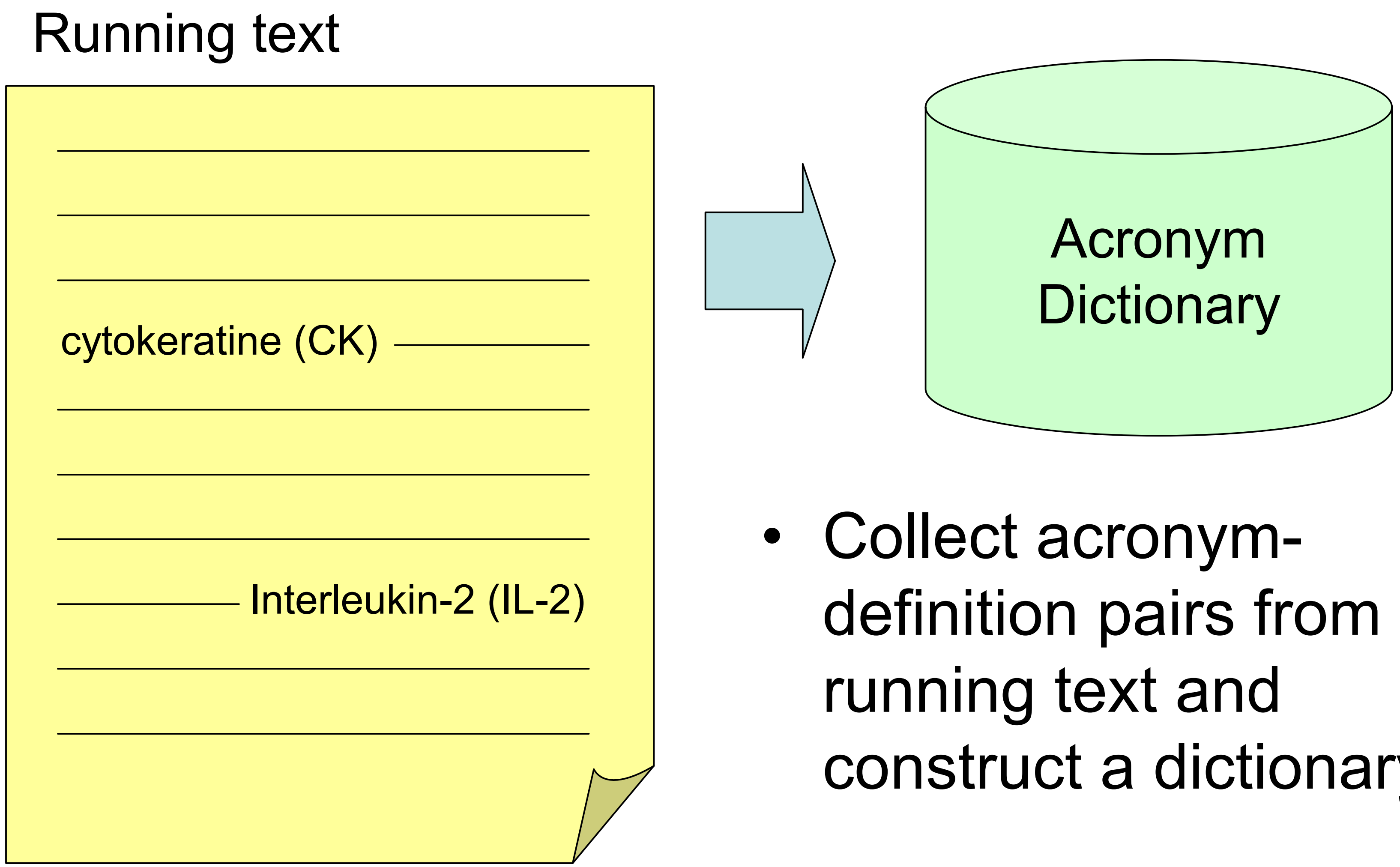
Acronym Generation



– The system generates possible acronyms from a given expanded form.

Term similarities for applications such as term clustering, term variation generator, etc.

Dictionary-Building Approaches



- Collect acronym-definition pairs from running text and construct a dictionary.

Problems of Dictionary-Building Approaches

- Coverage
  - Limited available resources (corpora) and lack of generalization
  - Dynamic nature of terms
- Term variation in expanded forms
  - We need to address the problems of term variations in which acronyms are mixed with other variations such as spelling, lexical variations, etc.

Our approach

- Machine learning-based
  - Acronym generation as sequer
  - Probabilistic modeling
- Advantages
  - Wide coverage can be achieved by generalization.
  - Similarities can be computed in a probabilistic form.
- Drawbacks
  - Needs training data
    - Unsupervised approach (future work)

MEMM can
features tha
intuition of rule-based
methods with statistical
modeling

Collection of
weak cues

Acronym Generation as Sequence Tagging

cytokeratines

Definition	Tag
c	UPPER
y	SKIP
t	SKIP
o	SKIP
k	UPPER
e	SKIP
r	SKIP
a	SKIP
t	SKIP
i	SKIP
n	SKIP
e	SKIP
s	LOWER

CKs

Acronym
C
K
s

Sequence Tagging with MEMM

Maximum Entropy Modeling with Inequality Constraints (Kazama and Tsujii 2003, 2005)

- Smoothing effects

  Performance is better or comparable to that achieved with the use of Gaussian prior.

- Smaller model size -> quick decoding

  Ex.) POS tagging

  - Gaussian prior: 12MB
  - Inequality constraints: 1.3MB

MEMM can integrate features that reflect intuition of rule-based methods with statistical modeling

maximum entropy classifier

(model size = 60kB)

Features (1)

target letter

lactate dehydrogenase

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)

- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (2)

target letter

lactate dehydrogenase

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (3)

target letter

lactate dehydrogenase

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (4)

target letter

lactate dehydrogenase

SKIP

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (5)

target letter

lactate dehydrogenase

Uppercase? → false

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (6)

target letter

lactate dehydrogenase

2 words

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (7)

target letter

lactate dehydrogenase

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Features (8)

target letter

lactate dehydrogenase

2 10

- Letter unigrams (UNI)
- Letter bigrams (BI)
- Letter trigrams (TRI)
- Tagging history (HIS)
- Orthographic features (ORT)
- Definition length (LEN)
- Letter sequence (SEQ)
- Distance (DIS)

Training data

- Acronym-definition pairs are extracted from running text, and position information is manually added to each pair.

Acronym	Definition	Position
IM	Intestinal metaplasia	1, 12
LDH	lactate dehydrogenase	1, 9, 11
CK	cytokeratine	1, 5
CKs	cytokeratines	1, 5,12
EBV	Epstein-Barr virus	1, 9, 14
:	:	:

Experiments

- Training data
  - 1,901 acronym-definition pairs extracted from MEDLINE abstracts published in 2001.
  - A simple deterministic method (Schwartz 2003) was used for extraction.
  - Position information is semi-manually added.
- Evaluation
  - 10-fold cross validation

Generated acronyms

- For “traumatic brain injury”

Rank	Probability	String
1	0.779	TBI
2	0.062	TUBI
3	0.028	TB
4	0.019	TbI
5	0.015	TB-I
6	0.009	tBI
7	0.008	TI
8	0.007	TBi
9	0.002	TUB
10	0.002	TUbI

Generated acronyms

- For “open reading frame 1”

Rank	Probability	String
1	0.423	ORF1
2	0.096	OR1
3	0.085	ORF-1
4	0.070	RF1
5	0.047	OrF1
6	0.036	OF1
7	0.025	ORf1
8	0.019	OR-1
9	0.016	R1
10	0.014	RF-1

Generated acronyms

- For “RNA polymerase”

Rank	Probability	String
1	0.163	RNA-P
2	0.147	RP
3	0.118	RNP
4	0.110	RNAP
5	0.064	RA-P
6	0.051	R-P
7	0.043	RAP
8	0.041	RN-P
9	0.034	RNA-PM
10	0.030	RPM

Generated acronyms

- For "meta-chlorophenylpiperazine"

Rank	Probability	String
1	0.405	MCPP
2	0.149	MCP
3	0.056	MCP
4	0.031	MPP
5	0.028	McPP
6	0.024	MchPP
7	0.020	MC
8	0.011	MP
9	0.011	mCPP
10	0.010	MCRPP

Generated acronyms

- For “Toscana virus”

Rank	Probability	String
1	0.811	TV
2	0.034	TSV
3	0.030	TCV
4	0.021	Tv
5	0.019	TVs
6	0.013	T-V
7	0.008	TOV
8	0.004	TSCV
9	0.002	T-v
10	0.001	TOSV

Coverage (recall)

- Coverage achieved with top-N candidates.
  – Below top 10

ex.)

melanoma cell adhesion molecule

↓

Mel-CAM

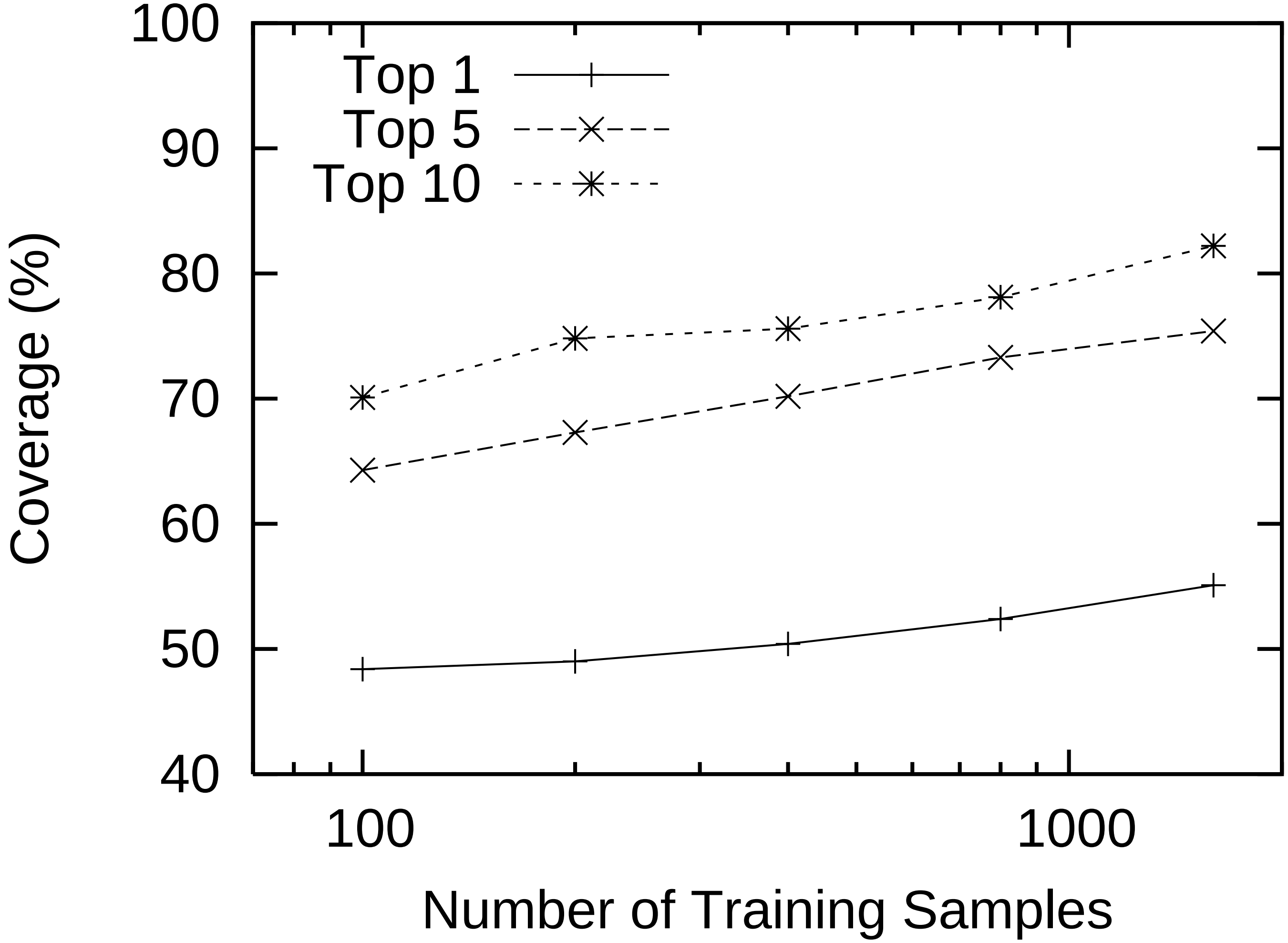
- Baseline
  – Rule-based
    - Take the initial letter of each word and capitalize them.
  – Coverage: 47.3%

Rank	Coverage
1	55.2%
2	65.8%
3	70.4%
4	73.2%
5	75.4%
6	76.7%
7	78.3%
8	79.8%
9	81.1%
10	82.2%

Effectiveness of Features

Features	Top1 Coverage	Top 5 Coverage	Top 10 Coverage
UNI	48.2%	66.2%	74.2%
UNI, BI	50.1%	71.2%	78.3%
UNI, BI, TRI	50.4%	72.3%	80.1%
UNI, BI, TRI, HIS	50.6%	73.6%	81.2%
UNI, BI, TRI, HIS, ORT	51.0%	73.9%	80.9%
UNI, BI, TRI, HIS, ORT, LEN	53.9%	74.6%	81.3%
UNI, BI, TRI, HIS, ORT, LEN, DIS	54.4%	75.0%	81.8%
UNI, BI, TRI, HIS, ORT, LEN, DIS, SEQ	55.1%	75.4%	82.2%

Learning curve



Conclusion

- Spelling variation in biomedical terms
- Acronym generation with a similarity measure
- Sequential tagging with MEMM
- Experiments
  - 1,901 acronym-definition pairs
  - Top 1 coverage: 55.1%
  - Top 5 coverage: 75.4%
- Future work
  - Unsupervised learning using acronym-definition pairs with unambiguous position information.
  - More features reflecting rule-based intuition such as specific combining forms, prefixes, suffixes, etc. and features of resultant acronyms such as consonant, vowel, etc.
  - Integration with larger systems (term variation generator, term clustering, etc)