BARY at the NTCIR-11 MedNLP-2 Task for complaints and diagnosis recognition

Yusuke Matsubara
Mizuki Morita
Koiti Hasida

Social ICT Research Center, The University of Tokyo

School of Information Science and Technology, The University of Tokyo
Social ICT Research Center

Overview

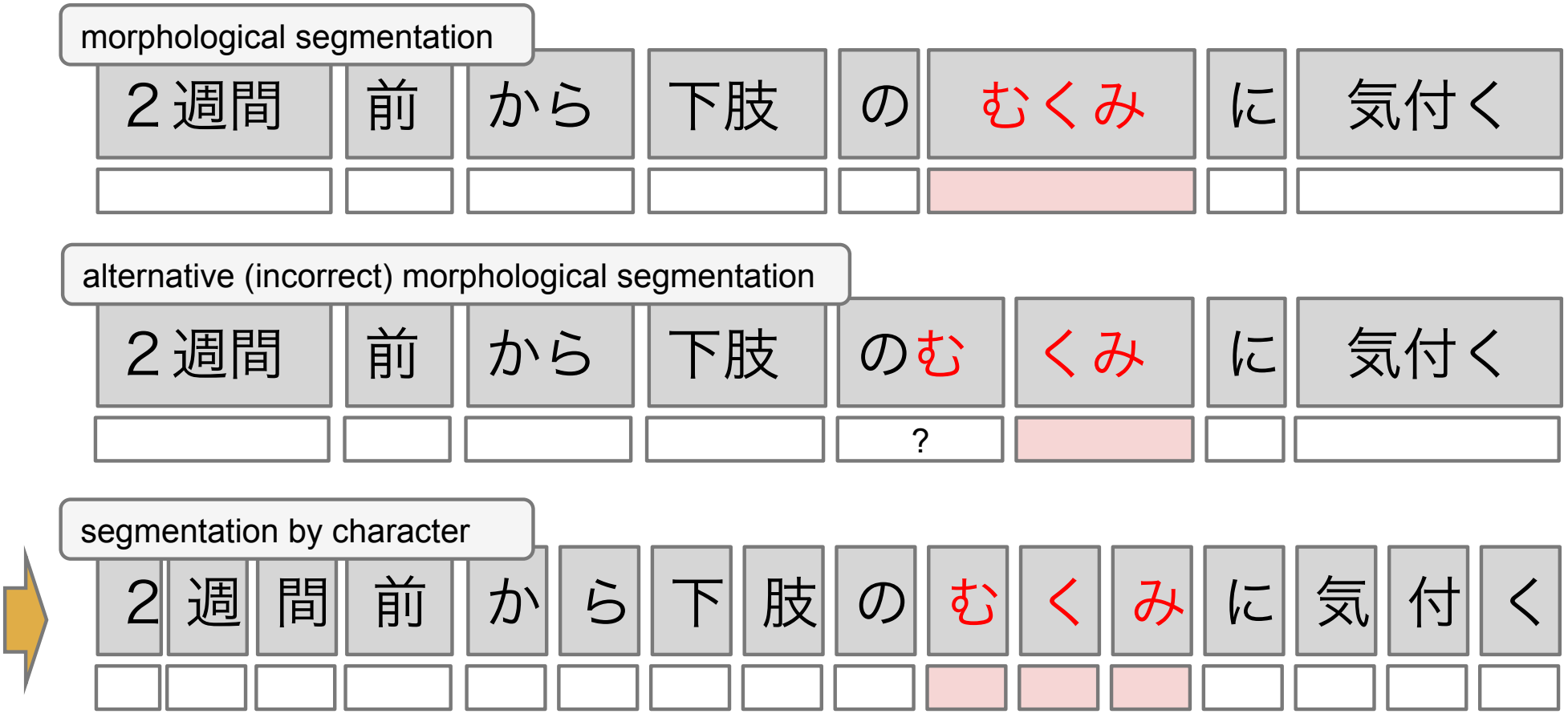
- Character-wise, two-stage, CRF-based sequence labeling
- Recognizes disease names, diagnoses with modalities, and temporal expressions
- Uses affixes extracted from the external terminology resources

Character-level labeling (BARY1-3)

２週間前から下肢のむくみに気付く

ni shu kan mae kara　kashi no mukumi ni kizuku

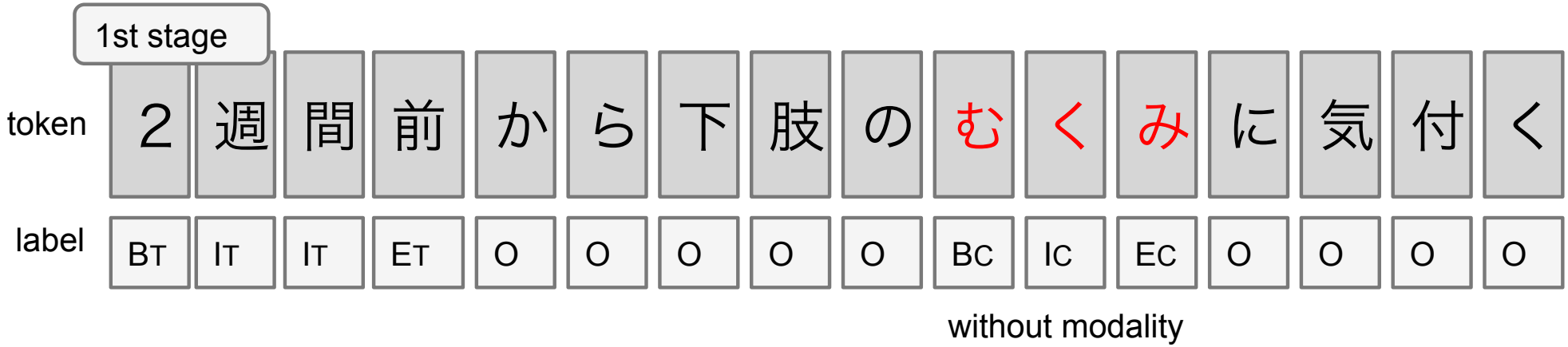
“Aware of a swelling on the lower limb since two weeks ago”

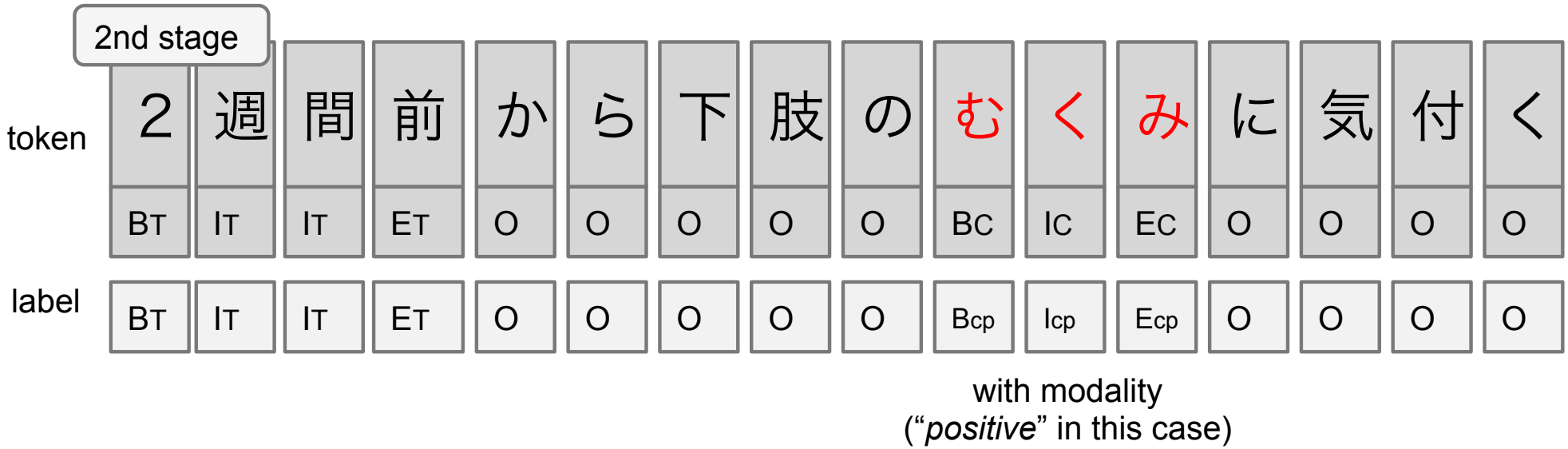


Characters over mophemes

- Pros
  - No propagation of segmentation errors
- Cons
  - Inconsistent amount of information in a fixed context window (3 kanji characters carries more than 3 latin letters)

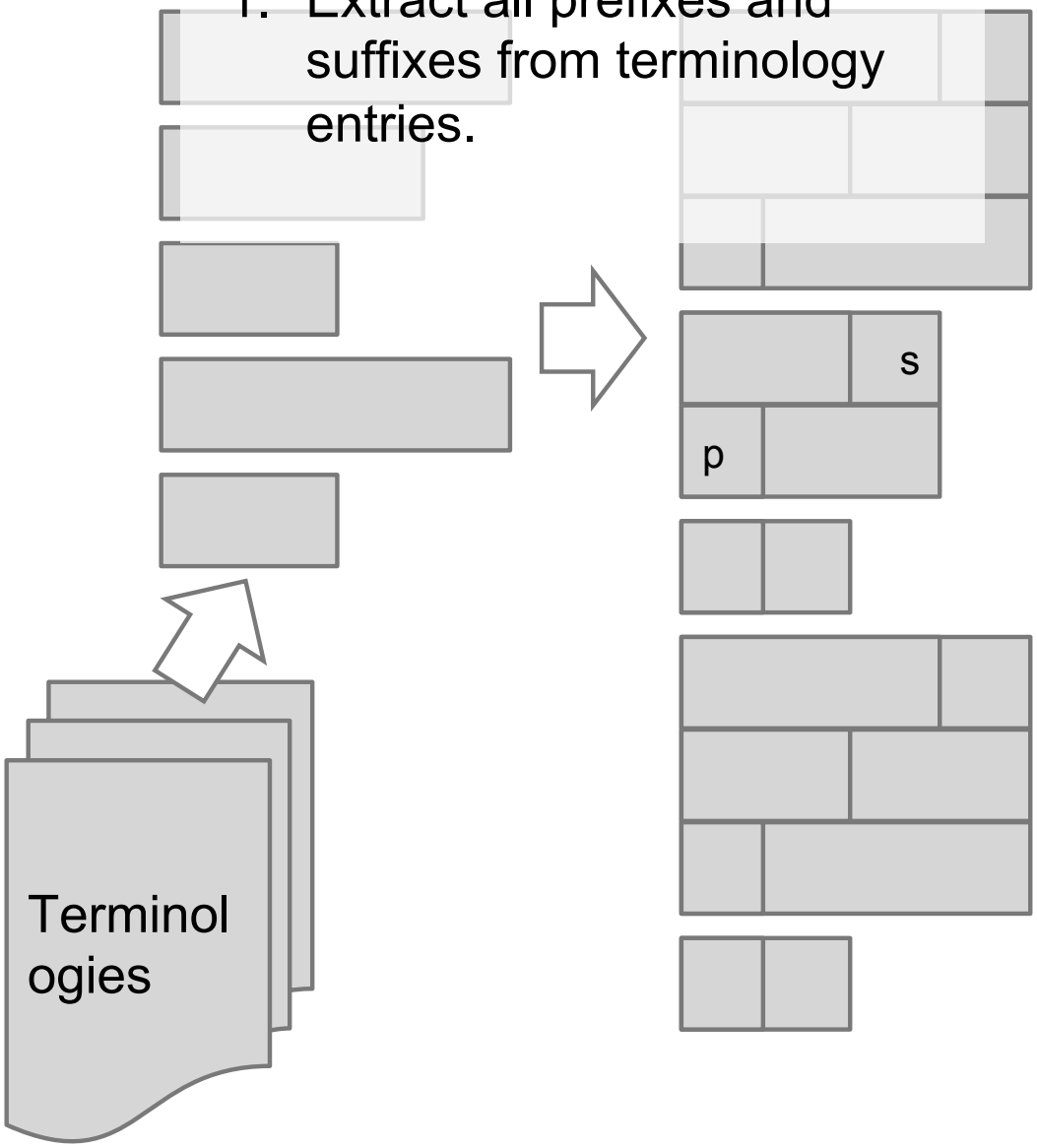
Two-stage model (BARY2,3)





Max-repeat term affixes (BARY-3)

1. Extract all prefixes and suffixes from terminology entries.



2. Count them and filter out low-frequency ones and redundant ones.

y is redundant if there exists x s.t. $y \subseteq x$ and freq(x) = freq(y).

Term affixes: preview of the results

Affix	Freq.	Source entries
高血圧 *high blood pressure, hypertension*	**16**	白衣**高血圧** *white coat* … / **高血圧**性網膜症 .... *retinopathy*
腫脹 *tumefacient*	**9**	上肢**腫脹** *upper limb* ... / 腹部**腫脹** *abdominal* ...
水疱 *bladder*	**4**	**水疱**性中耳炎 … otitis media / 非熱傷性**水疱** *burn* ….
萎縮 *atrophy*, *shrinkage*	**4**	骨**萎縮** *bone* … / **萎縮**性鼻炎 … *rhinitis*
肥厚 *incrassate, thickening, thickened*	**3**	胸膜**肥厚** *pleural* … / **肥厚**性瘢痕 … *cicatrix*
紅斑 *erythema, red patch*	**3**	遊走性**紅斑** *migrating* ... / **紅斑**症 … *disease*
感染症 *infectious disease*	**3**	細菌**感染症** *bacterial* … / 条虫**感染症** *tapeworm* ...
感染 *infection*	**3**	日和見**感染** *opportunistic* … / **感染**性鼻炎 … *rhinitis*
腹水 *ascites*	**2**	異常**腹水** abnormal ... / **腹水**症 *ascites*
肥満 *obesity*	**2**	**肥満**細胞症 ... *disease* / 小児**肥満** *infantile* ...
細菌 *bacteria*	**2**	**細菌**感染症 ... *infection* / **細菌**尿 *bacteriuria*
混濁 *cloudiness*	**2**	角膜**混濁** *corneal* ... / **混濁**尿 ... *urine*
リンパ節腫脹 *tumefacient lymph node*	**2**	肘**リンパ節腫脹** *cubital* ... / 下顎**リンパ節腫脹** *mandibular*
アルコール *alcohol*	**2**	**アルコール**性躁病 … *mania* / **アルコール**依存症 *alcoholism*

This shows affixes that coincide with <c> elements in the MedNLP-2 test set. Extracted from Hyojun Byomei Master, with a threshold of 10+ occurrences.

Other features (conventional)

Input token	Tag	Features					
2 ni (“two”)	**B-T**	**NUMBER**	**O**	**I**	2	2	Noun
日 nichi (“days”)	**I-T**	**CJK**	**O**	**I**	日	日	Noun
前 mae (“ago”)	**I-T**	**CJK**	**O**	**I**	前	前	Noun
か ka (“since”)	**I-T**	**HIRAGANA**	**O**	**I**	か	から	PostP.
ら ra	**I-T**	**HIRAGANA**	**O**	**I**	ら	から	PostP.
発 hatsu (“fever”)	**B-C**	**CJK**	**I**	**O**	発	発熱	Noun
熱 netsu	**I-C**	**CJK**	**I**	**O**	熱	発熱	Noun
。(EOS)	**O**	**KUTEN**	**O**	**O**	。	。	Punct.

Character type

Terminology (string match)

Temporal expression pattern

Character surface

Morphological analysis

Materials

- Terminological resources
  - Hyojun Byomei Master (標準病名マスター) ver 3.13
  - Shojo Shoken Master <Shintai Shoken Hen> (症状所見マスター【身体所見編】) ver 20140306
- Training data provided by the MedNLP-2 task
- Self-developed regular expression patterns

Machine learning setups

- Training using the CRF++ toolkit
- Hyperparameter tuning using 5-fold cross-validation
  - Feature 2-grams and 3-grams in a (-2,2) window
  - Regularization parameter C=1.0, frequency cutoff F=1

Experimental results: models

	Precision	Recall	F-measure
BARY1	89.44	77.41	82.99
BARY2	89.33	78.84	83.76
BARY3	89.66	78.92	83.95

Two-stage model

Affix features and pattern features

Performances on the 51 reports of the the MedNLP-2 test set.
Each system trained on the 102 reports of the training set.

Contributions of features (BARY1)

	Precision	Recall	F-measure
(1): Character surface	ALL: 73.94 C: 73.45 T: 75.80	ALL: 58.11 C: 56.45 T: 65.83	ALL: 65.04 C: 63.81 T: 70.36
(2): (1) + Character type	ALL: 88.00 C: 88.44 T: 86.01	ALL: 79.41 C: 79.57 T: 78.52	ALL: 83.46 C: 83.72 T: 82.02
(3): (2) + Morphological analysis	ALL: 89.40 C: 90.28 T: 85.70	ALL: 79.81 C: 79.90 T: 79.42	ALL: 84.28 C: 84.70 T: 82.34
(4): (3) + Medical terminology	ALL: 89.34 C: 90.08 T: 86.17	ALL: 80.00 C: 80.00 T: 80.03	ALL: 84.35 C: 84.67 T: 82.87

Results from 5-fold cross-validation using the 102 reports of the MedNLP-2 training set