NEC LABORATORIES AMERICA AT THE MEDICAL NATURAL LANGUAGE PROCESSING PILOT TASK

Pierre-François Laquerre

Christopher Malon

{pierre.francois,malon}@nec-labs.com

NEC Laboratories
America
Relentless passion for innovation

Task 1: de-identification

[Time]
8／13にはfollow upの頭部単純CTを施行し、新規病変の出現がないことを確認した。

[Time]
12月24日には低緊張性十二指腸造影を施行したとこ乳頭部直上から45mmに及ぶ十二指腸の不整
SpO298%（room air）
#2.糖尿病
失神・右眼霧視

[Age]
71歳；前立腺肥大症
NIHSSスケール18点（JCS10
1点、質問2点、従命1点、注視1点、視野1点、顔面麻痺1点、右上肢麻痺2点、右下肢麻痺3点、感覚1点

[Time]　　　　　　　　　　　　[Hospital]
4月21日頃より咳症状増悪微熱あったため、近医受診し、胸部レントゲン上右下肺野にコンソリデーシ

Our task: predict these labels in sentences

- Age: **小児期より**アトピー性皮膚炎
- Gender: 成人発症型の気管支喘息の**女性**
- Hospital: **大学病院**循環器科受診…
- Location: **■■健康管理センター**健診で胸部…
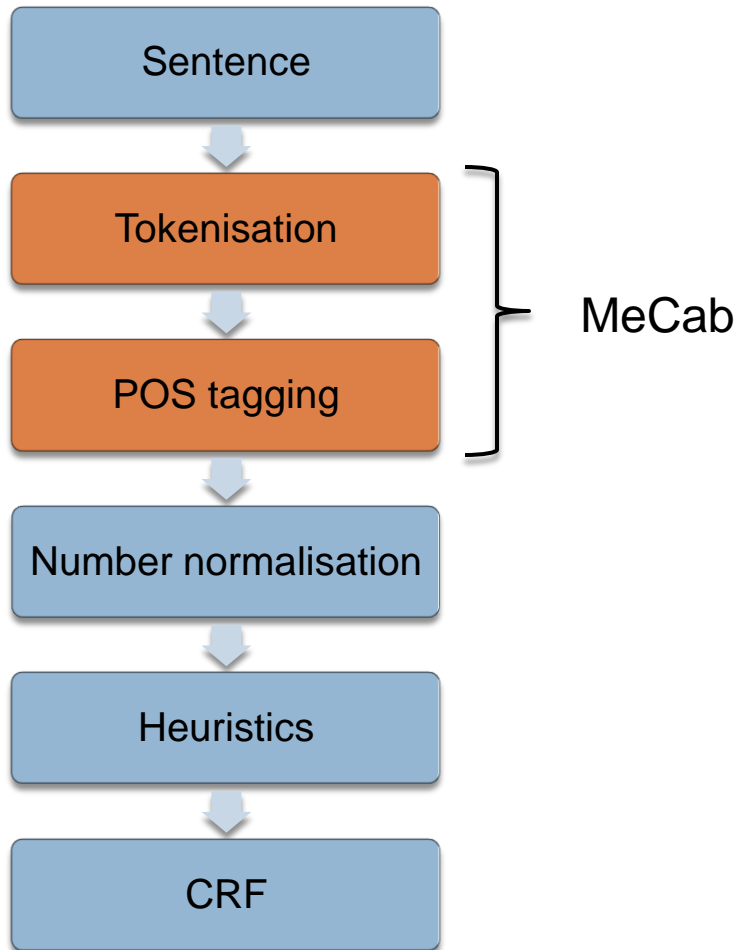- Time: **2020年頃**：顔面けいれん…

Small training set

- Age   ← 56 mentions
- Gender   ← 4 mentions
- Hospital   ← 75 mentions
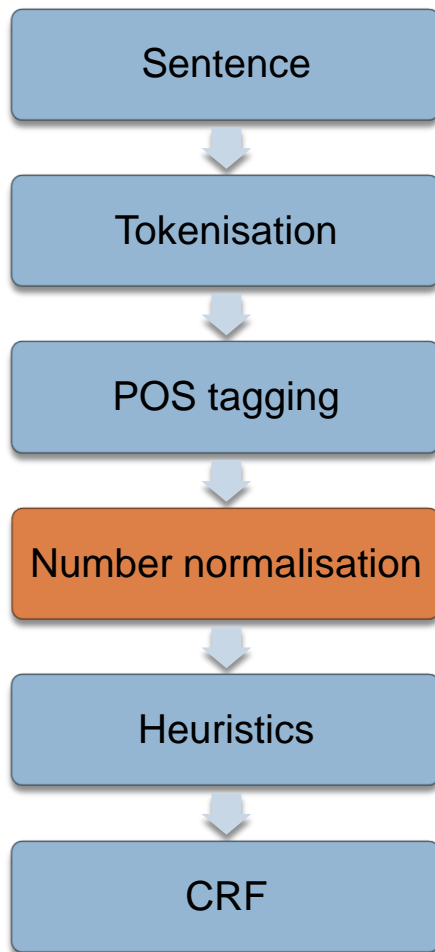- Location   ← 2 mentions
- Time   ← 355 mentions
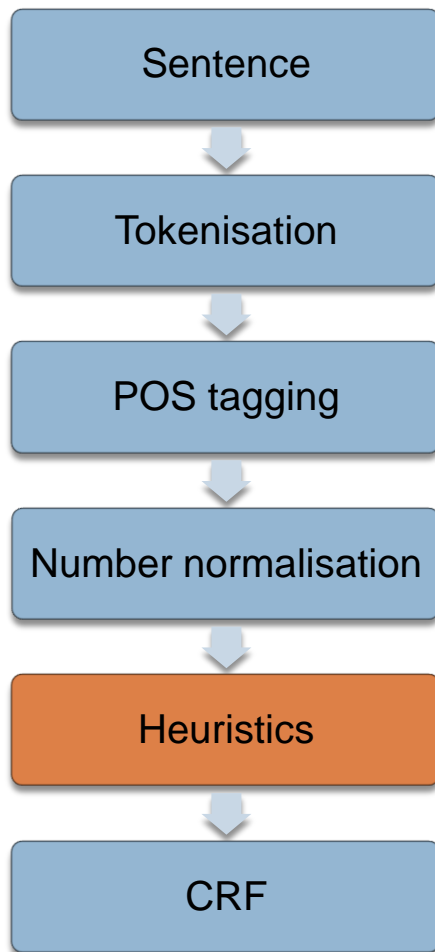
Pipeline

Sentence → Tokenisation → POS tagging → Number normalisation → Heuristics → CRF
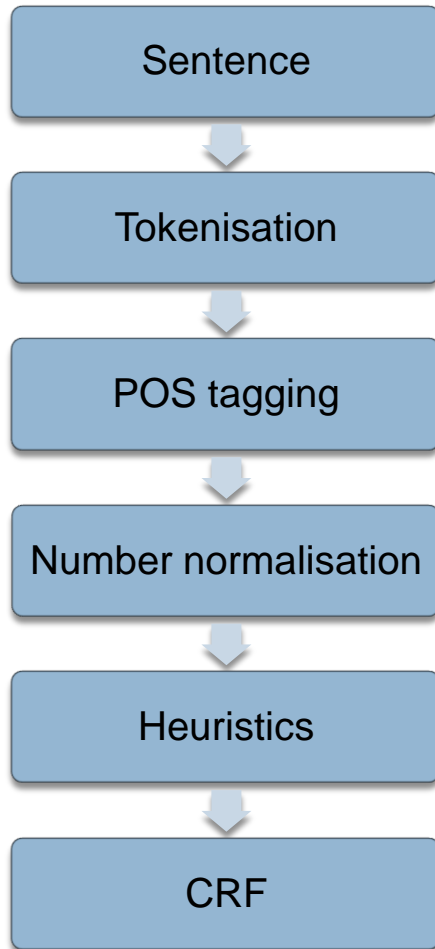
Pipeline – Tokenisation and POS tagging

```
┌─────────────────────────┐
│        Sentence         │
└─────────────────────────┘
             │
             ▼
┌─────────────────────────┐ ┐
│      Tokenisation       │ │
└─────────────────────────┘ │
             │              ├─ MeCab
             ▼              │
┌─────────────────────────┐ │
│      POS tagging        │ │
└─────────────────────────┘ ┘
             │
             ▼
┌─────────────────────────┐
│  Number normalisation   │
└─────────────────────────┘
             │
             ▼
┌─────────────────────────┐
│       Heuristics        │
└─────────────────────────┘
             │
             ▼
┌─────────────────────────┐
│          CRF            │
└─────────────────────────┘
```

Pipeline – Number normalisation

Sentence

Tokenisation

POS tagging

Number normalisation

Heuristics

CRF

Rough normalisation:

- 1.3 → FLOAT
- [1..12] → NUMBER_MONTH
- [13..31] → NUMBER_DAY
- [1940..2099] → NUMBER_YEAR
- \d+ → NUMBER

Pipeline – Heuristics

Sentence

↓

Tokenisation

↓

POS tagging

↓

Number normalisation

↓

Heuristics

↓

CRF

Trigger words:
- Gender: 男性 or 女性
- Time:
    - Numbers followed by a time suffix (月,日, …)
    - Prefixes: 今, 同, 当, 前, 翌
- Hospital:
    - Complete hospital mentions e.g. "this hospital"
    - Suffixes e.g. "hospital" or "clinic"
- Locations: ■ until the next particle

Pipeline – Classifier

Sentence

Tokenisation

POS tagging

Number normalisation

Heuristics

CRF
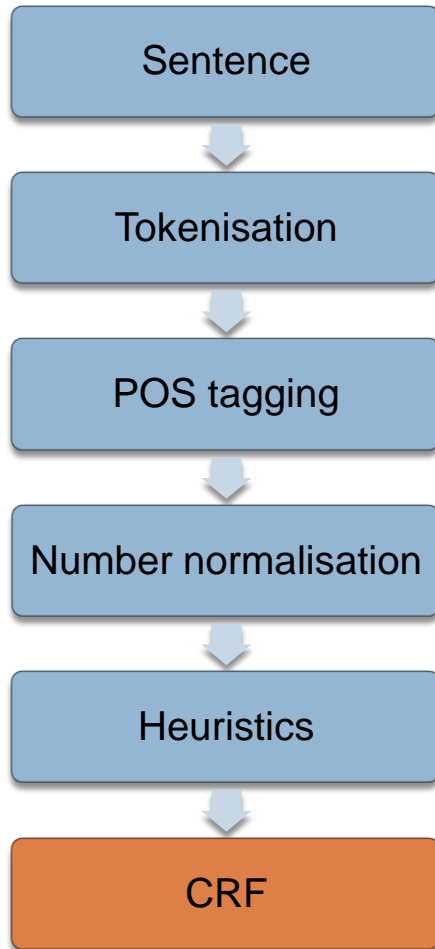
Summary of features

- Normalised tokens
- Part of speech tags
- Heuristics as binary features:
  - Time
  - Gender
  - Hospital
  - Location
- All combined with a CRF++ feature template

Results on the training set

Features	Precision	Recall	F_1
Raw tokens	89.74	74.14	81.09
Normalised tokens	91.00	79.35	84.74
+ POS tags	90.82	80.96	85.55
+ heuristics	90.58	87.12	88.77

Submitted systems

System	P	R	F_1	Ranking (/14)
simple	91.67	86.57	89.05	2
extras	90.82	87.04	88.89	3
extras-b	90.05	87.96	88.99	4

Per-class results

Class	P	R	F$_1$
Age	90.00	84.38	87.10
Hospital	91.89	89.47	90.67
Location	00.00	00.00	00.00
Time	91.24	88.65	89.93
Gender	100.00	100.00	100.00

Summary

- Simple heuristics performed well enough
- Adding more data would have made these heuristics less relevant

Task 2: complaint & diagnosis detection

Suspected complaint

以上の結果より何らかの筋原性疾患を疑い。

Complaint

①急性胆嚢炎

アスペルギルス抗原（－） クリプトコッカス抗原（－）オーム病クラミジア（－） C.ニューモニエ（－）

Negated complaint

結膜貧血・黄染なし、

Hb5.9mg／dlであり、4月1日、2日にMAP4単位輸血。

右下腿縦切開したところ、腓腹筋は脂肪に置換されていた。

4月よりmethotrexate（MTX） 4mg／week、prednisolone（PSL） 2mg／dayにて加療開始し、MTX

Negated complaint

6mg／weekまで増量され、関節症状は改善した。

Complaint Complaint

発症時期不明：高血圧、脂質異常症

Our task

- Detect these types of complaints & diagnoses:
  - Positive　　　　**ＬＤＨ高値**
  - Negated　　　　**頸部リンパ節腫脹**なし
  - Suspected　　　**筋原性疾患**を疑い
  - Family　　　　　息子：**気管支喘息**

Class frequencies

- Complaints and diagnoses:
  - Positive   ← 1314 mentions
  - Negated   ← 504 mentions
  - Suspected   ← 72 mentions
  - Family   ← 32 mentions

Observations

- Specialised terminology: **β細胞疲弊** (beta cell exhaustion)
- Language mix: **pneumatosis intestinalis**がみられた
- Long range effects: <u>**異常陰影**</u>、<u>**気胸**</u>の所見なし

- All four classes look similar – factors that affect the subclass are not part of the mention itself (なし is not part of a negated mention)

General approach

- Use dictionaries to cover special terminology
- Use heuristics to propagate long range negation and suspicion information
- 2 stages
  1) find complaint & diagnosis mentions
  2) determine their right subclass

Summary of features

- Normalised tokens
- Part of speech tags
- Life Science Dictionary membership
- UMLS membership and category (disease, symptom, medication, etc.)
- Binary heuristics

Life Science Dictionary

- Japanese translation of ~32000 biomedical terms

がん検診 [がんけんしん] /cancer screening/
頭痛 [ずつう] /headache/
日焼け [ひやけ] /sunburn (vs)/

- Greedy longest match to tag tokens

UMLS

- NLM effort to unify [bio-]medical vocabularies
- 3 components:
  - Metathesaurus: biomedical concepts linked to various sources
  - Semantic network: categories, relationships
  - Lexical information and NLP tools

UMLS metathesaurus

- > 100 vocabularies (MeSH, ICD-10, SNOMED)
- > 15 languages
- Gives us information about:
  - What concepts exist
  - How they are named
  - Their relationships

UMLS concept organisation
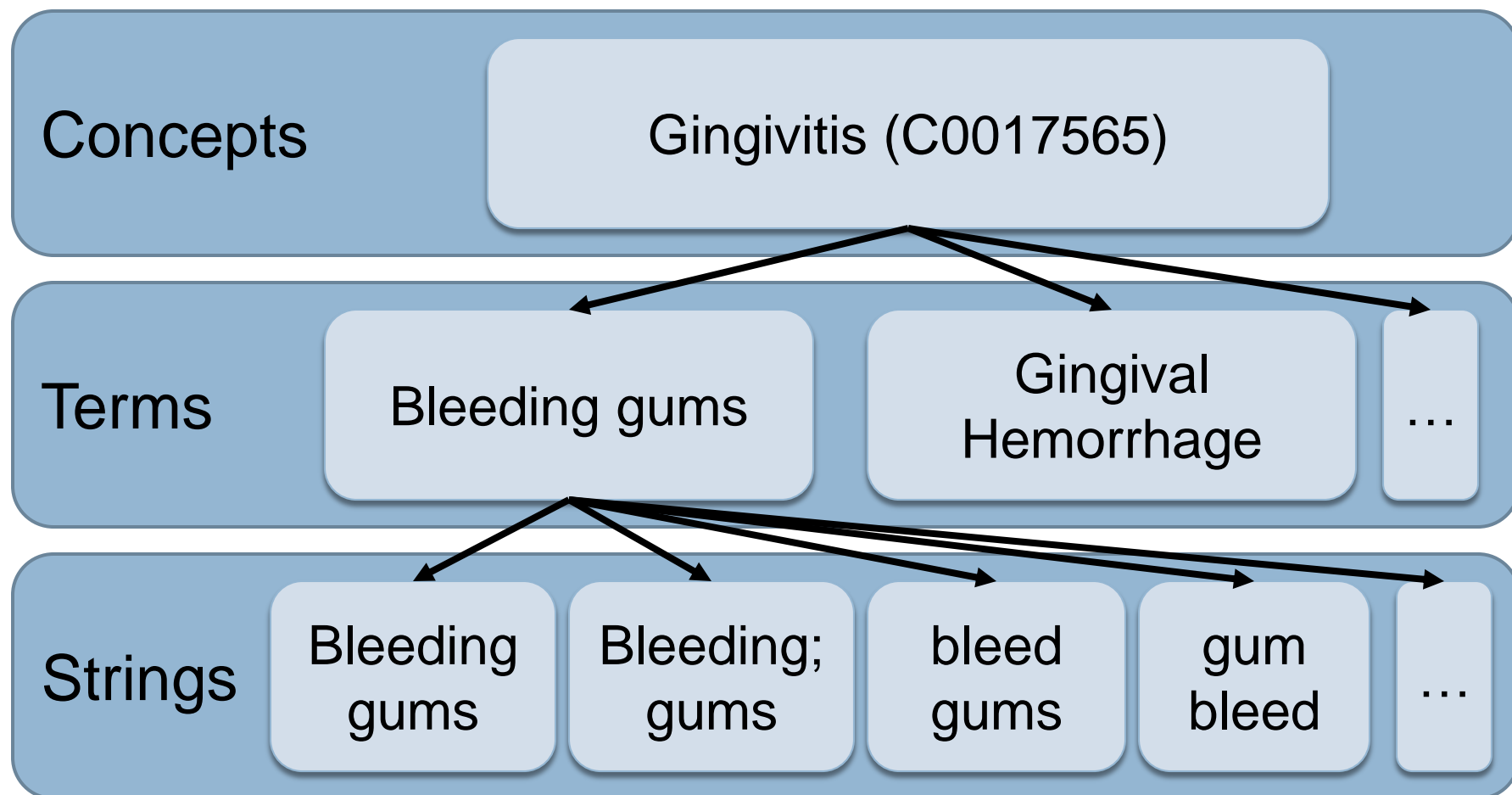
Japanese sources in UMLS

- MeSH (MEdical Subject Headings)
- MedDRA (Medical Dictionary for Regulatory Activities terminology)
- 179881 distinct strings
- 58528 distinct concepts

UMLS semantic network

- Broad categorisation of concepts
- 133 semantic types:

  …

  Physiology - Organ or Tissue Function

  Physiology - Physiologic Function

  Procedures - Diagnostic Procedure

  Procedures - Educational Activity

  …

Medical concept detection

- Process both languages independently:
  - English: MetaMap
  - Japanese: Longest match

UMLS features

- IOB2 membership feature
- Category (disease, medication, …)

Heuristics

- Assume we have a good generic complaint & diagnosis detector
- Provide hints about what sub-type they should be

Family heuristics

- Flag the whole sentence if we encounter one of

息子, 娘, 父, 母, 兄, 姉, 妹, 弟, 祖母, 祖父, 母親, 父親

Negation heuristics

- Negation suffixes: ない, 弱い, 改善, ぬ
- Propagate polarity changes through dependency tree

Suspicion heuristics

- Triggers: 疑う, 考える, 可能
- Propagate to dependents
- Exceptions for justifications:
  - "The patient may have **brain cancer** because of his **cluster headaches**"

Heuristics performance

□ Performance only on complaint tokens

Class	Precision	Recall
Positive	0.92	0.85
Family	0.86	0.94
Negative	0.79	0.78
Suspicion	0.30	0.74

Classification: two approaches

1) Single stage: predict all classes at once

2) Two stages:

    - Detect complaints and diagnoses

    - Subclassify into the correct class using the heuristics

Single class case

Features	P	R	F_1
Normalised tokens	88.81	66.17	75.82
+ POS	87.41	74.20	80.21
with dictionaries	87.15	79.19	82.97

Test set performance

Class	Stages	P	R	F$_1$
No modalities	Single	89.76	77.81	83.36
	Two	89.68	79.98	84.55
All	Single	81.15	70.35	75.36
	Two	75.97	67.75	71.62
Positive	Single	80.92	73.28	76.91
	Two	80.80	68.00	73.85
Family	Single	82.35	63.64	71.79
	Two	65.22	68.18	66.67
Negation	Single	84.50	68.42	75.62
	Two	75.23	67.61	71.22
Suspicion	Single	50.00	30.00	37.50
	Two	35.85	63.33	45.78

Conclusion

Summary

- Heuristics helped on both tasks
- If we add more training data:
  - Task 1 heuristics will become less relevant
  - Task 2 will still be hard for local classifiers
  - → Investigate better ways to propagate long range information (RNN, Tree CRF)
- 2 stage architecture not necessary
- Need better terminology resources

Thank you for your attention

- Questions?

Resources

- CRF++:   https://code.google.com/p/crfpp/

- MeCab: http://code.google.com/p/mecab/

- J.DepP: http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/jdepp/

- UMLS: http://www.nlm.nih.gov/research/umls/

- Life Science Dictionary: http://lsd.pharm.kyoto-u.ac.jp/

- MetaMap: http://metamap.nlm.nih.gov/