UT-FX at NTCIR-10: Incorporating Medical Knowledge to Enhance Medical Information Extraction

Yasuhide Miura[1], Tomoko Ohkuma[1], Hiroshi Masuichi[1],
Emiko Yamada-Shinohara[2], Eiji Aramaki[3,4], and Kazuhiko Ohe[2,3]

2013/6/20

[1]Fuji Xerox Co., Ltd., Japan
[2]The University of Tokyo Hospital, Japan
[3]The University of Tokyo, Japan
[4]JST PRESTO, Japan

Introduction (1)

- Team Name
  - UT-FX
- Participated Subtasks
  - Task 1: De-identification
  - Task 2: Complaint and Diagnosis

Introduction (2)

- We submitted two systems for Task 1 and three systems for Task 2.
  - Task 1: De-identification
    - Scores (in overall PRF)
      - Our best system (A1): P=91.50%, R=84.72%, F_1=87.98
        - » 6th system (3rd in team)
  - Task 2: Complaint and Diagnosis
    - Scores (ignoring modalities)
      - Our best system (A1): P=90.73%, R=81.6%, F_1=85.93
        - » 1st system

Overview of this presentation

- Introduction
- System Description
  - 2-stage Named Entity Recognition
  - Integration of Knowledge Resources
  - System Overview
  - Features
- Experiments
- Test Runs
- Conclusion

System Description (1)

- Basic Architecture
  - A named entity recognizer
    - Statistical system
      - Conditional random fields (CRF) based
    - Character (NOT word) level process
- Characteristics
  - 2-stage named entity recognition

    For modality detection
  - Integration of knowledge resources
    - Terminologies
    - Named entities (on a different scheme)

System Description (2)

- Named entity recognition (NER) is well studied. Many techniques to improve NER is already known.
  - Character level process
    - Asahara and Matsumoto[3] showed a state of the art Japanese NER can be realized by character level process.
  - Integration of knowledge resources
    - Kazama and Torisawa[5] exploited Wikipedia to enhance NER.
  - 2-stage NER
    - Alex et al.[1] investigated several layering techniques of NER.

2-stage Named Entity Recognition

- Used to detect modalities
  1. Complaint and Diagnosis tags (*c* tags) are converted to nested representations.
  2. A 2-stage NER is performed to extract *c* tags and their modalities.

EN No <c modality="negation">edema</c> on the front shin bone part.

JA 前脛骨部に<c modality="negation">浮腫</c>なし。

nested representation

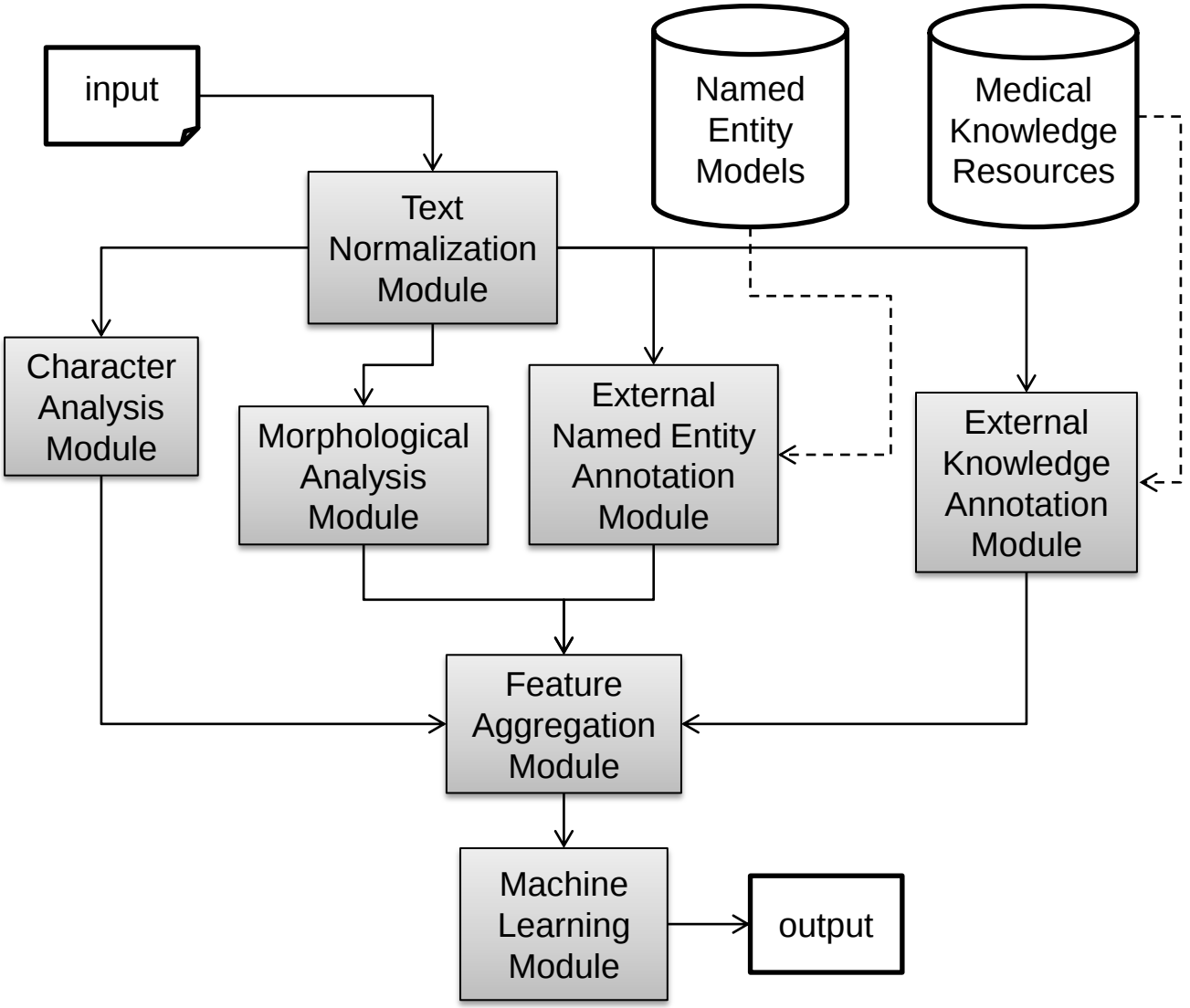
EN No <c><c-neg>edema</c-neg></c> on the front shin bone part.

JA 前脛骨部に<c><c-neg>浮腫</c-neg></c>なし。

Integration of Knowledge Resources

- We integrated two kinds of medical knowledge resources to enhance our system.
  - Terminologies
    - MedDRA
      - A pragmatic, medically valid multilingual terminology.
    - MEDIS Byomei Master
      - A disease name terminology in Japanese.
    - MEDIS Shojyo Shoken Master <Shintai Shoken Hen>
      - A diagnosis terminology in Japanese.
  - Named Entities of Discharge Summary Corpus (DS Corpus)
    - *date* (including time) *and symptom* (including disease names)
    - DS Corpus is a corpus of Japanese discharge summary texts (detail in Aramaki et al.[2])

System Overview



Features (1)

- Various character or morphology features are defined for each characters.

Feature	Brief Description
C-SURF	The surface form of a character.
C-TYPE	The type of a character.
M-SURF	The surface form of a morpheme.
M-BASE	The base form of a morpheme.
M-POS1, M-POS2, M-POS3	The POS layer 1, 2, and 3 of a morpheme, respectively.
M-CJ-FORM	The conjugation form of a morpheme.
M-CJ-TYPE	The conjugation type of a morpheme.

Features (2)

- Knowledge resources are integrated as features.

Feature	Brief Description
MEDDRA	The BIO-style matching result of a character with MedDRA/J entries
MEDIS-BM	The BIO-style matching result of a character with MEDIS Byomei Master entries.
MEDIS-SSM	The BIO-style matching result of a character with MEDIS Shojo Shoken Master <Shintai Shoken Hen> entries.
NE-DT	The matching with recognized DS Corpus *date* named entities.
NE-SD	The matching with recognized DS Corpus *symptom* named entities.
NE-C	The matching with *c* tags of the MedNLP task.

Experiments (1)

- Task 1: De-identification Task
  - Evaluation Method
    - 5-fold cross validation on the sample data*.
  - Evaluation Target
    - *a* (age), *h* (hospital), and *t* (time) tags which appeared more than 50 times in the sample data.
  - Features
    - 2 system compositions

Composition	Features
BASELINE	{C-SURF, C-TYPE, M-SURF, M-BASE, M-POS1, M-POS2, M-POS3, M-CJ-FORM, M-CJ-TYPE}
DATETIME	BASELINE ∪ {NE-DT}

* 2,244 sentences of the NTCIR-10 MedNLP sample data.

Experiments (2)

- Task 1: De-identification Task
  - Results

Composition	Tag	Precision	Recall	F_1 Score
BASELINE	a	86.67%	69.94%	77.23
	h	98.51%	88.00%	92.96
	t	90.42%	85.07%	87.66
	overall	91.26%	83.74%	87.34
DATETIME	A	79.55%	62.50%	70.00
	H	98.53%	89.33%	93.71
	T	91.52%	85.07%	88.18
	overall	91.40%	83.13%	87.07

Experiments (3)

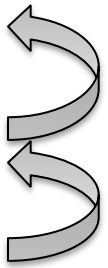
- Task 2: Complaint and Diagnosis Task
  - Evaluation Method
    - 5-fold cross validation on the sample data (same as Task 1)
  - Evaluation Target
    - *c* (complaint and diagnosis) tag
  - Features
    - 4 system compositions

Composition	Features
BASELINE	{C-SURF, C-TYPE, M-SURF, M-BASE, M-POS1, M-POS2, M-POS3, M-CJ-FORM, M-CJ-TYPE}
SYMPDIS	BASELINE ∪ {NE-SD}
MEDDIC	BASELINE ∪ {MEDDRA, MEDIS-BM, MEDIS-SSM}
FULL	BASELINE ∪ SYMPDIS ∪ MEDDIC

Experiments (4)

- Task 2: Complaint and Diagnosis Task
  - Results (*c* tags ignoring modalities)
    - Underlined values denote statistically significant improvements.

Composition	Precision	Recall	F_1 Score
BASELINE	87.87%	81.43%	84.53
SYMPDIS	87.46%	84.18%	85.79
MEDDIC	88.57%	83.45%	85.94
FULL	88.39%	84.76%	86.54



statistical significance tests

SYMPDIS and MEDDIC against BASELINE

FULL against MEDDIC

Experiments (5)

- Task 2: Complaint and Diagnosis Task
  - Results (*c* tag modalities)
    - Gold *c* tags used.
    - Statistical improvements not observed.

Composition	Modality	Precision	Recall	F_1 Score
BASELINE	none	87.46%	94.98%	91.06
	negation	87.50%	76.39%	81.57
	suspicion	61.11%	30.56%	40.74
	family	78.57%	34.38%	47.83
FULL	none	87.81%	93.76%	90.69
	negation	84.65%	76.59%	80.42
	suspicion	55.81%	33.33%	41.74
	family	75.00%	37.50%	50.00

Test Runs

- We submitted two systems for De-identification Task and three systems for Complaint and Diagnosis Task.

Task	Composition	ID*	Precision**	Recall**	F_1 Score**
De-identification	BASELINE	A1	91.50%	84.72%	87.98
	DATETIME	A2	90.15%	84.72%	87.35
Complaint and Diagnosis	BASELINE	A3	88.26%	79.76%	83.80
	MEDDIC	A1	90.73%	81.60%	85.93
	FULL	A2	88.34%	82.03%	85.07

*ID in the overview paper.

** The scores correspond to overall results in De-identification Task and 2-way results in Complaint and Diagnosis Task.

Conclusion

- We presented a system that utilizes external medical knowledge into a state-of-the-art named entity recognizer.
  - About 2.03 improvement in F_1 score (2-way) in Complaint and Diagnosis Task.
    - However, the best feature compositions differs between the experiment and the test run.
- The result suggests the promising future of a natural language processing in medical fields.
  - Numerous knowledge resources are available in the medical fields.

References

[1] B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. In Proceedings of the Workshop on BioNLP 2007, pages 65—72, 2007.

[2] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Mashuichi, and K. Ohe. TEXT2TABLE: Medical text summarization system based on named entity recognition and modality identication. In Proceedings of the BioNLP 2009 Workshop, pages 185—192, 2009.

[3] M. Asahara and Y. Matsumoto. Japanese named entity extraction with redundant morphological analysis. In Proceedings of HLT-NAACL 2003, pages 8—15, 2003.

[4] N. Chinchor. The statistical significance of the MUC-4 results. In Proceedings of MUC-4, pages 30—50, 1992.

[5] J. Kazama and K. Torisawa. Exploiting Wikipedia as external knowledge for named entity recognition. In Proceedings of EMNLP-CoNLL 2007, pages 698—707, 2007.

[6] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, pages 282—289, 2001.

[7] M. Morita, Y. Kano, T. Ohkuma, M. Miyabe, and E. Aramaki. Overview of the NTCIR-10 MedNLP task. In Proceedings of NTCIR-10, 2013. To appear.

Thank you for listening!