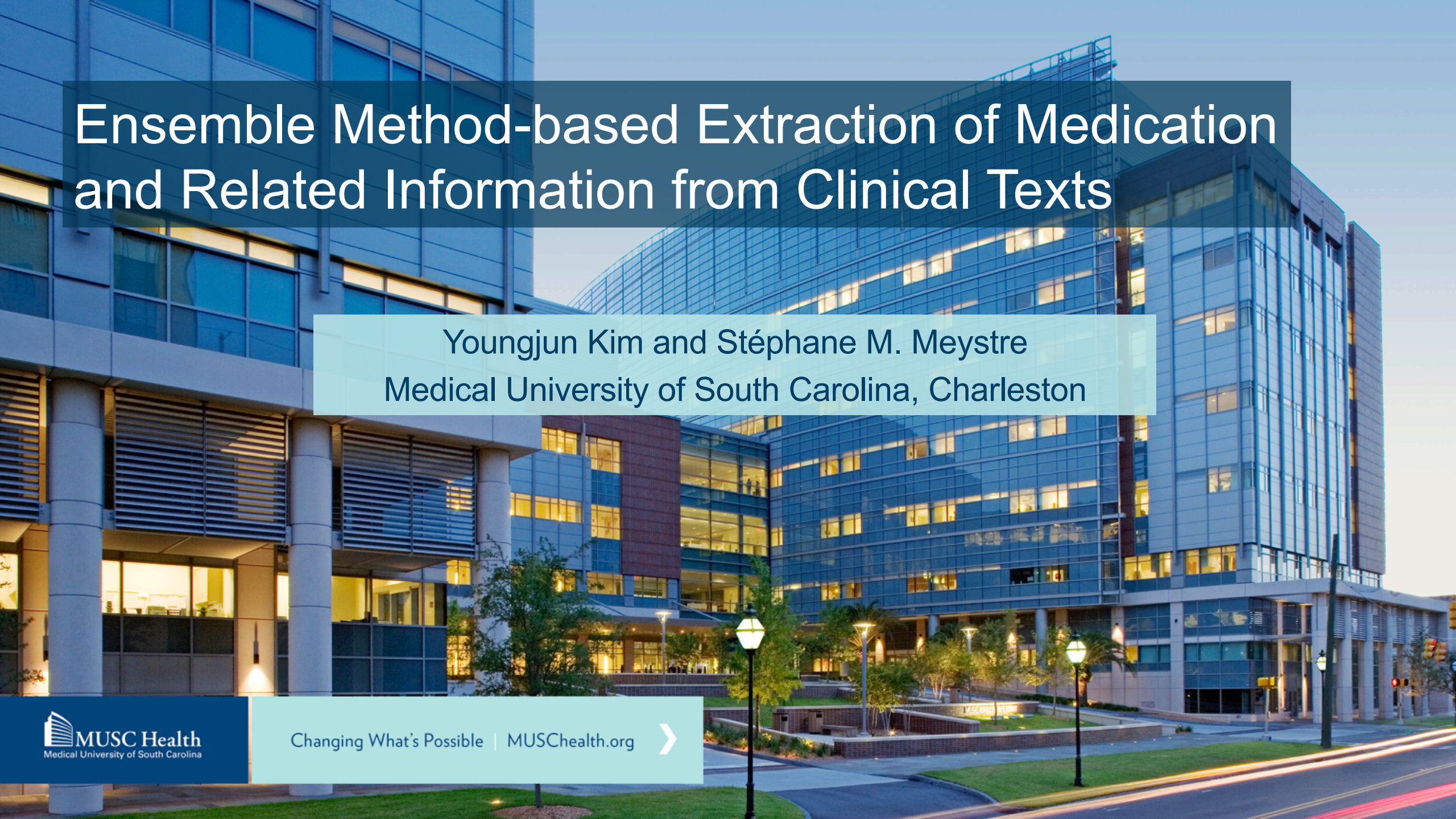


Ensemble Method-based Extraction of Medication and Related Information from Clinical Texts

Youngjun Kim and Stéphane M. Meystre
Medical University of South Carolina, Charleston

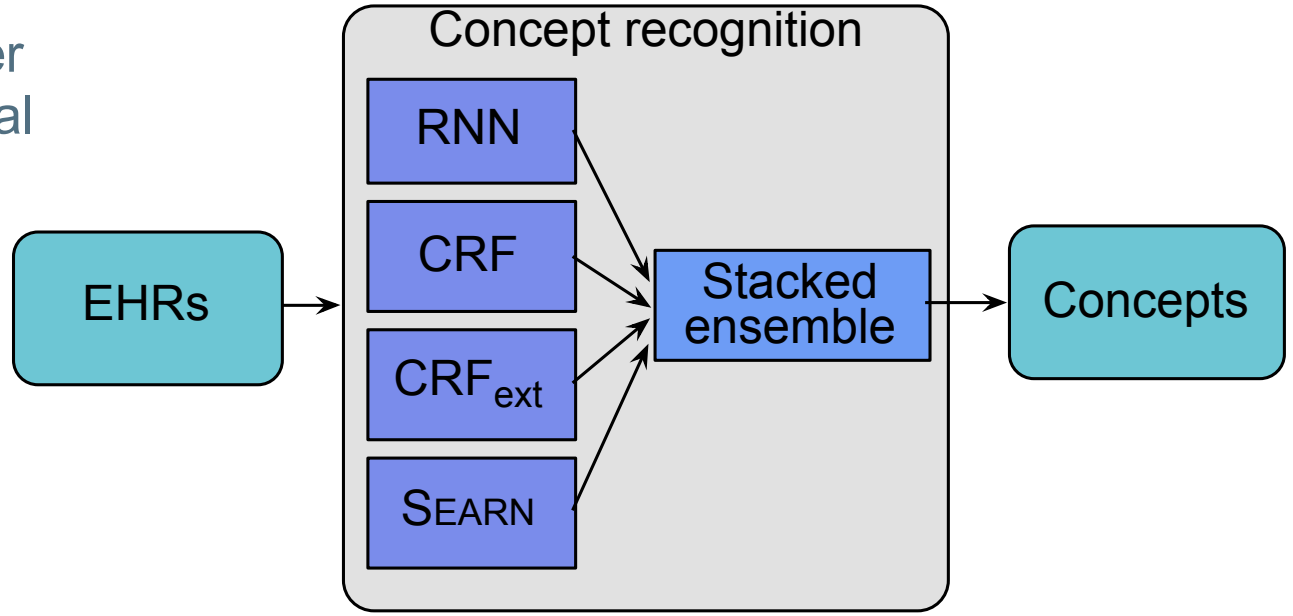
2018 n2c2 Shared-Task

Track 2: Adverse Drug Events and Medication Extraction in EHRs

- Subtask 1: Concept identification to identify drug names, dosages, durations and other entities
- Subtask 2: Relation classification to identify relations of drugs with adverse drugs events and other entities given gold standard entities
- Subtask 3: End-to-end to identifying relations of drugs with ADEs and other entities on system predicted entities

Subtask 1

- Identification of drugs and related entities
  - Multiple stage of ensemble learning
  - Four individual sequential taggers
  - A sequential model-based meta-classifier trained on the predictions of the individual classifiers



External resources

- Pre-processing
  - Stanford CoreNLP tool
  - Tokenization, lemmatization, part-of-speech tagging, and NER System
- Standalone System
  - MedEx
- Corpora
  - The 2009 i2b2 challenge corpus
  - The 2010 i2b2 challenge corpus
  - Cadec corpus (from posts on social media)

- MedEx: Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. JAMIA. 2010;17(1):19-24
- Cadec: Karimi S, Metke-Jimenez A, Kemp M, Wang C. Cadec: A corpus of adverse drug event annotations. Journal of biomedical informatics. 2015 Jun 1;55:73-81.

External resources

MedEx: Drug, Strength, Duration, Route, Frequency, Form, Dosage, ~~ADE, Reason~~

i2b2 2009: Drug, Strength, Duration, Route, Frequency, ~~Form, Dosage, ADE, Reason~~

i2b2 2010: Drug, ~~Strength, Duration, Route, Frequency, Form, Dosage, ADE, Reason~~

Cadec: Drug, ~~Strength, Duration, Route, Frequency, Form, Dosage, ADE, Reason~~

	Strict			Lenient		
Overall (micro avg.)	Precision	Recall	F_1 score	Precision	Recall	F_1 score
MedEx	72.19	53.50	61.46	85.45	61.95	71.83
i2b2 2009	72.56	44.86	55.44	90.47	55.81	69.03
Drug only						
i2b2 2010	52.01	69.92	59.65	66.03	88.3	75.56
Cadec	65.18	35.08	45.61	78.51	41.78	54.54

Information extraction models

- RNN (recurrent neural network)
- CRF (conditional random fields)
- CRF with the outputs of external resources
- SEARN (search-based structured prediction)[1] with the outputs of external resources

1. Daumé H, Langford J, Marcu D. Search-based structured prediction. Mach Learn. 2009;75(3):297–325.

RNN

- Train a LSTM (long short-term memory network) model
- Employing the NER system of Lample et al[1]
- Sequences of word tokens were used

1. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. Proceedings of the 15th NAACL; 2016. p. 260–70.

CRFs

- Train two CRF models
  - Wapiti[1]: a simple and fast discriminative sequence labeling toolkit
  - Features
    - Lexical string
    - Lemma
    - POS tag
    - Affix(es),
    - Orthographic features
    - Named entity tag
    - Cluster identifier of word embeddings clusters
- Another version of a CRF classifier using the predictions of external resources

1. Lavergne T, Capp´e O, Yvon F. Practical very large scale CRFs. Proceedings of the 48th ACL; 2010. p. 504–13.

SEARN

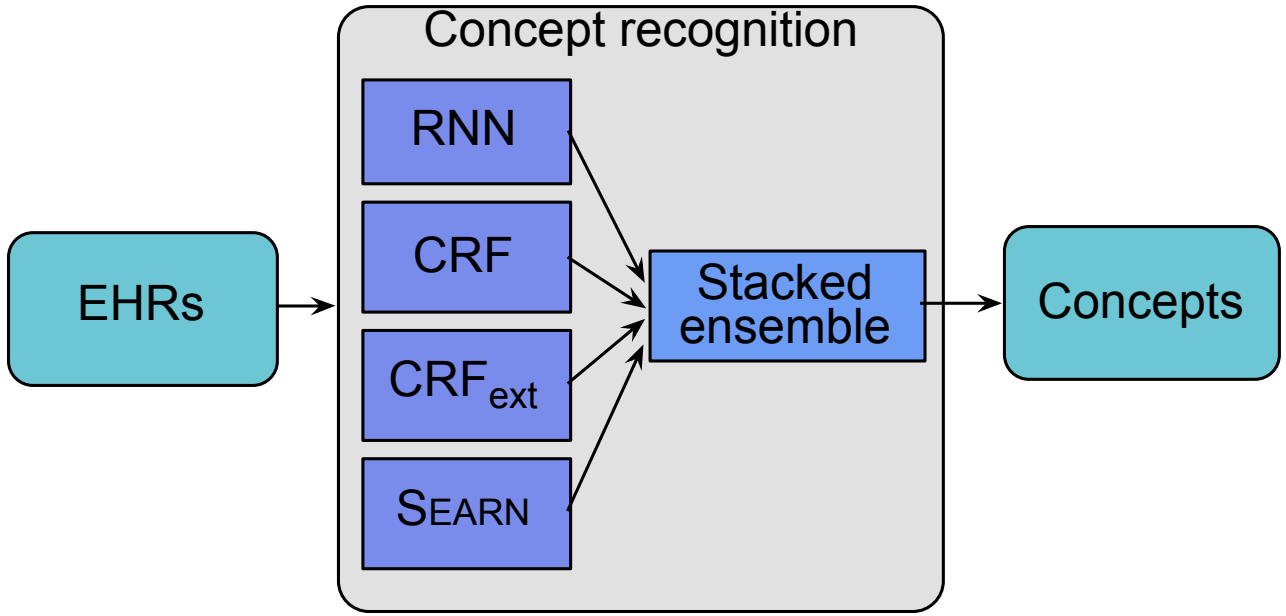
- Vowpal Wabbit[1]
  - A fast out-of-core online learning system
  - Supports a number of machine learning algorithms
- Features: trigram with skips
  - Word, Lemma
  - Cluster identifier of each word
    - Used Word2Vec[1] to perform K-means clustering on word embeddings
    - First billion characters from wikipedia[2]
  - Predictions of external resources

1: Langford J. Vowpal Wabbit, http://hunch.net/~vw/, 2011.

2: http://mattmahoney.net/dc/enwik9.zip

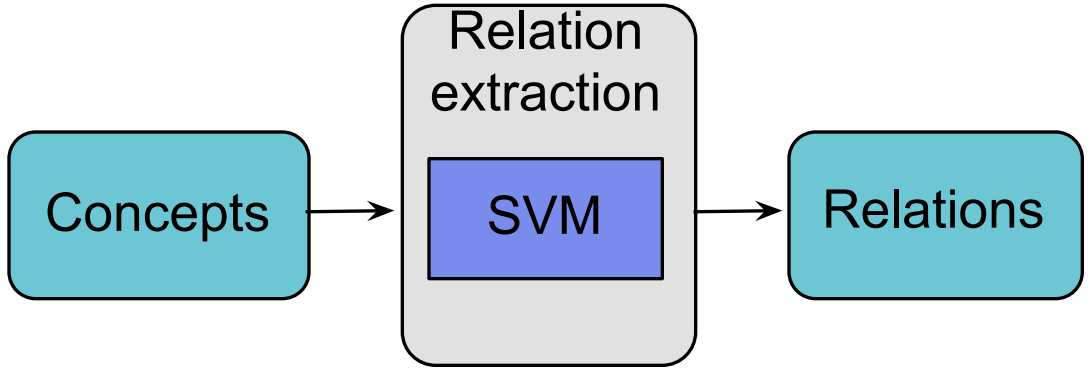
Stacked ensemble

- Created a meta-classifier
- Training a SEARN classifier based on the predictions from the individual classifiers
- Performed a 10-fold cross-validation on the training set to obtain predictions for each classifier



Subtask 2

- Relation classification
  - Between medication attributes and adverse drugs events
  - Support Vector Machine (SVM) model



Relation classification

- Trained an SVM classifier with a linear kernel
- Used the LIBLINEAR software[1]
- Created a binary-class SVM classifier

- Filtering
  - Do not consider when two concepts are away from each other more than
    - 200 words
    - 10 sentences
    - 1000 characters (including whitespaces)

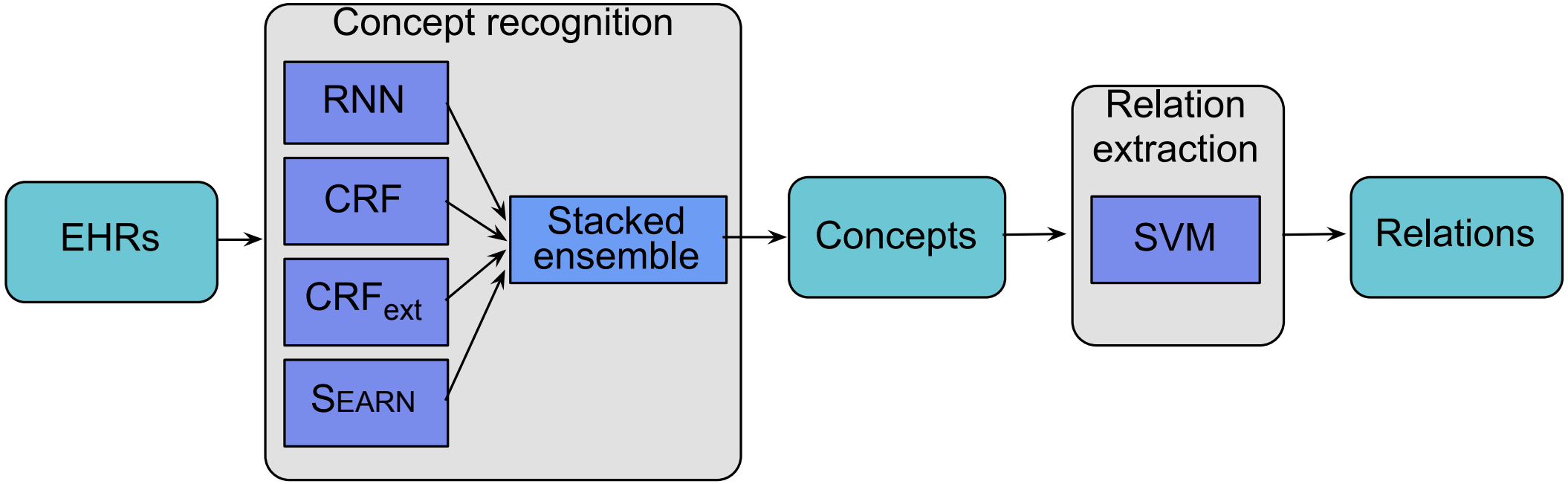
1: Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. J Mach Learn Res. 2008;9(Aug):1871–4.

Relation classification

- Features
  - distance between two concepts:
    - Counting the number of characters, words, and concepts between them
    - A feature that indicates whether the concept pair is closest in the file
  - Lexical features
    - Three preceding and three following words for each concept
    - Words between the two concepts
  - Word embedding
    - The cluster identifier of each word contained in two concepts
    - The cluster identifier of each word between the two concepts

System architecture

End-to-end concept and relation identification system



Results

Subtask1: Individual models results

	Strict			Lenient		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
CRF	90.91	84.28	87.47	95.54	88.34	91.80
CRFext	91.13	84.49	87.69	95.71	88.52	91.97
SEARN	91.38	84.48	87.80	96.07	88.53	92.15
RNN	89.33	85.14	87.19	94.65	**89.78**	92.15
Stacked ensemble	**91.93**	**85.46**	**88.58**	**96.30**	89.29	**92.66**

Testing for Statistical Significance:

- paired t-tests to compare individual classifiers and stacked ensemble
- Ran 2^{20} trials to calculate the significance between two methods[1]

The performance of the stacked ensemble was significantly better than all of the other methods (p = 0.05) except RNN recall

1. Yeh A. More accurate tests for the statistical significance of result differences. Proceedings of the 18th COLING; 2000. p. 947–53

Stacked learning results

	Strict			Lenient		
	Precision	Recall	F_1 score	Precision	Recall	F_1 score
Drug	93.15	92.50	92.83	96.13	94.98	95.55
Strength	95.64	95.41	95.53	98.08	97.68	97.88
Duration	80.81	63.49	71.11	92.93	73.02	81.78
Route	95.01	92.71	93.85	96.64	94.19	95.40
Form	94.68	90.71	92.65	97.60	93.42	95.46
ADE	61.15	15.36	24.55	67.52	16.96	27.11
Dosage	91.31	90.94	91.12	94.12	93.73	93.93
Reason	84.42	39.61	53.92	90.56	42.20	57.57
Frequency	83.92	82.75	83.33	97.77	96.36	97.06
Overall (micro avg.)	91.93	85.46	88.58	96.30	89.29	92.66
Overall (macro avg.)	91.72	83.89	87.42	96.09	87.69	91.48

Relation classification results

With given concepts

	Precision	Recall	F_1 score
Strength	96.31	97.90	97.10
Dosage	95.89	97.77	96.82
Duration	87.91	88.73	88.32
Frequency	96.75	96.68	96.71
Form	98.58	98.58	98.58
Route	96.29	97.97	97.12
Reason	76.64	77.57	77.10
ADE	75.75	79.67	77.66
Overall (micro avg.)	93.07	94.12	93.59
Overall (macro avg.)	92.30	93.55	92.83

Relation classification results

With detected concepts

	Precision	Recall	F_1 score
Strength	94.55	95.71	95.13
Dosage	92.12	91.58	91.85
Duration	83.14	65.96	73.56
Frequency	95.09	93.58	94.33
Form	96.96	92.52	94.69
Route	93.87	92.44	93.15
Reason	74.97	34.87	47.60
ADE	48.46	15.01	22.92
Overall (micro avg.)	92.52	81.88	86.88
Overall (macro avg.)	91.04	78.15	83.35

Discussion and future work

- SEARN and RNN performed well
  - SEARN produced high precision, RNN produced high recall
- External resources did not contribute much to the performance
  - The amount of training data might be already sufficient
  - Content differences between different corpora

- Future work
  - Apply n2c2 models to the other corpora
    - e.g. the 2009 i2b2 data
    - Employ domain adaptation models

Thank you.

Questions?

Acknowledgments: supported by the SmartState Program
(Translational Biomedical Informatics Chair Endowment)

Contact information: kimy@musc.edu