Sieve-Based Entity Linking for the Biomedical Domain

Jennifer D'Souza and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas

Entity Linking

- Given an entity mention in a text document and a knowledge base (KB) of entities,
  - find the entity in the KB the entity mention refers to

    or

  - determine that such entity does not exist in the KB

Entity Linking

- challenging because
  - mentions with the same word/phrase can refer to different entities
  - mentions with different words/phrases can refer to the same entity
- known as normalization for the biomedical domain
  - Map a word/phrase in a document to a concept in an ontology after disambiguating potential ambiguous words/phrases
- **Our goal**: normalize disorder mentions

Plan for the Talk

- Datasets
- Multi-pass sieve approach to normalization
- Evaluation

Plan for the Talk

- Datasets
- Multi-pass sieve approach to normalization
- Evaluation

Datasets

- Two standard evaluation datasets from two genres
- The ShARe eHealth Challenge corpus (Pradhan et al., 2013)
  - 298 de-identified **clinical reports** from US Intensive Care
- The NCBI disease corpus (Dogan et al., 2014)
  - 793 **biomedical abstracts**

Datasets: Statistics

	ShARe (Clinical reports)	NCBI (Biomedical abstracts)
Documents	298	792
Disorder mentions	11167	6885
Mentions with ID	7793	6885
ID-less mentions	3374	0

Datasets: Statistics

	ShARe (Clinical reports)	NCBI (Biomedical abstracts)
Documents	298	792
Disorder mentions	11167	6885
Mentions with ID	7793	6885
ID-less mentions	3374	0

Datasets: Statistics

	ShARe (Clinical reports)	NCBI (Biomedical abstracts)
Documents	298	792
Disorder mentions	11167	6885
Mentions with ID	7793	6885
ID-less mentions	3374	0

Datasets: Statistics

	ShARe (Clinical reports)	NCBI (Biomedical abstracts)
Documents	298	792
Disorder mentions	11167	6885
Mentions with ID	7793	6885
ID-less mentions	3374	0

Datasets: Statistics

	ShARe (Clinical reports)	NCBI (Biomedical abstracts)
Documents	298	792
Disorder mentions	11167	6885
Mentions with ID	7793	6885
ID-less mentions	3374	0

Datasets: Statistics

	ShARe (Clinical reports)	NCBI (Biomedical abstracts)
Documents	298	792
Disorder mentions	11167	6885
Mentions with ID	7793	6885
ID-less mentions	3374	0

- Ontologies
  - ShARe: UMLS Metathesaurus (128,430 disorder concepts)
  - NCBI: MEDIC (11,915 disorder concepts)

Ontology Concepts

- Each concept in these two ontologies is described by:
  - the concept ID
  - the list of terms commonly used to refer to the concept
  - its definition
  - …

Ontology Concepts

- Each concept in the two ontologies is described by:
  - the concept ID
  - the list of terms commonly used to refer to the concept
  - its definition
  - …

Our multi-pass sieve approach only uses this information

Example Ontology Concept

- preprocessed the ontologies so that for each concept we retain only the concept ID and the associated terms

- UMLS Metathesaurus

  C0000731 | swollen abdomen | abdominal distension | abdomen distended | abdominal distention | abdominal swelling

- NCBI

  D008288 | Malaria | Fever, Marsh | Fever, Remittent | Infection, Plasmodium | MALS | Plasmodium Infection | Remittent Fever

Example Ontology Concept

- preprocessed the ontologies so that for each concept we retain only the concept ID and the associated terms

- UMLS Metathesaurus

  C0000731 | swollen abdomen | abdominal distension | abdomen distended | abdominal distention | abdominal swelling

- NCBI

  D008288 | Malaria | Fever, Marsh | Fever, Remittent | Infection, Plasmodium | MALS | Plasmodium Infection | Remittent Fever

Example Ontology Concept

- preprocessed the ontologies so that for each concept we retain only the concept ID and the associated terms

- UMLS Metathesaurus

  C0000731 | swollen abdomen | abdominal distension | abdomen distended | abdominal distention | abdominal swelling

- NCBI

  D008288 | Malaria | Fever, Marsh | Fever, Remittent | Infection, Plasmodium | MALS | Plasmodium Infection | Remittent Fever

Plan for the Talk

- Datasets
- Multi-pass sieve approach to normalization
- Evaluation

Overview of the Sieve Approach

- A sieve is composed of one or more heuristic rules
  - In the context of normalization, each rule normalizes (i.e., assigns a concept ID) to a disorder mention in a document

- Sieves are ordered as a pipeline, in decreasing order of precision

Sieve 1 → Sieve 2 → Sieve 3 → Sieve 4 → Sieve 5

- Later sieves can exploit decisions made by earlier sieves
  - Cannot undo earlier mistakes: errors can propagate

Applying Sieves for Normalization

- The normalizer makes multiple passes over the mentions in a document
  - In the i-th pass, it uses only the rules in the i-th sieve for normalization

Applying Sieves for Normalization

- The normalizer makes multiple passes over the mentions in a document
  - In the i-th pass, it uses only the rules in the i-th sieve for normalization
  - If the i-th sieve cannot normalize a mention unambiguously (i.e., the sieve normalizes it to more than one concept in the ontology), the sieve will leave it unnormalized

Applying Sieves for Normalization

- The normalizer makes multiple passes over the mentions in a document
  - In the i-th pass, it uses only the rules in the i-th sieve for normalization
  - If the i-th sieve cannot normalize a mention unambiguously (i.e., the sieve normalizes it to more than one concept in the ontology), the sieve will leave it unnormalized
  - If a mention is normalized, it will be added to the list of terms associated with the ontology concept to which it's normalized
    - so later sieves can exploit the decisions made by earlier sieves
    - but earlier normalization decisions cannot be overridden later

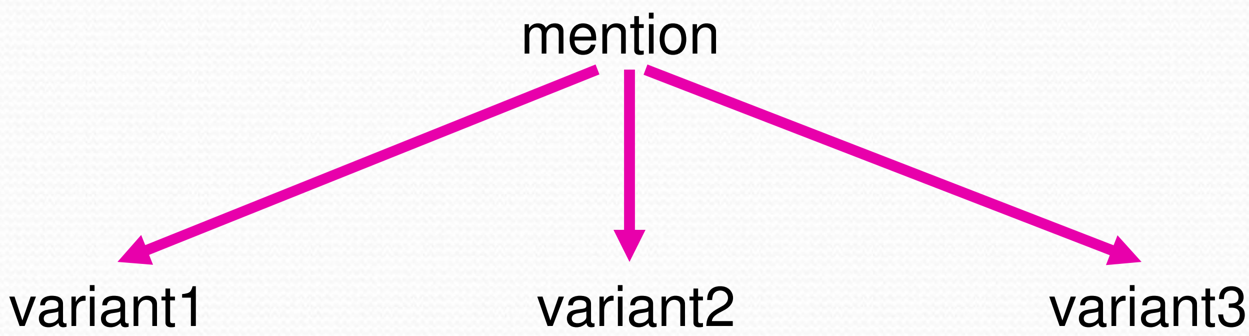
Ten Sieves for Normalization

- General idea

mention

- Sieve 1: mention has exact match with any concept terms?
  - If yes, link mention to the concept associated with the term

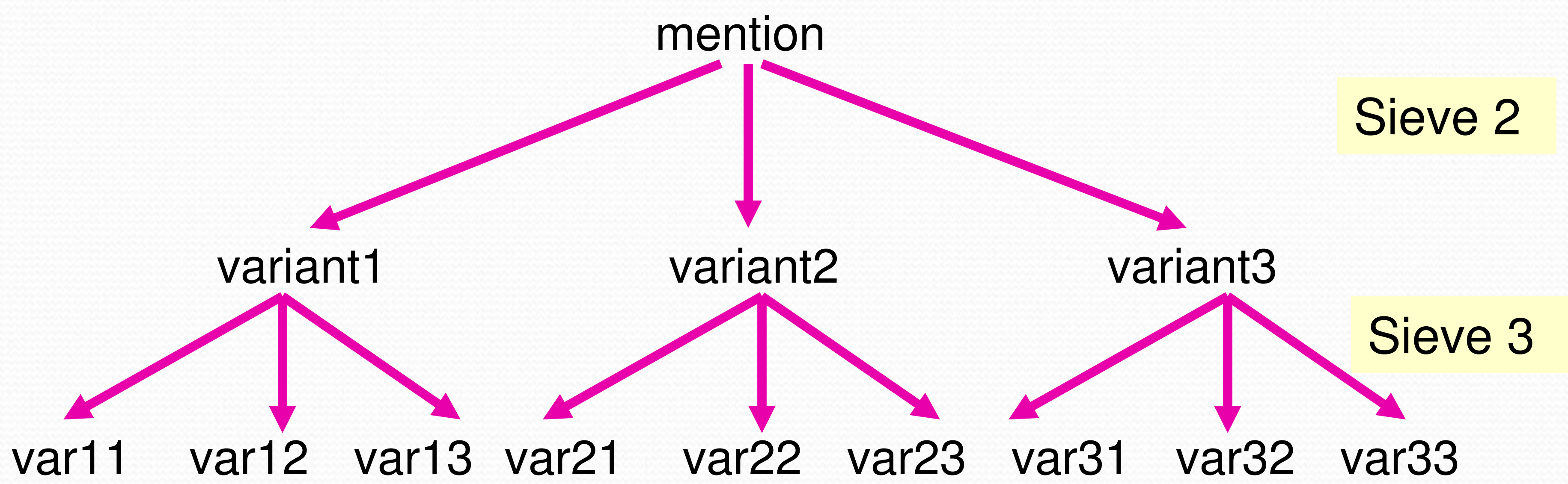
If no, the next sieve creates variants



Sieve 2

- Does any of these variants have an exact match with any concept terms?
  - If yes, link mention to the concept associated with the term

If no, the next sieve creates variants



- Does any of these new variants have an exact match with any concept terms?
  - If yes, link mention to the concept associated with the term

If no, the process repeats

- The next sieve generates more lexico-syntactic variants for each variant generated by the previous sieve

Sieve 1: Exact Match

- Performs exact match of the given disorder mention with the concept terms

Sieve 2: Abbreviation Expansion

- Variants are generated by expanding abbreviated disorder mentions

Sieve 3: Word Reordering

- Variants of a disorder mention are generated by
  - replacing any preposition(s) with other prepositions
    - e.g., “changes on ekg” → “changes in ekg”
  - dropping a preposition and swapping substrings surrounding it
    - e.g., “changes on ekg” → “ekg changes”

Sieve 4: Numbers Replacement

- Variants are generated by replacing each number in the mention with other forms of the same number
  - e.g., “three vessel disease”

  → “3 vessel disease”, “iii vessel disease”, “triple vessel disease”

Sieve 5: Hyphenation

- Variants are generated by **hyphenation** or **dehypenation**

- Hyphenation
  - consecutive words are hyphenated one pair at a time
    - e.g., “ventilator associated pneumonia”

    → “ventilator-associated pneumonia”,
    “ventilator associated-pneumonia”

- Dehypenation
  - hyphens are removed one at a time
    - e.g., “saethre-chotzen syndrome” → “saethre chotzen syndrome”

Sieve 6: Suffixation

- Variants are generated by applying suffixation patterns manually derived from the training data
  - e.g., “infectious source” → “source of infectious” (Sieve 3)
    → “source of infection”

Sieve 7: Disorder Synonym Replacement

- Variants are generated by
  - replacing the disorder term with its synonyms
    - e.g., “presyncopal events”

    → “presyncopal disorders”, “presyncopal episodes”, …
    - synonyms are manually compiled based on the training data

Sieve 8: Stemming

- Variants are generated by stemming the mention using the Porter stemmer

Sieve 9: Composite Mentions and Terms

- A disorder mention or concept term is composite if it contains more than one concept term

- To increase the likelihood of an exact match, we split each composite mention/concept term into its constituent mentions/concept terms before matching
  - E.g., “common eye and/or eyelid symptom”

  → “common eye symptom”, “common eyelid symptom”

Sieve 10: Partial Match

- Rules are different for the two datasets
  - in part because NCBI has no ID-less disorder mentions

- For NCBI, a mention is normalized to the concept containing a term it shares most tokens with

- For ShARe, a mention *m* is normalized to a concept *c* if
  - all tokens in *m* appear in one of the terms in *c* or vice versa
  - *m* has more than 3 tokens and has an exact match with a term in *c* after dropping its 1st token or 2nd to last token; or
  - *c* has a term with three tokens and *m* has an exact match with this term after dropping its 1st or middle token; or

Plan for the Talk

- Datasets
- Multi-pass sieve approach to normalization
- Evaluation

Experimental Setup

- Datasets
  - ShARe (Pradhan et al., 2013)
    - 199 clinical reports for training, 99 reports for testing
  - NCBI (Dogan et al., 2014)
    - 693 biomedical abstracts for training, 100 abstracts for testing

- Evaluation measure: **Accuracy**
  - Percentage of gold mentions correctly normalized

Baseline Systems: Supervised Approach

- DNorm (Leaman et al., 2013)
  - best result to date on NCBI
- Ghiasvand and Kate (2014)
  - best result to date on ShARe

Results: Baseline Systems

	ShARe	NCBI
BASELINE	89.5	82.2

Results: Our Approach

	ShARe	NCBI
BASELINE	89.5	82.2
OUR SYSTEM		
Sieve 1 (Exact Match)	84.04	69.71
+ Sieve 2 (Abbreviation)	86.13	74.17
+ Sieve 3 (Word Reordering)	86.40	74.27
+ Sieve 4 (Numbers Replacement)	86.45	75.00
+ Sieve 5 (Hyphenation)	86.62	75.21
+ Sieve 6 (Suffixation)	88.11	75.62
+ Sieve 7 (Synonyms Replacement)	88.45	76.56
+ Sieve 8 (Stemming)	90.47	77.70
+ Sieve 9 (Composite Mentions/Terms)	90.53	78.00
+ Sieve 10 (Partial Match)	**90.75**	**84.65**

Results: Our Approach

	ShARe	NCBI
BASELINE	89.5	82.2
OUR SYSTEM		
Sieve 1 (Exact Match)	84.04	69.71
+ Sieve 2 (Abbreviation)	86.13	74.17
+ Sieve 3 (Word Reordering)	86.40	74.27
+ Sieve 4 (Numbers Replacement)	86.45	75.00
+ Sieve 5 (Hyphenation)	86.62	75.21
+ Sieve 6 (Suffixation)	88.11	75.62
+ Sieve 7 (Synonyms Replacement)	88.45	76.56
+ Sieve 8 (Stemming)	90.47	77.70
+ Sieve 9 (Composite Mentions/Terms)	90.53	78.00
+ Sieve 10 (Partial Match)	**90.75**	**84.65**

Results: Our Approach

	ShARe	NCBI
BASELINE	89.5	82.2
OUR SYSTEM		
Sieve 1 (Exact Match)	84.04	69.71
+ Sieve 2 (Abbreviation)	86.13	74.17
+ Sieve 3 (Word Reordering)	86.40	74.27
+ Sieve 4 (Numbers Replacement)	86.45	75.00
+ Sieve 5 (Hyphenation)	86.62	75.21
+ Sieve 6 (Suffixation)	88.11	75.62
+ Sieve 7 (Synonyms Replacement)	88.45	76.56
+ Sieve 8 (Stemming)	90.47	77.70
+ Sieve 9 (Composite Mentions/Terms)	90.53	78.00
+ Sieve 10 (Partial Match)	**90.75**	**84.65**

Results: Our Approach

	ShARe	NCBI
BASELINE	89.5	82.2
OUR SYSTEM		
Sieve 1 (Exact Match)	84.04	69.71
+ Sieve 2 (Abbreviation)	86.13	74.17
+ Sieve 3 (Word Reordering)	86.40	74.27
+ Sieve 4 (Numbers Replacement)	86.45	75.00
+ Sieve 5 (Hyphenation)	86.62	75.21
+ Sieve 6 (Suffixation)	88.11	75.62
+ Sieve 7 (Synonyms Replacement)	88.45	76.56
+ Sieve 8 (Stemming)	90.47	77.70
+ Sieve 9 (Composite Mentions/Terms)	90.53	78.00
+ Sieve 10 (Partial Match)	**90.75**	**84.65**

Results: Our Approach

	ShARe	NCBI
BASELINE	89.5	82.2
OUR SYSTEM		
Sieve 1 (Exact Match)	84.04	69.71
+ Sieve 2 (Abbreviation)	86.13	74.17
+ Sieve 3 (Word Reordering)	86.40	74.27
+ Sieve 4 (Numbers Replacement)	86.45	75.00
+ Sieve 5 (Hyphenation)	86.62	75.21
+ Sieve 6 (Suffixation)	88.11	75.62
+ Sieve 7 (Synonyms Replacement)	88.45	76.56
+ Sieve 8 (Stemming)	90.47	77.70
+ Sieve 9 (Composite Mentions/Terms)	90.53	78.00
+ Sieve 10 (Partial Match)	**90.75**	**84.65**

Two Major Sources of Error

- occurs when a mention is mapped to more than one concept in the Partial Match sieve
  - E.g., aspiration → pulmonary aspiration, aspiration pneumonia
- accounts for 11-13% of the errors
- ambiguity arose typically when a shortened form of the entity was used (e.g., when the mention is anaphoric)
  - can be addressed by employing a coreference resolver to find its full name, and normalize the full name instead

Two Major Sources of Error

- occurs when a disorder mention’s string is so lexically dissimilar with the concept terms that none of our heuristics can syntactically transform it into any of them
- accounts for 64-71% of the errors
- Additional information is needed for normalization
  - E.g., query Wikipedia for the mention’s alternate names

Summary

- Presented a simple, modular approach to normalizing disorder mentions, the multi-pass sieve approach
- Achieved state-of-the-art normalization results on two standard datasets
- Released the source code of our system