

Mining Terminological Knowledge in Large Biomedical Corpora

Hongfang Liu, Carol Friedman
Department of Medical Informatics
Columbia University

Presented by:
Azadeh Shakery

Motivation

- Terminological knowledge of biomedical domain is important for NLP and IR applications
- New terms and symbols are continually being created and used without being in the resources.
- Abbreviations are widely used in biomedical domain
- Previous work use manually crafted patterns or rules to find abbreviations
- In this paper they use collocation to automatically extract related terms

Outline

- Background
- Proposed method
- Experiments
- Discussion

Background

Parenthetical Expressions

- In the biomedical literature domain, parenthetical expressions are used to:
 - Define abbreviations : estrogen receptor (**ER**)
 - Semantic relations:
 - Synonymy : natural toxin (i.e. **aflatoxin**)
 - Hypernymy: an inactive HRas protein (**RasN17**)
 - Citations : by using a recently developed ultra sensitive HPLC technique (**Sakhi et al. J. Chromatogr. A 828:451-460, 1998**)
 - Measure : CGRP failed to inhibit glucose-stimulated (**16.7mM**)

Background Collocation

- Collocation: a set of words such that the presence of one or several words of the set often implies or suggests the rest of the collocation.
- Parenthetical expressions are collocations:
 - congestive heart failure (CHF)
- Select collocations using a complex frequency-based method

Proposed Method

- COLLECT
- DETECT
- SEPERATE

Method

- **COLLECT** :
 - Collect parenthetical expressions from a large collection of text and filter out certain expressions
 - Use heuristics to determine sentence boundaries
 - Expressions where the inner-text is occurred only once are eliminated
- **DETECT**
- **SEPERATE**

Method

- COLLECT
- **DETECT** : use the results of the first step to derive a set of pairwise terms
- SEPERATE

DETECT

All unique outer-text strings
that correspond to the same
inner-text and their frequencies



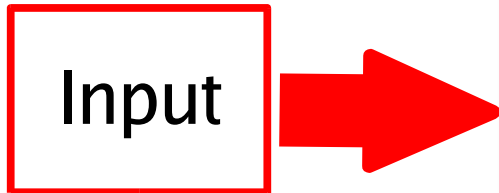
DETECT



Set of pair-wise terms

Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia Patients with community acquired pneumonia patients with community acquired pneumonia group of patients with community acquired pneumonia blood culture of community- acquired pneumonia hospitalized community acquired pneumonia including pneumonia	14 10 3 2 2 1 1
Normalization	(treatment, community, acquired, pneumonia) (14) (patient, community, acquired, pneumonia) (13) (group, patient, community, acquired, pneumonia) (2) (blood, culture, community, acquired, pneumonia) (2) (hospitalize, community, acquired, pneumonia) (1) (include, pneumonia)	14 13 2 2 1 1
Potential Collocations	pneumonia acquired pneumonia include pneumonia community acquired pneumonia treatment community acquired pneumonia patient community acquired pneumonia culture community acquired pneumonia hospitalize community acquired pneumonia blood culture community acquired pneumonia group patient community acquired pneumonia	33 32 1 32 14 15 2 1* 2 2
After Eliminating	pneumonia acquired pneumonia community acquired pneumonia	33 32 32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32

DETECT

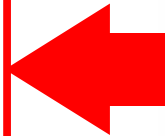


Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia	14
	Patients with community acquired pneumonia	10
	patients with community acquired pneumonia	3
	group of patients with community acquired pneumonia	2
	blood culture of community- acquired pneumonia	2
	hospitalized community acquired pneumonia	1
	including pneumonia	1
Normalization	(treatment, community, acquired, pneumonia) (14)	14
	(patient, community, acquired, pneumonia) (13)	13
	(group, patient, community, acquired, pneumonia) (2)	2
	(blood, culture, community, acquired, pneumonia) (2)	2
	(hospitalize, community, acquired, pneumonia) (1)	1
	(include, pneumonia)	1
Potential Collocations	pneumonia	33
	acquired pneumonia	32
	include pneumonia	1
	community acquired pneumonia	32
	treatment community acquired pneumonia	14
	patient community acquired pneumonia	15
	culture community acquired pneumonia	2
	hospitalize community acquired pneumonia	1*
	blood culture community acquired pneumonia	2
	group patient community acquired pneumonia	2
After Eliminating	pneumonia	33
	acquired pneumonia	32
	community acquired pneumonia	32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32

DETECT

- Purpose: Unify textual variants
- Module:
 - Change an outer-text string into lower case
 - Remove all non-letter characters
 - Remove a small set of stop words
 - Normalize each word by transforming it to the base form

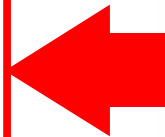
Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia Patients with community acquired pneumonia patients with community acquired pneumonia group of patients with community acquired pneumonia blood culture of community- acquired pneumonia hospitalized community acquired pneumonia including pneumonia	14 10 3 2 2 1 1
Normalization	(treatment, community, acquired, pneumonia) (14) (patient, community, acquired, pneumonia) (13) (group, patient, community, acquired, pneumonia) (2) (blood, culture, community, acquired, pneumonia) (2) (hospitalize, community, acquired, pneumonia) (1) (include, pneumonia)	14 13 2 2 1 1
Potential Collocations	pneumonia acquired pneumonia include pneumonia community acquired pneumonia treatment community acquired pneumonia patient community acquired pneumonia culture community acquired pneumonia hospitalize community acquired pneumonia blood culture community acquired pneumonia group patient community acquired pneumonia	33 32 1 32 14 15 2 1* 2 2
After Eliminating	pneumonia acquired pneumonia community acquired pneumonia	33 32 32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32



DETECT

- Purpose: Generate candidate collocations associated with frequency information
- Method:
 - For each array generate potential collocations
 - Count the number of occurrences for each potential collocation
 - Eliminate all potential collocations that occur only once

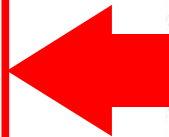
Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia	14
	Patients with community acquired pneumonia	10
	patients with community acquired pneumonia	3
	group of patients with community acquired pneumonia	2
	blood culture of community- acquired pneumonia	2
	hospitalized community acquired pneumonia	1
	including pneumonia	1
Normalization	(treatment, community, acquired, pneumonia) (14)	14
	(patient, community, acquired, pneumonia) (13)	13
	(group, patient, community, acquired, pneumonia) (2)	2
	(blood, culture, community, acquired, pneumonia) (2)	2
	(hospitalize, community, acquired, pneumonia) (1)	1
Potential Collocations	(include, pneumonia)	1
	pneumonia	33
	acquired pneumonia	32
	include pneumonia	1
	community acquired pneumonia	32
	treatment community acquired pneumonia	14
	patient community acquired pneumonia	15
	culture community acquired pneumonia	2
	hospitalize community acquired pneumonia	1*
	blood culture community acquired pneumonia	2
group patient community acquired pneumonia	2	
After Eliminating	pneumonia	33
	acquired pneumonia	32
	community acquired pneumonia	32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32



DETECT

- pc : a collection of k words
- pc': last k - 1 words in pc
- PC(pc): set of potential collocations by adding a word to pc
- If $|PC(pc)| > t_0$
 - Prefix words occur with pc by chance
 - Eliminate collocations where the last k words are the same as pc
- $\text{freq}(pc)/\text{freq}(pc') < t_1$
 - pc is considered to be less frequent
 - Delete pc and all potential collocations where the last k words are the same as pc

Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia Patients with community acquired pneumonia patients with community acquired pneumonia group of patients with community acquired pneumonia blood culture of community- acquired pneumonia hospitalized community acquired pneumonia including pneumonia	14 10 3 2 2 1 1
Normalization	(treatment, community, acquired, pneumonia) (14) (patient, community, acquired, pneumonia) (13) (group, patient, community, acquired, pneumonia) (2) (blood, culture, community, acquired, pneumonia) (2) (hospitalize, community, acquired, pneumonia) (1) (include, pneumonia)	14 13 2 2 1 1
Potential Collocations	pneumonia acquired pneumonia include pneumonia community acquired pneumonia treatment community acquired pneumonia patient community acquired pneumonia culture community acquired pneumonia hospitalize community acquired pneumonia blood culture community acquired pneumonia group patient community acquired pneumonia	33 32 1 32 14 15 2 1* 2 2
After Eliminating	pneumonia acquired pneumonia community acquired pneumonia	33 32 32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32



DETECT

Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia	14
	Patients with community acquired pneumonia	10
	patients with community acquired pneumonia	3
	group of patients with community acquired pneumonia	2
	blood culture of community- acquired pneumonia	2
	hospitalized community acquired pneumonia	1
	including pneumonia	1
Normalization	(treatment, community, acquired, pneumonia) (14)	14
	(patient, community, acquired, pneumonia) (13)	13
	(group, patient, community, acquired, pneumonia) (2)	2
	(blood, culture, community, acquired, pneumonia) (2)	2
	(hospitalize, community, acquired, pneumonia) (1)	1
	(include, pneumonia)	1
Potential Collocations	pneumonia	33
	acquired pneumonia	32
	include pneumonia	1
	community acquired pneumonia	32
	treatment community acquired pneumonia	14
	patient community acquired pneumonia	15
	culture community acquired pneumonia	2
	hospitalize community acquired pneumonia	1*
	blood culture community acquired pneumonia	2
group patient community acquired pneumonia	2	
After Eliminating	pneumonia	33
	acquired pneumonia	32
	community acquired pneumonia	32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32

- pc : a collection of k words
- sc: summation of the frequency of all collocations formed by adding one more word to the left of pc
- If $sc / \text{freq}(pc) > t_2$
 - pc is subsumed
 - Delete pc



DETECT

Step	Examples	FREQ
Outer-texts for CAP	treatment of community acquired pneumonia	14
	Patients with community acquired pneumonia	10
	patients with community acquired pneumonia	3
	group of patients with community acquired pneumonia	2
	blood culture of community- acquired pneumonia	2
	hospitalized community acquired pneumonia	1
	including pneumonia	1
Normalization	(treatment, community, acquired, pneumonia) (14)	14
	(patient, community, acquired, pneumonia) (13)	13
	(group, patient, community, acquired, pneumonia) (2)	2
	(blood, culture, community, acquired, pneumonia) (2)	2
	(hospitalize, community, acquired, pneumonia) (1)	1
	(include, pneumonia)	1
Potential Collocations	pneumonia	33
	acquired pneumonia	32
	include pneumonia	1
	community acquired pneumonia	32
	treatment community acquired pneumonia	14
	patient community acquired pneumonia	15
	culture community acquired pneumonia	2
	hospitalize community acquired pneumonia	1*
	blood culture community acquired pneumonia	2
	group patient community acquired pneumonia	2
After Eliminating	pneumonia	33
	acquired pneumonia	32
	community acquired pneumonia	32
After Subsuming	community acquired pneumonia	32
Final Collocation	(CAP, community acquired pneumonia)	32



Output

Method

- COLLECT
- DETECT
- **SEPERATE** : assess the set of pair-wise terms and separate them into two sets:
 - (abbreviation, expansion) pairs
 - Other types of related terms (synonyms and hyponyms)

Experiments

- Used the MEDLINE free-text collection for the experiments
- Used two abbreviation collections:
 - Berman's collection (pathology related abbreviations)
 - LocusLink collection (pairs of gene symbols with their definitions)

Results

- Expansions associated with 96.3% of the pairs were detected correctly
- The recall of the method was around 88.5%
- The coverage of the acquired abbreviation knowledge:
 - 38.3% for Berman's list
 - 3% for LocusLink collection

Discussion

- Their method has the advantage of not requiring manually crafted patterns or rules
- They have not compared their method with the existing methods
- The method does not recognize expansions that occur only once in the corpora
- They do not study the case where pairs are not abbreviations

Discussion Questions

- Can their method extend to cases other than parenthetical expressions?
- How can we distinguish between relations other than abbreviations?
- How can we cover expansions that occur only once in the corpora?

Thank You!