QuickUMLS: a Fast, Unsupervised Approach for Medical Concept Extraction

Luca Soldaini and **Nazli Goharian**

@soldni

Information Retrieval Lab
Georgetown University

2016-07-21 *MedIR '16*

Task

Medical Information Extraction (*MIE*):

Extract concepts and their location from medical document

```
A 47 year old male who fell on his left arm
presents with pain and bruising on the elbow,
swelling, and inability to bend the arm.
```

Task

Medical Information Extraction (*MIE*):

Extract concepts and their location from medical document

C0230347

A 47 year old male who fell on his left arm

C0030193 C0009938 C0013769

presents with pain and bruising on the elbow,

C0013604 C0560887 C0446516

swelling, and inability to bend the arm.

State of the Art

- *MetaMap* (Aronson 2001, Aronson & Lang 2010)
  - Designed for biomedical text, handles negation, word sense disambiguation
- *cTAKES* (Savova et al 2010)
  - Created for clinical notes
- Focus is on accuracy, not performance
  - OK if 1000s of documents, challenging if more
  - Is real-time analysis a goal?

This Work

- Introduce *QuickUMLS*: unsupervised IE algorithm
- Compared to state-of-the-art:
  - Similar or better performance (Prec, Rec, F1)
  - Significantly faster (2 to 135 times)
  - 500 - 1000 tokens processed per second
- Tests on three datasets: *i2b2*, *THYME*, drug reviews
- Python 2/3 implementation available at: https://github.com/Georgetown-IR-Lab/QuickUMLS

System Overview

Input text

A 47 year old male who fell on his left arm presents with pain and bruising on the elbow, swelling, and inability to bend the arm.

Candidates generation

old male
pain and bruising
inability
inability to bend
bruising
old male who fell
47 year
left arm presents
left arm
pain
47 year old

UMLS concept matching

- **left arm → Left arm structure, C0230347**
- **pain → Pain, C0030193**
- **bruising → Contrusions, C0009938**
- **inability to bend → Ability to bend, C0560887**

Candidates generation

1. Document tokenization and PoS extraction

2. Generate all seq. of tokens with length up to w such that:

   1. contains at least one word & it is not a stopword
   2. not span across sentences
   3. does not start or end with conjunction, adposition, determiner, or punctuation

Woke up in <u>pain</u>

<u>nine</u> bruises

bruise on the <u>left arm</u>.

mark on the <u>left. Arm</u> was bruised

pain in <u>the right arm</u>.

the patient was <u>in pain.</u>

UMLS concept matching

- *CPMerge* (Okazaki and Tsujii, 2010) used for matching sequences to *UMLS* concepts
- For each sequence d, determine the set of concepts C_K such that:

$$StringSimilarity(d, c_{iK}) \geq \alpha \quad \forall c_{iK} \in C_K$$

- For efficiency:
  - strings are tokenized in trigrams and indexed using an inverted index
  - each trigram posting list is partitioned by length of the strings containing the trigram
- In our experiments: *Jaccard* similarity, $0.6 \leq \alpha \leq 1.0$

Experimental Setup

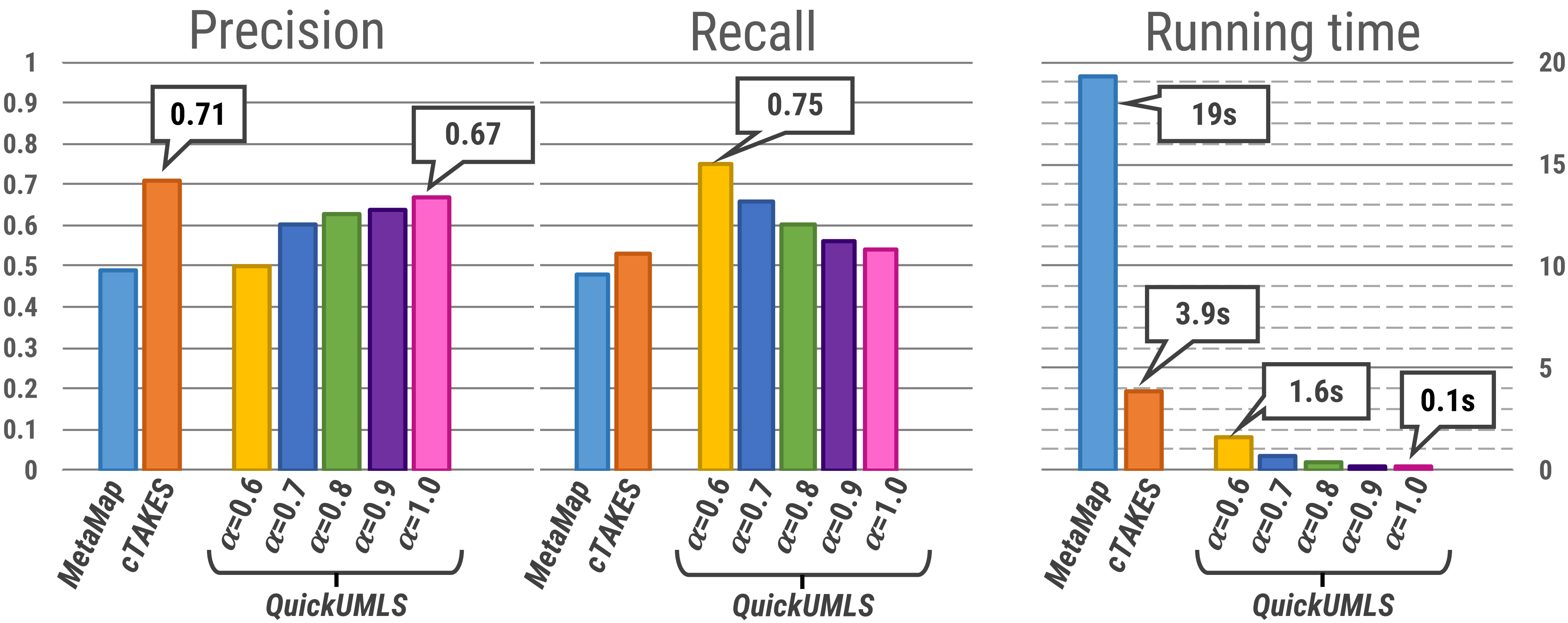
- 2010 *i2b2/VA* Challenge Dataset (Uzuner et al., 2011)
  - 169 annotated with medical concepts for US VA dept.
- *THYME* Corpus (Styler et al., 2014)
  - 1,254 de-identified clinical reports from Mayo Clinic
- Drug Reviews (Yates and Goharian, 2013)
  - 2,500 reviews for *Anastrozole*, *Exemestane*, *Letrozole*, *Raloxifene*, and *Tamoxifen*
  - *generated by laypeople, annotated for drugs side effects*

	Tokens per doc	**Concepts** per doc
i2b2 dataset	1,040	99
THYME corpus	1,035	172
***Drug* Reviews**	131	2

Experimental Setup

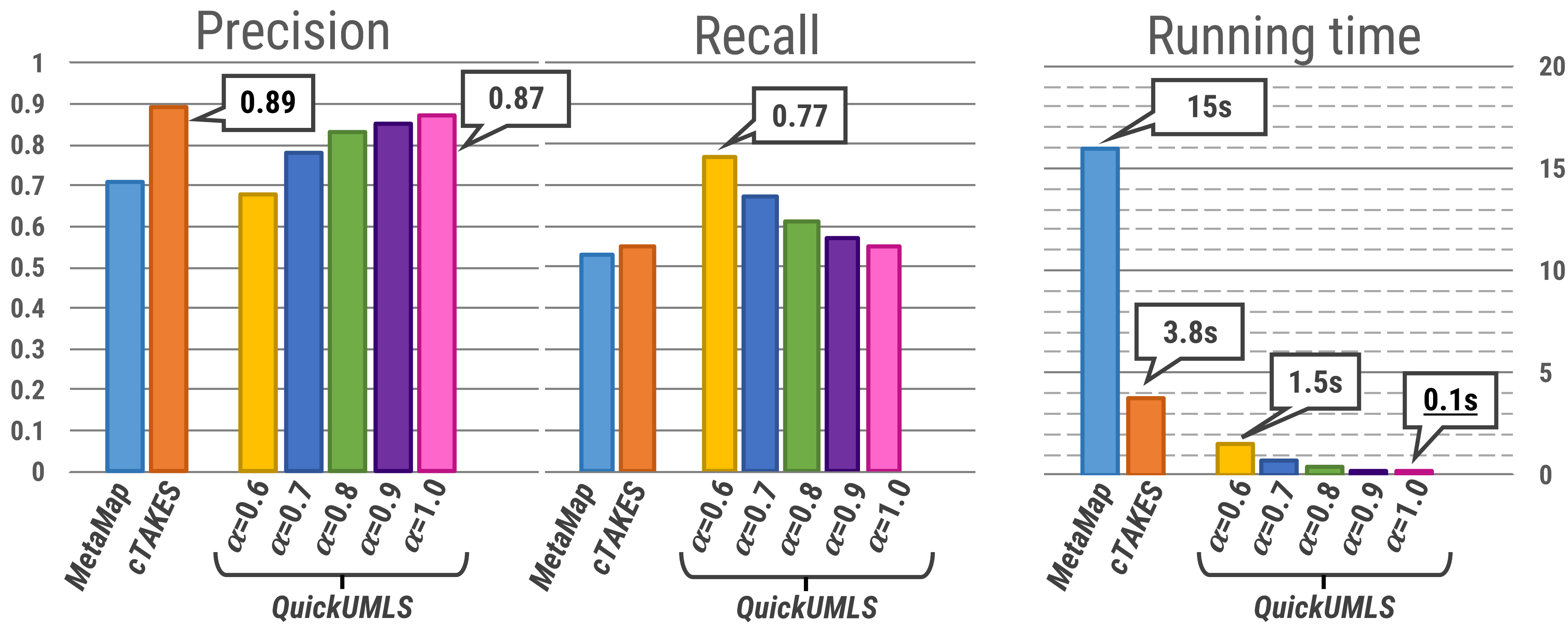
- *SpaCy* for tokenization, parsing, and chunking
  - v.0.100.7, https://spacy.io/
- *MetaMap*
  - v.2016, UMLS 2015AB release, NegEx processing
  - Phrase chunking done with SpaCy (much faster)
- *cTAKES*
  - v.3.2.2, FastUMLSProcessor pipeline.

Results – *i2b2*



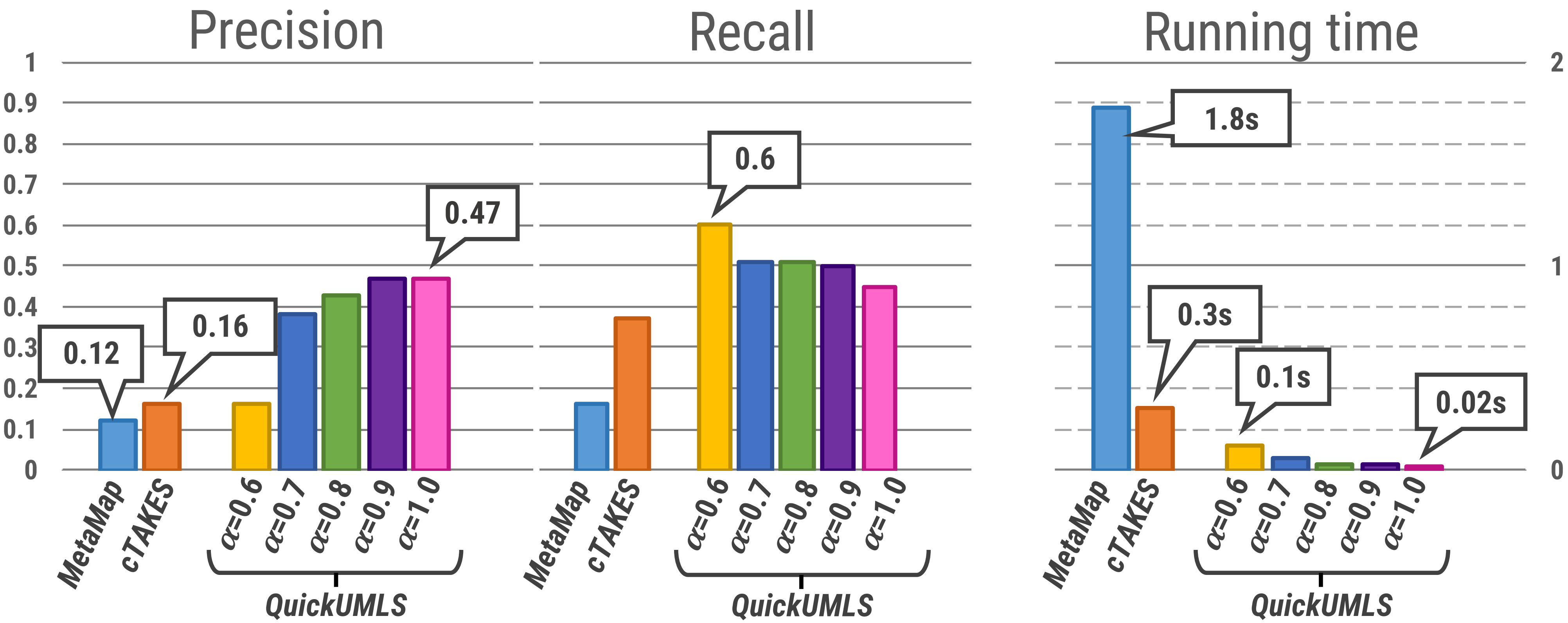
- *cTAKES* has the best precision
- *QuickUMLS* has best recall, close to *cTAKES* when $\alpha = 1.0$
- F1: QuickUMLS = 0.63, cTAKES = 0.61, MetaMap = 0.48
- Small α: more matches, better recall, lower precision, slower

Results – *Thyme*



- Results are similar to i2b2
- *cTAKES* has still the best precision, *QuickUMLS* best recall
- F1: *QuickUMLS* = 0.72, *cTAKES* = 0.68*, *MetaMap* = 0.61*
- *QuickUMLS* is 2-26 times faster than *cTAKES*

Results – *Drug Reviews*



- Results are worse
  - laypeople content is harder to parse
  - only adverse symptoms are annotated
- F1: QuickUMLS = 0.48, cTAKES = 0.22*, MetaMap = 0.14*
- *QuickUMLS* has the best precision and recall

Conclusions

- *QuickUMLS*: unsupervised concept extraction
  - Uses approximate dictionary mapping to match sequences of tokens to UMLS concepts
- Proposed method performs similarly or better than the state of the art
- 2 to 135 times faster than *cTAKES* or *MetaMap*
- Available at: https://github.com/Georgetown-IR-Lab/QuickUMLS

@soldni luca@ir.cs.georgetown.edu

Method		*Prec*	*Rec*	*F-1*	*ms/doc*
MetaMap		0.49*	0.48*	0.48*	19,295*
cTAKES		**0.71**	0.53*	0.61	3,852*
QuickUMLS	$\alpha = 0.6$	0.50*	**0.75**	0.60	1,594*
	$\alpha = 0.7$	0.60*	0.66*	**0.63**	680*
	$\alpha = 0.8$	0.63*	0.60*	0.61	332*
	$\alpha = 0.9$	0.64*	0.56*	0.60	193*
	$\alpha = 1.0$	0.67*	0.54*	0.60	**143**

i2b2

Method		*Prec*	*Rec*	*F-1*	*ms/doc*
MetaMap		0.71*	0.53*	0.61*	15,935
cTAKES		**0.89**	0.55*	0.68*	3,765*
QuickUMLS	$\alpha = 0.6$	0.68*	**0.77**	**0.72**	1,536*
	$\alpha = 0.7$	0.78*	0.67*	**0.72**	646*
	$\alpha = 0.8$	0.83*	0.61*	$0.70^{\dagger}$	340*
	$\alpha = 0.9$	0.85*	0.57*	0.68*	174*
	$\alpha = 1.0$	0.87*	0.55*	0.67*	**141**

THYME

Method		*Prec*	*Rec*	*F-1*	*ms/doc*
MetaMap		0.12*	0.16*	0.14*	1,774*
cTAKES		0.16*	0.37*	0.22*	301*
QuickUMLS	$\alpha = 0.6$	0.16*	**0.6**	0.25*	116*
	$\alpha = 0.7$	0.38*	0.51	0.44*	57*
	$\alpha = 0.8$	0.43	0.51	0.47	32*
	$\alpha = 0.9$	**0.47**	0.50	**0.48**	22
	$\alpha = 1.0$	**0.47**	0.45*	0.46	**18**

Drug Reviews