A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain

Denis R Griffis, Chaitanya Shivade,
Eric Fosler-Lussier, Albert M Lai

AMIA Joint Summits on Translational Science
March 22, 2016

The Ohio State University
Department of Computer Science and Engineering
Department of Biomedical Informatics
NIH National Institutes of Health Clinical Center

Outline

What is Sentence Boundary Detection (SBD)?

UNIX SYSTEM LABS PICKS JUNE D-DAY, WOOS IBM, HP, DEC

Unix System Laboratories Inc has picked Tuesday June 16 to launch Destiny, its desktop system now officially designated SVR4.2. A roll-out is expected on the West Coast in either San Francisco or around San Jose, California, near the time of the Xhibition X-Windows show which will be held there that week. USL is hoping to collect an impressive array of godparents to stand witness. DEC, Hewlett-Packard Co and IBM have yet to agree to adopt the software, but USL is trying to get their representatives there in a show of solidarity and support for the operating system. A magnanimous gesture from the founders of the Open Software Foundation is needed now to heal any lingering breeches in the industry. Destiny is also their one chance to beat back the forces of the Baron of Bellevue, Bill Gates, and his gathering Microsoft NT hordes. Closed ranks would be USL's pay-off for recent concessions made to the Open Software Foundation's most important technologies.

What is Sentence Boundary Detection (SBD)?

UNIX SYSTEM LABS PICKS JUNE D-DAY, WOOS IBM, HP, DEC

Unix System Laboratories Inc has picked Tuesday June 16 to launch Destiny, its desktop system now officially designated SVR4.2. A roll-out is expected on the West Coast in either San Francisco or around San Jose, California, near the time of the Xhibition X-Windows show which will be held there that week. USL is hoping to collect an impressive array of godparents to stand witness. DEC, Hewlett-Packard Co and IBM have yet to agree to adopt the software, but USL is trying to get their representatives there in a show of solidarity and support for the operating system. A magnanimous gesture from the founders of the Open Software Foundation is needed now to heal any lingering breeches in the industry. Destiny is also their one chance to beat back the forces of the Baron of Bellevue, Bill Gates, and his gathering Microsoft NT hordes. Closed ranks would be USL's pay-off for recent concessions made to the Open Software Foundation's most important technologies.

UNIX SYSTEM LABS PICKS JUNE D-DAY, WOOS IBM, HP, DEC

Unix System Laboratories Inc has picked Tuesday June 16 to launch Destiny, its desktop system now officially designated SVR4.2.

A roll-out is expected on the West Coast in either San Francisco or around San Jose, California, near the time of the Xhibition X-Windows show which will be held there that week.

USL is hoping to collect an impressive array of god-parents to stand witness.

DEC, Hewlett-Packard Co and IBM have yet to agree to adopt the software, but USL is trying to get their representatives there in a show of solidarity and support for the operating system.

A magnanimous gesture from the founders of the Open Software Foundation is needed now to heal any lingering breeches in the industry.

Destiny is also their one chance to beat back the forces of the Baron of Bellevue, Bill Gates, and his gathering Microsoft NT hordes.

Closed ranks would be USL's pay-off for recent concessions made to the Open Software Foundation's most important technologies.

SBD faces challenges in the clinical domain

6/10/1999 12:00:00 AM
GASTROINTESTINAL BLEED
DISCHARGE DIAGNOSIS: SEPSIS.
HISTORY OF THE PRESENT ILLNESS :
She takes lisinopril / hydrochlorothiazide 20/25 mg p.o. q.d. , Vioxx 50 mg p.o. q.d. , Lipitor 10 mg p.o. q.d. , Nortriptyline 25 mg p.o. q.h.s. , Neurontin 300 mg p.o. t.i.d. She had a regular heart rate and rhythm. Her gastrointestinal bleeding issues were investigated with an upper endoscopy which revealed multiple superficial gastic ulcerations consistent with an non-steroidal anti-inflammatory drugs gastopathy.
Dictated By: MAULPLACKAGNELEEB, M. INACHELLE, M.D.

SBD faces challenges in the clinical domain

6/10/1999 12:00:00 AM
GASTROINTESTINAL BLEED
DISCHARGE DIAGNOSIS: SEPSIS.
HISTORY OF THE PRESENT ILLNESS :
She takes lisinopril / hydrochlorothiazide 20/25 mg p.o. q.d. , Vioxx 50 mg p.o. q.d. , Lipitor 10 mg p.o. q.d. , Nortriptyline 25 mg p.o. q.h.s. , Neurontin 300 mg p.o. t.i.d. She had a regular heart rate and rhythm. Her gastrointestinal bleeding issues were investigated with an upper endoscopy which revealed multiple superficial gastic ulcerations consistent with an non-steroidal anti-inflammatory drugs gastopathy.
Dictated By: MAULPLACKAGNELEEB, M. INACHELLE, M.D.

6/10/1999 12:00:00 AM

GASTROINTESTINAL BLEED

DISCHARGE DIAGNOSIS :

SEPSIS.

HISTORY OF THE PRESENT ILLNESS :

She takes lisinopril / hydrochlorothiazide 20/25 mg p.o. q.d. , Vioxx 50 mg p.o. q.d. , Lipitor 10 mg p.o. q.d. , Nortriptyline 25 mg p.o. q.h.s. , Neurontin 300 mg p.o. t.i.d.

She had a regular heart rate and rhythm.

Her gastrointestinal bleeding issues were investigated with an upper endoscopy which revealed multiple superficial gastic ulcerations consistent with an non-steroidal anti-inflammatory drugs gastopathy.

Dictated By :

MAULPLACKAGNELEEB, M. INACHELLE, M.D.

Example "sentences" from different domains

Newswire

USL has had Destiny, initially conceived for Intel Corp platforms, in beta test for some weeks and should start regular deliveries to its OEM customers in July.

Biomedical abstracts

The 5' sequences up to nucleotide -120 of the human and murine IL-16 genes share >84% sequence homology and harbor promoter elements for constitutive and inducible transcription in T cells.

Speech (telephone)

Yeah. Uh-huh. W-, uh, the, the call was probably for her.

Clinical text

The hCG on admission was 30,710 and on 1/19 was 805.

Example "sentences" from different domains

Newswire

USL has had Destiny, initially conceived for Intel Corp platforms, in beta test for some weeks and should start regular deliveries to its OEM customers in July.

Biomedical abstracts

The 5' sequences up to nucleotide -120 of the human and murine IL-16 genes share >84% sequence homology and harbor promoter elements for constitutive and inducible transcription in T cells.

Speech (telephone)

Yeah. Uh-huh. W-, uh, the, the call was probably for her.

Clinical text

The hCG on admission was 30,710 and on 1/19 was 805.

Note: the term "sentence" doesn't always make sense.
Different domains prefer different kinds of segmentation.

SBD needs to adapt to different assumptions

Different text domains have different expectations of

- structure (long/short sentences, discrete sections)
- formatting (variable case, unusual numeric patterns)

GENIA

The 5' sequences up to nucleotide -120 of the human and murine IL-16 genes share >84% sequence homology and harbor promoter elements for constitutive and inducible transcription in T cells.

i2b2

ALT (SGPT) - 249 AST (SGOT) - 147 LD (LDH) - 241 ALK PHOS - 230 AMYLASE - 28 TOT BILI - 0.9 LIPASE - 12 ALBUMIN - 2.6

SBD needs to adapt to different assumptions

Different text domains have different expectations of

- structure (long/short sentences, discrete sections)
- formatting (variable case, unusual numeric patterns)

GENIA

The 5' sequences up to nucleotide -120 of the human and murine IL-16 genes share >84% sequence homology and harbor promoter elements for constitutive and inducible transcription in T cells.

i2b2

ALT (SGPT) - 249 AST (SGOT) - 147 LD (LDH) - 241 ALK PHOS - 230 AMYLASE - 28 TOT BILI - 0.9 LIPASE - 12 ALBUMIN - 2.6

There is no one-size-fits-all approach!

SBD errors have impact far downstream

SBD

Lisinopril./ Hydrochlorothiazide 10 mg., po t.i.d.

SBD errors have impact far downstream

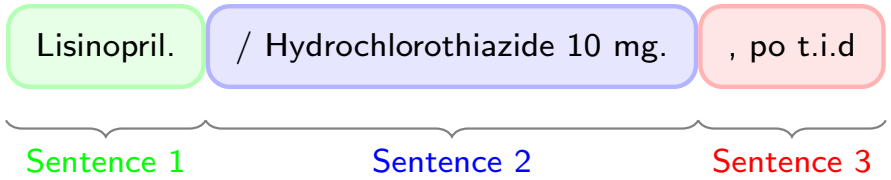
SBD

Lisinopril./ Hydrochlorothiazide 10 mg., po t.i.d.

Sentence 1

SBD errors have impact far downstream

SBD

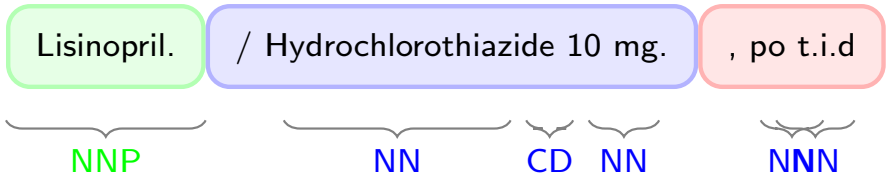


SBD errors have impact far downstream

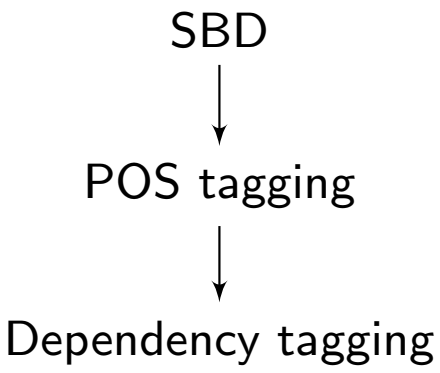
SBD

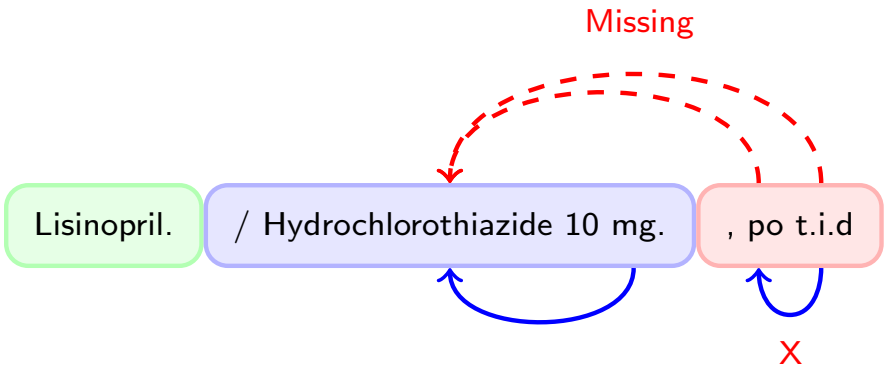
↓

POS tagging

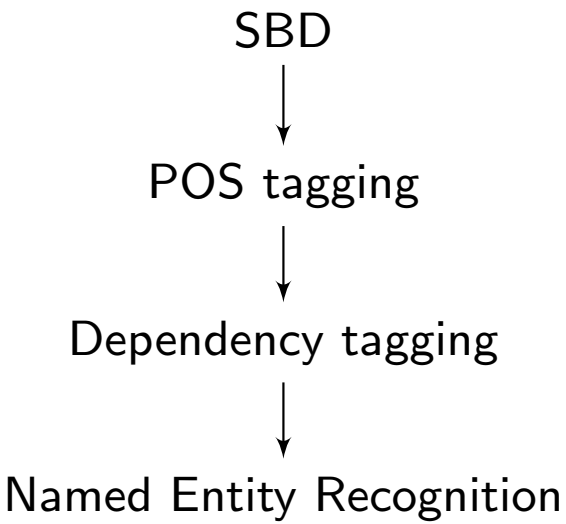


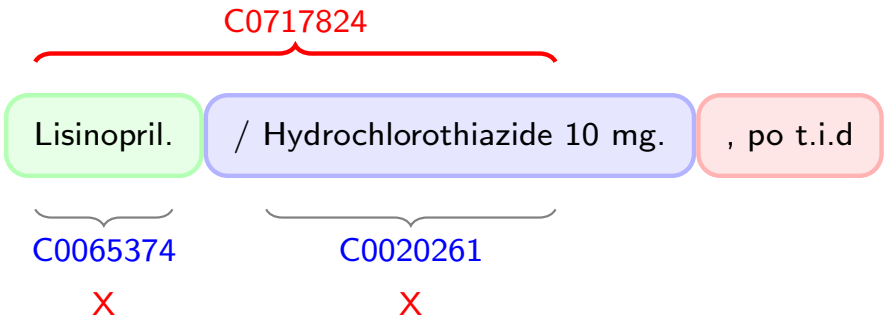
SBD errors have impact far downstream



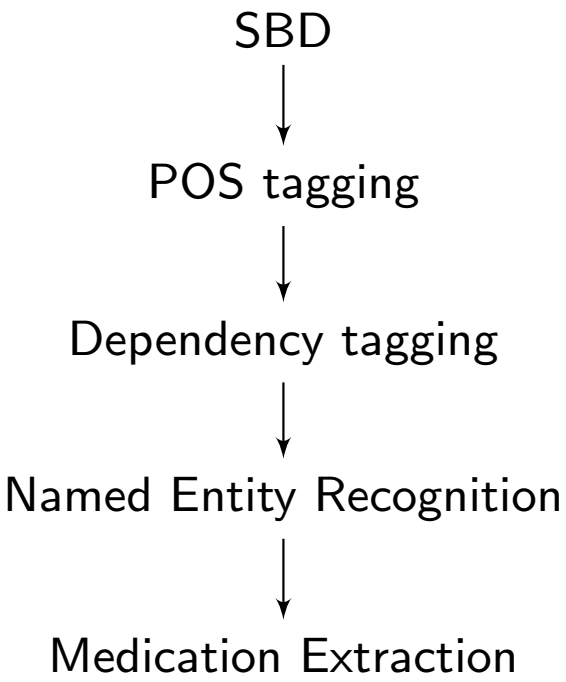


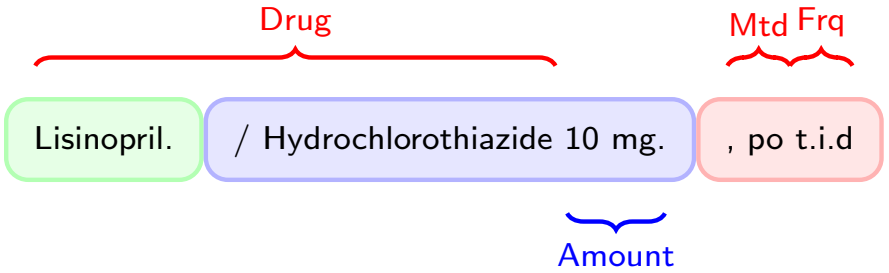
SBD errors have impact far downstream





SBD errors have impact far downstream





Amount: 10 mg

~~Drug: Lisinopril/Hydrochlorothiazide~~

~~Method: po~~

~~Frequency: t.i.d~~

SBD errors have impact far downstream

SBD

↓

POS tagging

↓

Dependency tagging

↓

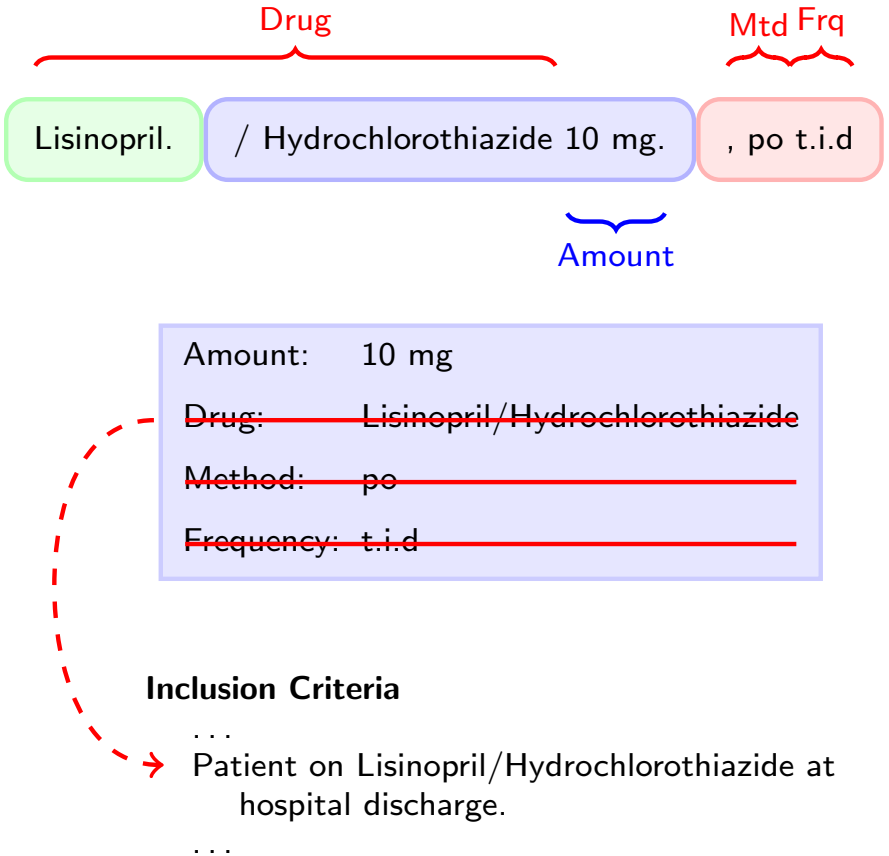
Named Entity Recognition

↓

Medication Extraction

↓

Clinical Trial Eligibility



Why is it time to re-evaluate SBD?

Why is it time to re-evaluate SBD?

SBD is a critical first step for many NLP tasks.

Why is it time to re-evaluate SBD?

SBD is a critical first step for many NLP tasks.

SBD is often treated as “solved” and done with off-the-shelf toolkits.

Why is it time to re-evaluate SBD?

SBD is a critical first step for many NLP tasks.

SBD is often treated as "solved" and done with off-the-shelf toolkits.

But this can lead to serious errors!

Why is it time to re-evaluate SBD?

SBD is a critical first step for many NLP tasks.

SBD is often treated as "solved" and done with off-the-shelf toolkits.

But this can lead to serious errors!

Our goal:
Evaluate off-the-shelf toolkits on SBD,
focusing on clinical text.

Outline

The toolkits

Toolkit	Training Corpora
Stanford CoreNLP	PTB[1], GENIA[2], other Stanford corpora
Lingpipe	MEDLINE abstracts, general text
Splitta	PTB
SPECIALIST	SPECIALIST lexicon[3]
cTAKES	GENIA, PTB, Mayo Clinic EMR

[1] Penn Treebank (PTB): corpus of Wall Street Journal articles
[2] GENIA: corpus of biomedical abstracts
[3] SPECIALIST lexicon: vocabulary from biomedical and general English

General-domain corpora Biomedical corpora Clinical text corpora

The datasets

	Well-formed text corpora	*Non-standard text corpora*
General-domain	BNC	Switchboard
Biomedical	GENIA	i2b2

BNC Mixed-domain British English

Switchboard Spoken English telephone transcripts

GENIA Biomedical abstracts

i2b2 Clinical EHR notes

How we evaluated the toolkits

1. Run each toolkit on each corpus

How we evaluated the toolkits

1. Run each toolkit on each corpus
2. Extract predicted sentence bounds from the output

How we evaluated the toolkits

1. Run each toolkit on each corpus
2. Extract predicted sentence bounds from the output
3. Compare beginning and ending of each sentence against gold standard

How we evaluated the toolkits

1. Run each toolkit on each corpus
2. Extract predicted sentence bounds from the output
3. Compare beginning and ending of each sentence against gold standard

Example text

Patient exhibits mild symptoms.
12.3* m.g. of aspirin administered.

Gold standard		Predicted	
Bgn	End	Bgn	End

How we evaluated the toolkits

1. Run each toolkit on each corpus
2. Extract predicted sentence bounds from the output
3. Compare beginning and ending of each sentence against gold standard

Example text

[Patient exhibits mild symptoms.]
[12.3* m.g. of aspirin administered.]

Gold standard		Predicted	
Bgn	End	Bgn	End
10	40		
41	75		

How we evaluated the toolkits

1. Run each toolkit on each corpus
2. Extract predicted sentence bounds from the output
3. Compare beginning and ending of each sentence against gold standard

Example text

[Patient exhibits mild symptoms.]
[12.] [3* m.g. of aspirin administered.]

Gold standard		Predicted	
Bgn	End	Bgn	End
10	40	10	40
41	75	41	43
-	-	44	75

True Positives: 4
False Positives: 2

F1-score of each toolkit on each corpus

	Well-formed		*Non-standard*	
Toolkit	BNC	GENIA	SWB	i2b2
Stanford	**0.82**	0.98	0.45	0.43
Lingpipe$_{General}$	0.73	0.96	0.42	0.42
Lingpipe$_{Medline}$	0.72	**0.99**	0.43	0.41
Splitta$_{SVM}$	-	0.97	0.39	0.43
Splitta$_{NaiveBayes}$	-	**0.99**	0.39	0.43
SPECIALIST	0.74	0.92	0.46	0.56
cTAKES	0.74	0.68	**0.55**	**0.95**

F1-score of each toolkit on each corpus

	Well-formed		*Non-standard*	
Toolkit	BNC	GENIA	SWB	i2b2
Stanford	**0.82**	0.98	0.45	0.43
Lingpipe$_{General}$	0.73	0.96	0.42	0.42
Lingpipe$_{Medline}$	0.72	**0.99**	0.43	0.41
Splitta$_{SVM}$	-	0.97	0.39	0.43
Splitta$_{NaiveBayes}$	-	**0.99**	0.39	0.43
SPECIALIST	0.74	0.92	0.46	0.56
cTAKES	0.74	0.68	**0.55**	**0.95**

F1-score of each toolkit on each corpus

	Well-formed		*Non-standard*	
Toolkit	BNC	GENIA	SWB	i2b2
Stanford	**0.82**	0.98	0.45	0.43
Lingpipe$_{General}$	0.73	0.96	0.42	0.42
Lingpipe$_{Medline}$	0.72	**0.99**	0.43	0.41
Splitta$_{SVM}$	-	0.97	0.39	0.43
Splitta$_{NaiveBayes}$	-	**0.99**	0.39	0.43
SPECIALIST	0.74	0.92	0.46	0.56
cTAKES	0.74	0.68	**0.55**	**0.95**

F1-score of each toolkit on each corpus

	Well-formed		*Non-standard*	
Toolkit	BNC	GENIA	SWB	i2b2
Stanford	**0.82**	0.98	0.45	0.43
Lingpipe$_{General}$	0.73	0.96	0.42	0.42
Lingpipe$_{Medline}$	0.72	**0.99**	0.43	0.41
Splitta$_{SVM}$	-	0.97	0.39	0.43
Splitta$_{NaiveBayes}$	-	**0.99**	0.39	0.43
SPECIALIST	0.74	0.92	0.46	0.56
cTAKES	0.74	0.68	**0.55**	**0.95**

F1-score of each toolkit on each corpus

	Well-formed		*Non-standard*	
Toolkit	BNC	GENIA	SWB	i2b2
Stanford	**0.82**	0.98	0.45	0.43
Lingpipe$_{General}$	0.73	0.96	0.42	0.42
Lingpipe$_{Medline}$	0.72	**0.99**	0.43	0.41
Splitta$_{SVM}$	-	0.97	0.39	0.43
Splitta$_{NaiveBayes}$	-	**0.99**	0.39	0.43
SPECIALIST	0.74	0.92	0.46	0.56
cTAKES	0.74	0.68	**0.55**	**0.95**

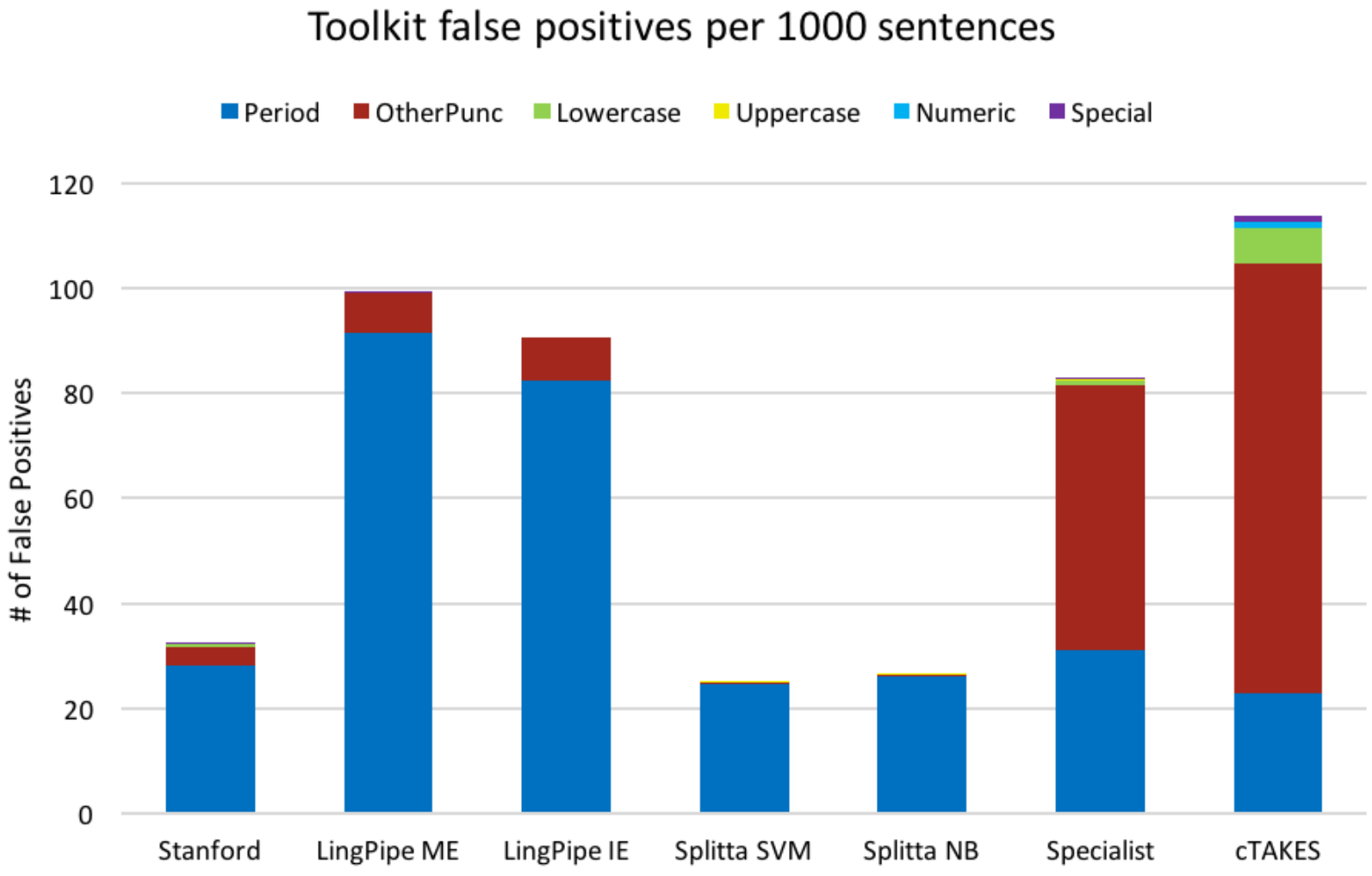
F1-score of each toolkit on each corpus

	Well-formed		*Non-standard*	
Toolkit	BNC	GENIA	SWB	i2b2
Stanford	**0.82**	0.98	0.45	0.43
Lingpipe$_{General}$	0.73	0.96	0.42	0.42
Lingpipe$_{Medline}$	0.72	**0.99**	0.43	0.41
Splitta$_{SVM}$	-	0.97	0.39	0.43
Splitta$_{NaiveBayes}$	-	**0.99**	0.39	0.43
SPECIALIST	0.74	0.92	0.46	0.56
cTAKES	0.74	0.68	**0.55**	**0.95**

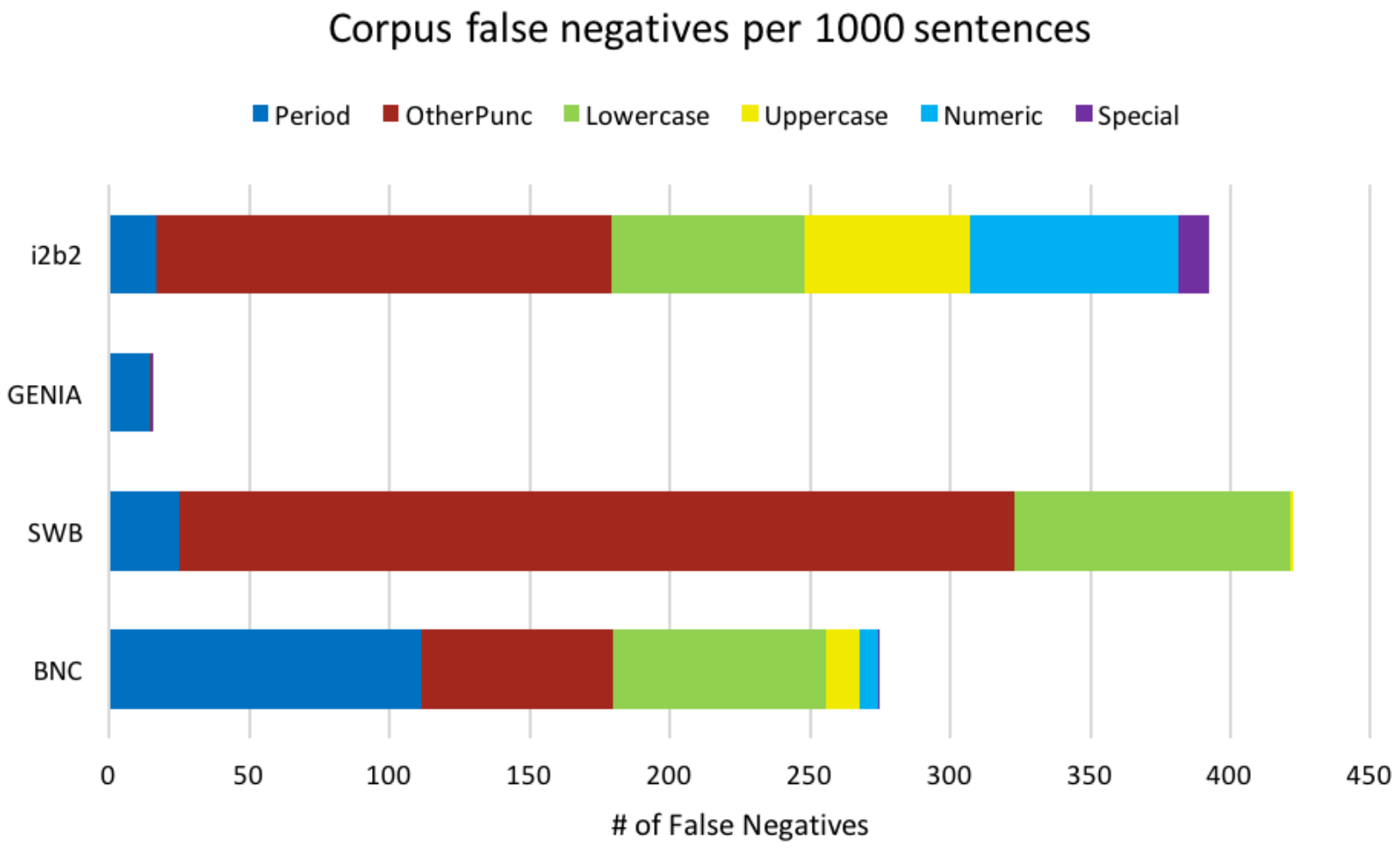
What's going on here?

Outline

SBD is sensitive to punctuation usage



SBD is sensitive to punctuation usage



Sentence length matters

Signed by: DR. Robert Downey on: (WED 2016-05-18 5:18PM)

Sentence 1

Sentence length matters

Signed by: DR. Robert Downey on: (WED 2016-05-18 5:18PM)

Sentence 1

cTAKES

Signed by: DR. | Robert Downey on: | (WED 2016-05-18 5:18PM)

Sentence length matters

Signed by: DR. Robert Downey on: (WED 2016-05-18 5:18PM)

Sentence 1

cTAKES

Signed by: DR. | Robert Downey on: | (WED 2016-05-18 5:18PM)

Stanford

Signed by: DR. Robert Downey on: (WED 2016-05-18 05:18PM) ...

Abbreviations are still problematic

Abbreviations are still problematic

Unfamiliar variants

"t.i.d" is fine, "t.i.d." considered sentence terminal.

Abbreviations are still problematic

Unfamiliar variants

"t.i.d" is fine, "t.i.d." considered sentence terminal.

New abbreviations and initials

St. Cyres

Joseph R. Cowdon

Non-standard formatting causes errors

Non-standard formatting causes errors

Case errors

. . . by DR. Melvin N.I. LICHTENBERGER. . . → Extra breaks

. . . human nm23-H2 gene product. nm23 gene. . . → Missed break

Non-standard formatting causes errors

Case errors

. . . by DR. Melvin N.I. LICHTENBERGER. . . → Extra breaks

. . . human nm23-H2 gene product. nm23 gene. . . → Missed break

---

Numeric formatting (e.g. readings)

. . . 1.1 +/- 0.4 10(-18) . . .

. . . 11.4* . . .

} cause false positives

Non-standard formatting causes errors

Case errors

. . . by DR. Melvin N.I. LICHTENBERGER. . . → Extra breaks

. . . human nm23-H2 gene product. nm23 gene. . . → Missed break

---

Numeric formatting (e.g. readings)

. . . 1.1 +/- 0.4 10(-18) . . .
. . . 11.4* . . .

} cause false positives

Extra headers

. . . murine erythroleukemia (MEL) cells . . . → Caps in parens

(i) item 1
(ii) item 2

→ Lists

Outline

Recap: evaluation of SBD toolkits

Our goal

Evaluate off-the-shelf toolkits on SBD
with a focus towards clinical text.

Recap: evaluation of SBD toolkits

Our goal

Evaluate off-the-shelf toolkits on SBD
with a focus towards clinical text.

To that end

We ran several
popular tools on a
variety of datasets

Recap: evaluation of SBD toolkits

Our goal

Evaluate off-the-shelf toolkits on SBD
with a focus towards clinical text.

To that end

We ran several popular tools on a variety of datasets

We found **domain sensitivity** and **poor overall performance on clinical text**.

Main takeaways

SBD in clinical text faces the challenges of:

Main takeaways

SBD in clinical text faces the challenges of:

- Different patterns of punctuation usage

Main takeaways

SBD in clinical text faces the challenges of:

- Different patterns of punctuation usage
- Different structure and length of “sentences”

Main takeaways

SBD in clinical text faces the challenges of:

- Different patterns of punctuation usage
- Different structure and length of "sentences"
- Different assumptions about text formatting

Main takeaways

SBD in clinical text faces the challenges of:

- Different patterns of punctuation usage
- Different structure and length of "sentences"
- Different assumptions about text formatting
- **Different definition of a useful "sentence"**

Main takeaways

SBD in clinical text faces the challenges of:

- Different patterns of punctuation usage
- Different structure and length of "sentences"
- Different assumptions about text formatting
- **Different definition of a useful "sentence"**

SBD errors negatively impact downstream biomedical applications, e.g.

- Medical event recognition
- Drug interaction mining
- Automated clinical trial eligibility screening
- etc.

Where do we go from here?

Where do we go from here?

Long-term

Where do we go from here?

Long-term

- Work on developing lightweight SBD methods that can be easily adapted to new domains.

Where do we go from here?

Long-term

- Work on developing lightweight SBD methods that can be easily adapted to new domains.
- Create and share datasets of clinical text annotated for SBD.

Where do we go from here?

Long-term

- Work on developing lightweight SBD methods that can be easily adapted to new domains.
- Create and share datasets of clinical text annotated for SBD.

Short-term

Where do we go from here?

Long-term

- Work on developing lightweight SBD methods that can be easily adapted to new domains.
- Create and share datasets of clinical text annotated for SBD.

Short-term

- Add a training pipeline to cTAKES and allow for custom abbreviation lists.

Where do we go from here?

Long-term

- Work on developing lightweight SBD methods that can be easily adapted to new domains.
- Create and share datasets of clinical text annotated for SBD.

Short-term

- Add a training pipeline to cTAKES and allow for custom abbreviation lists.
- Explore new rules for Stanford CoreNLP to work well on clinical text.

Acknowledgments

Co-authors:

Chaitanya Shivade

Eric Fosler-Lussier*

Albert M Lai*

*Co-advisors

Supported by:

- Award Number Grant R01LM011116 from the National Library of Medicine
- The Intramural Research Program of the National Institutes of Health, Clinical Research Center
- Inter-Agency Agreement with the US Social Security Administration

Source code available at:
http://github.com/drgriffis/sbd-evaluation

The Ohio State University

Department of Computer Science and Engineering
Department of Biomedical Informatics

NIH National Institutes of Health Clinical Center

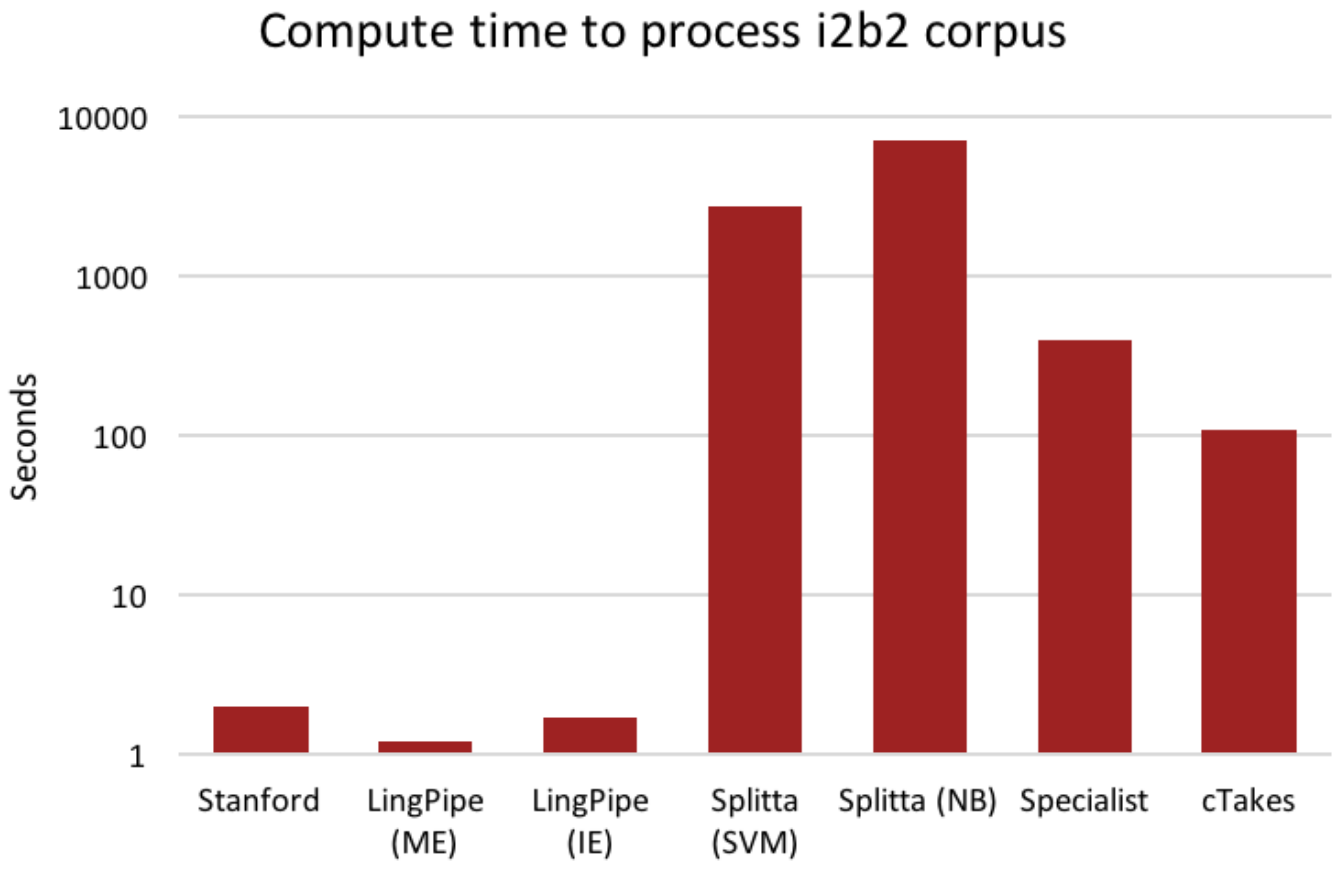
Thank you!

Source code available at:
http://github.com/drgriffis/sbd-evaluation

Contact Info:
Denis Griffis | griffis.30@osu.edu

Department of Computer Science and Engineering
Department of Biomedical Informatics

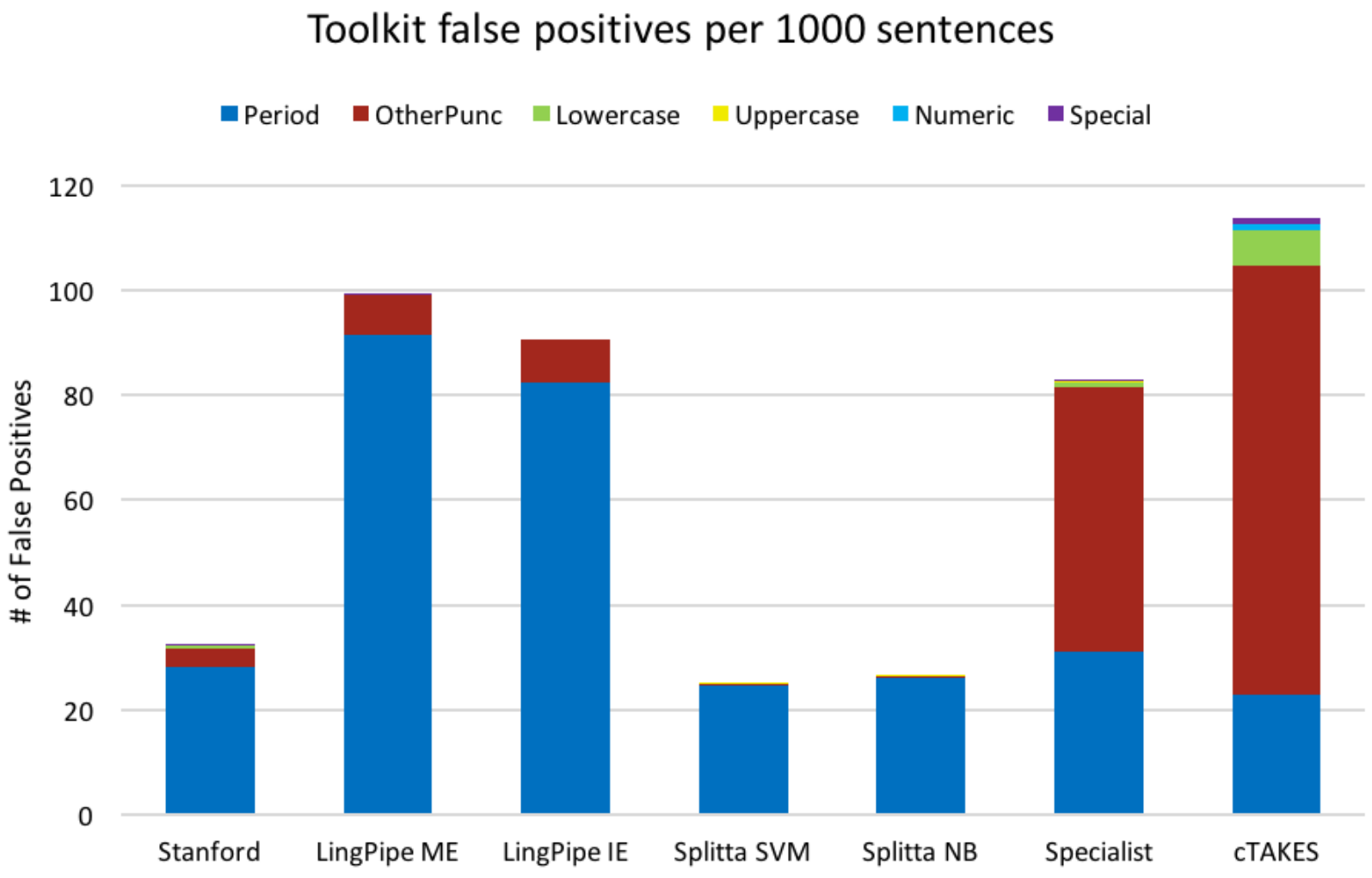
Supplemental: Runtime of each toolkit on i2b2 corpus



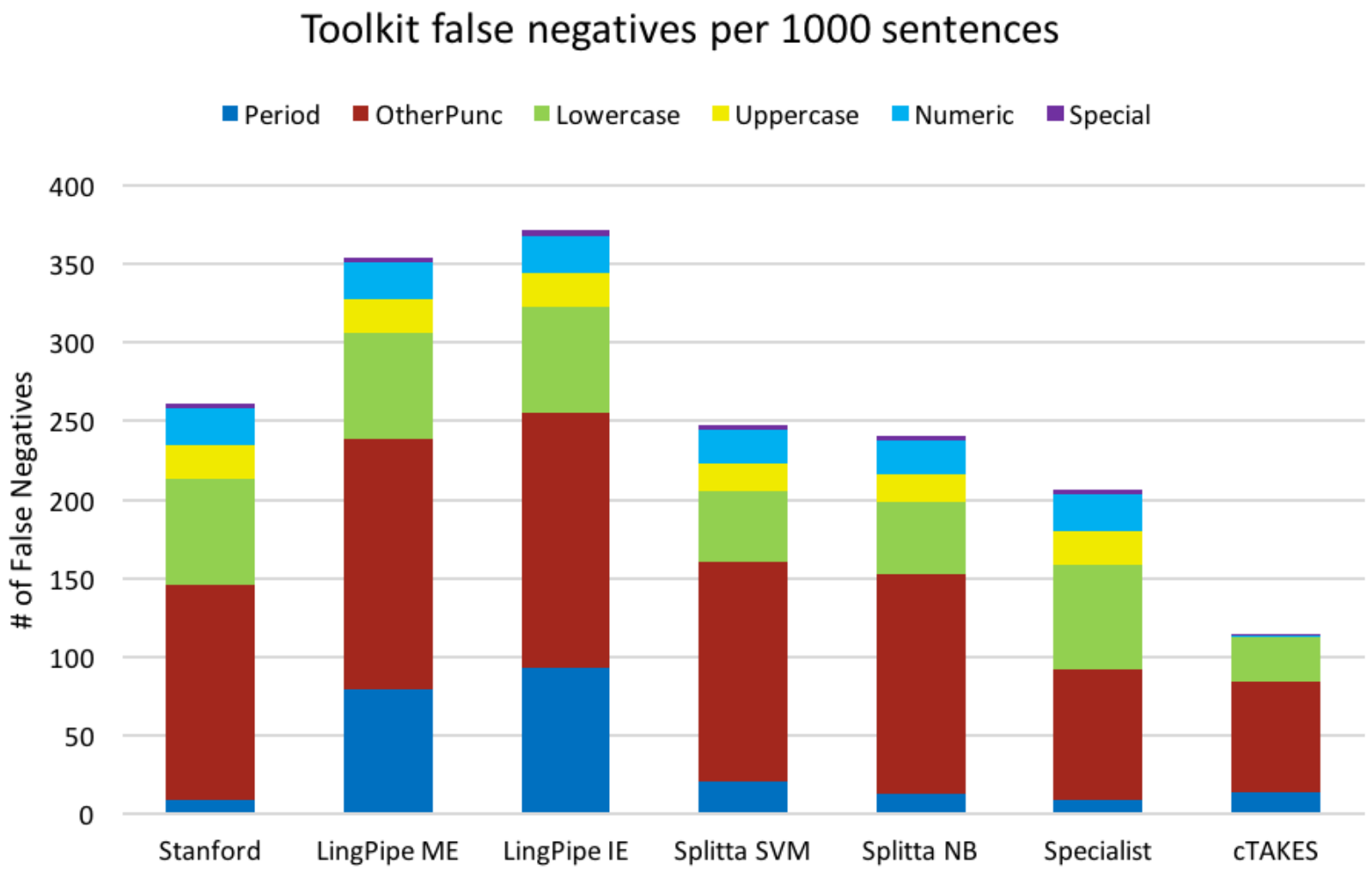
Supplemental: Corpus details

Corpus	**# of Documents**	**# of Sentences**	**Avg. # tokens**
BNC	4,049	6,027,378	16.1
Switchboard	650	110,504	7.4
GENIA	1,999	16,479	24.4
i2b2	426	43,940	9.5

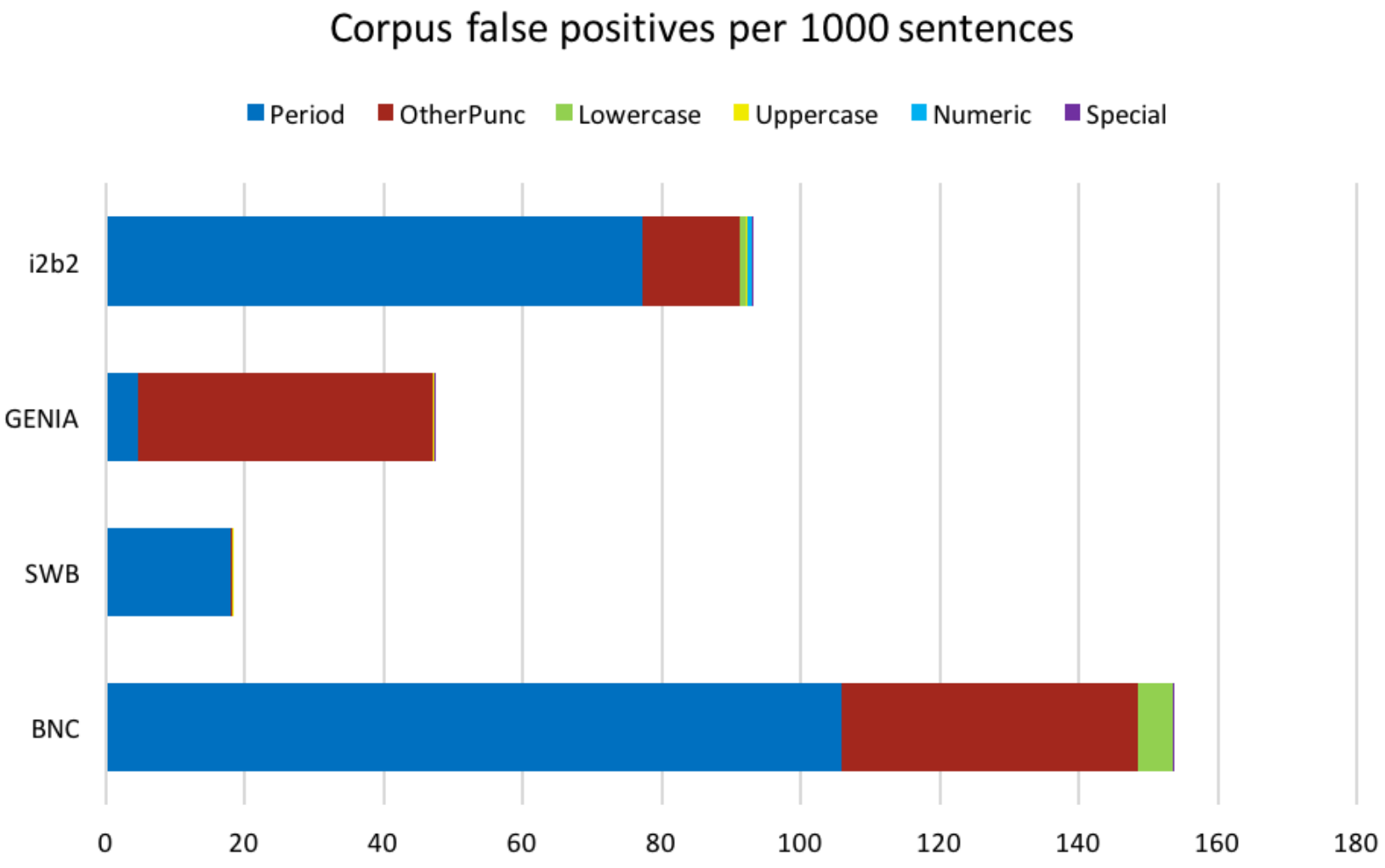
Supplemental: Detailed errors



Supplemental: Detailed errors



Supplemental: Detailed errors



Supplemental: Detailed errors