Identification of medication side effects in clinical records: an experiment based on the 2014 i2b2/UTHealth corpus

Cyril Grouin — Pierre Zweigenbaum

LIMSI–CNRS

14th November 2014

1 Introduction

2 Medication side effects identification

3 Clinical records reformatting

4 Conclusion

Introduction

Possible tracks → *Funny and useful*

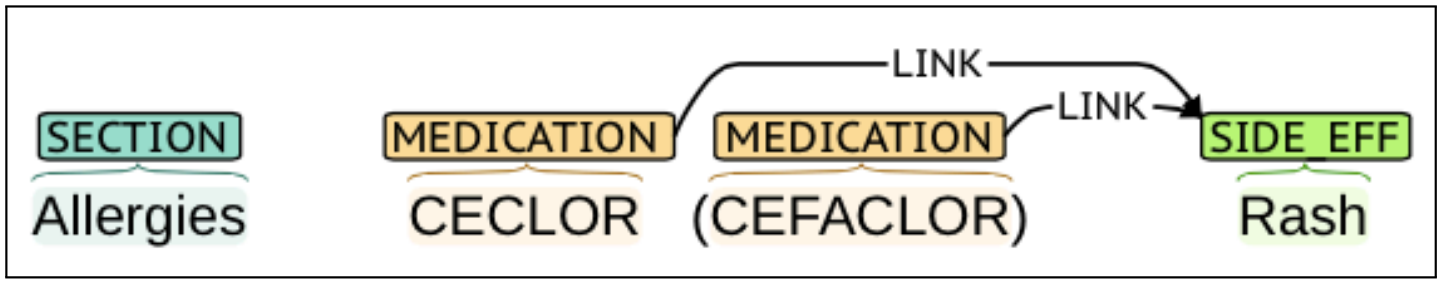
A few possible tracks:

- identification of **patient cohorts**:
  - patient who develop CAD (files 100-199)
  - patient who have CAD in their first record (files 200-299)
  - patient who never develop CAD (files 300-400)
- identification of files belonging to the **same patient** (e.g., based on the content of section *history of present illness*)
  - clinical record de-identification (patient and clinical staff names)
  - file renaming: 100-01, 100-02, 100-03, 100-04 → same patient

Introduction

Two tracks w/ immediate interest → *Useful and useful*

1. detection of **medication side effects**

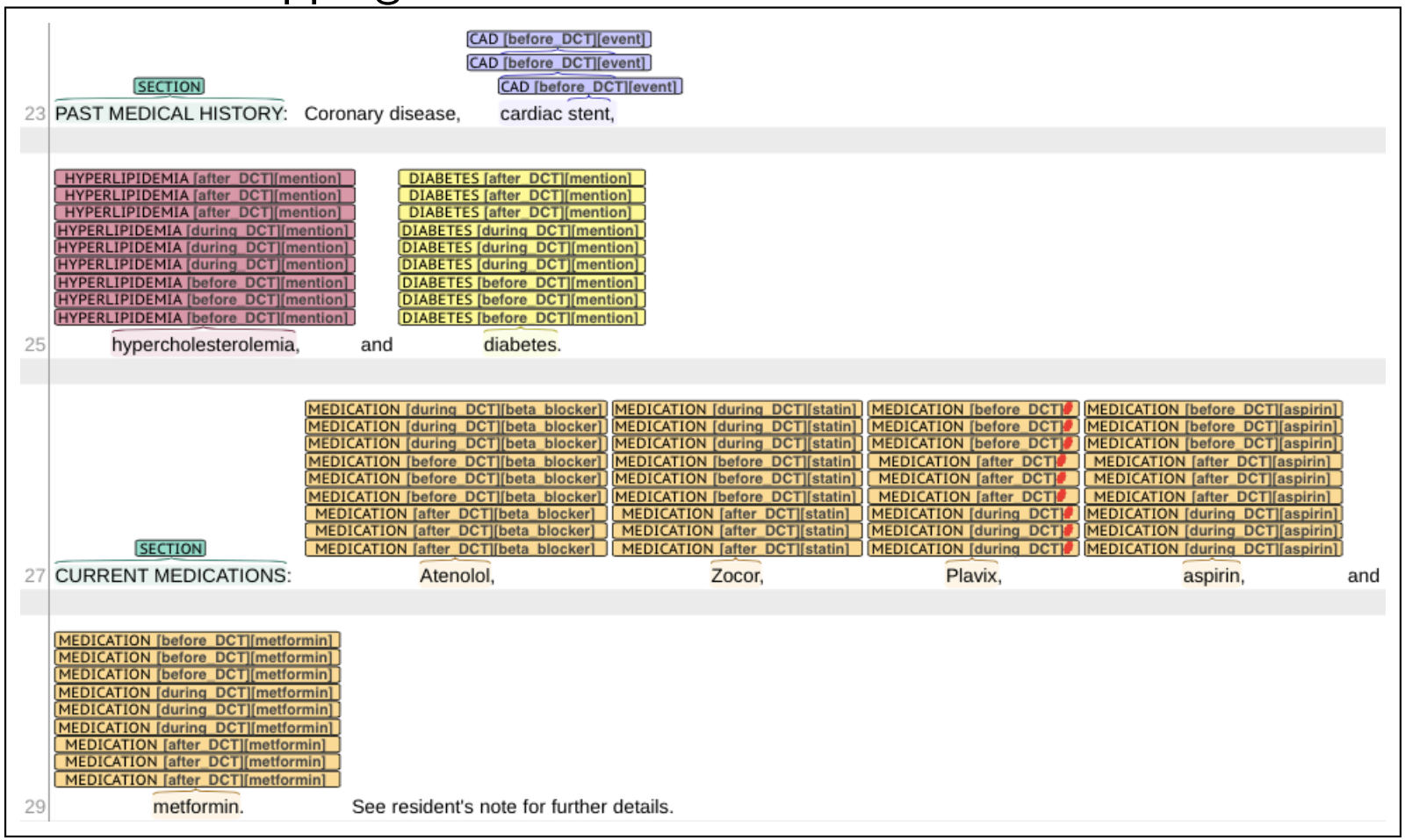


Introduction

Two tracks w/ immediate interest → *Useful and useful*

❷ clinical records reformatting

- double spacing
- line wrapping



Introduction

Footer/Header within clinical texts → *Useful and not funny*

the current time unless asked, and we will focus primarily on the bone situation. Otherwise, we will see the patient back or at least talk to RACHAEL G. OBRYAN, M.D.

Clinical text

VALLEJO GENERAL HOSPITAL
ENDOCRINOLOGY DIVISION Workman, Edwin
315 Lunar Lane 61248040
Brookshire, MO 15345 4-12-84
Page 4

Footer/Header

the patient after we get a bone density and decide when he next needs to be seen.

Clinical text

Introduction

Footer/Header within clinical texts → *Useful and not funny*

the current time unless asked, and we will focus primarily on the bone situation. Otherwise, we will see the patient back or at least talk to

Clinical text

Footer/Header

the patient after we get a bone density and decide when he next needs to be seen.

Clinical text

1 Introduction

2 Medication side effects identification

3 Clinical records reformatting

4 Conclusion

Medication side effects identification

Corpus annotation

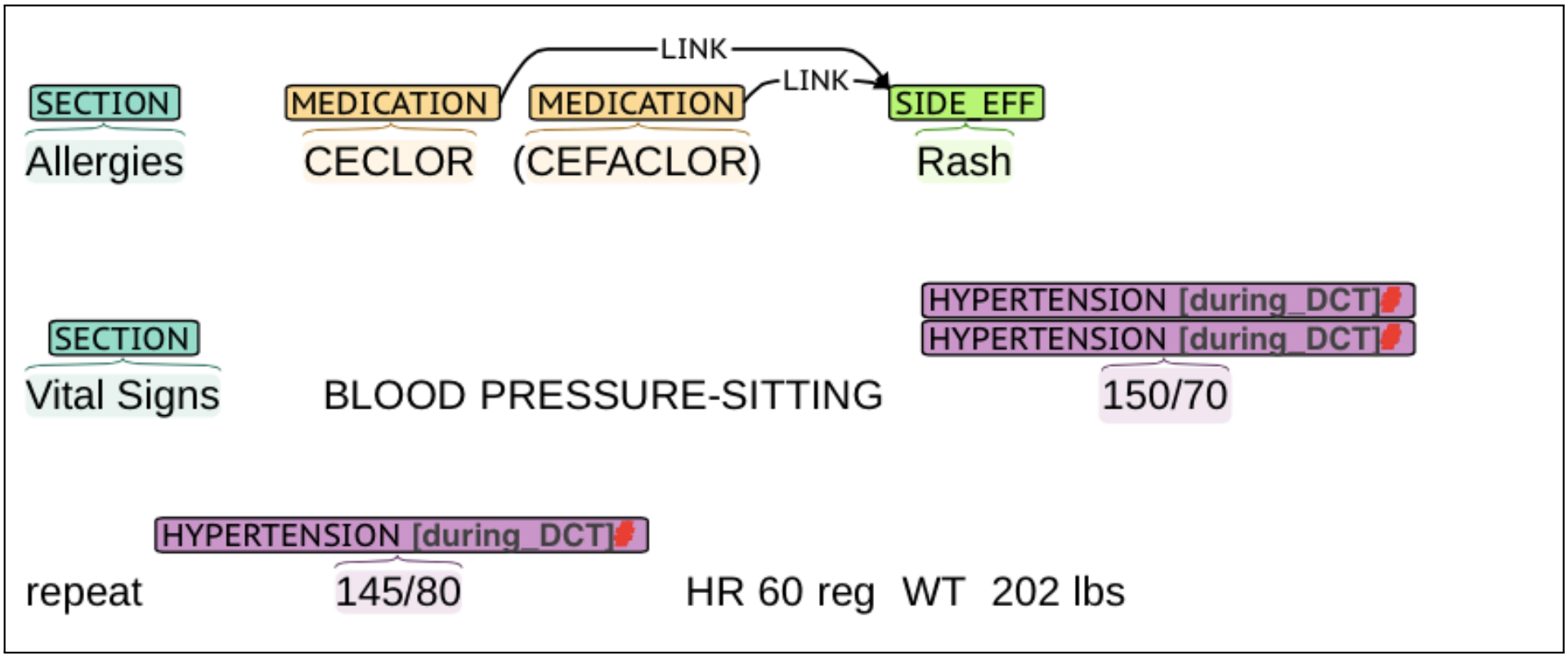
Categories:

- Sections of content: 25 sections based on the SOAP schema
  - annotation of titles of section
- Side effects (positive or negative) produced by the use of drugs
  - side effects are disorders, sign or symptoms, etc.
- Medications when they occur with a side effect
  - medications were generally annotated from the i2b2 annotations

Medication side effects identification

Motivation: medication side effects are found in a few sections

Allergies: known side effects experienced by the patient

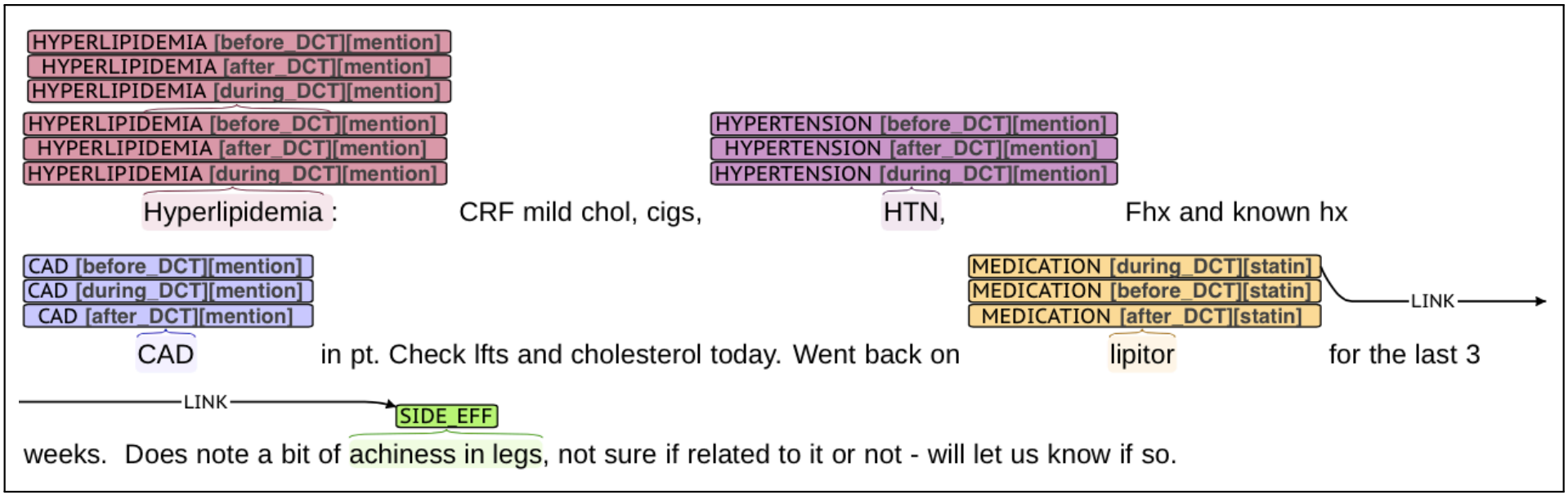


CAD | HYPERLIPIDEMIA | HYPERTENSION | MEDICATION | PHI : existing annotations

Medication side effects identification

Motivation: medication side effects are found in a few sections

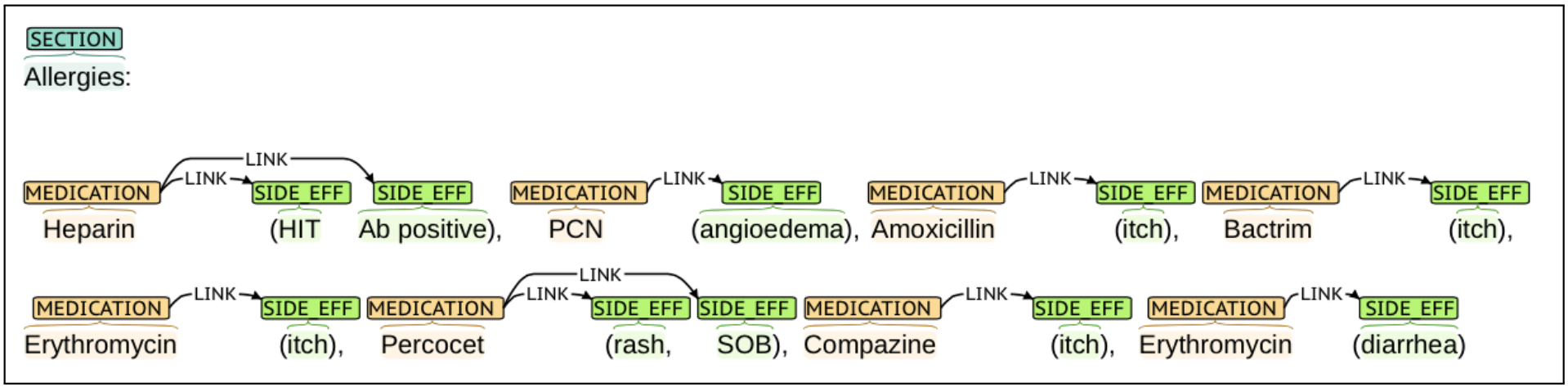
Other sections, mainly at the beginning of records: existing problem possibly linked to a medication

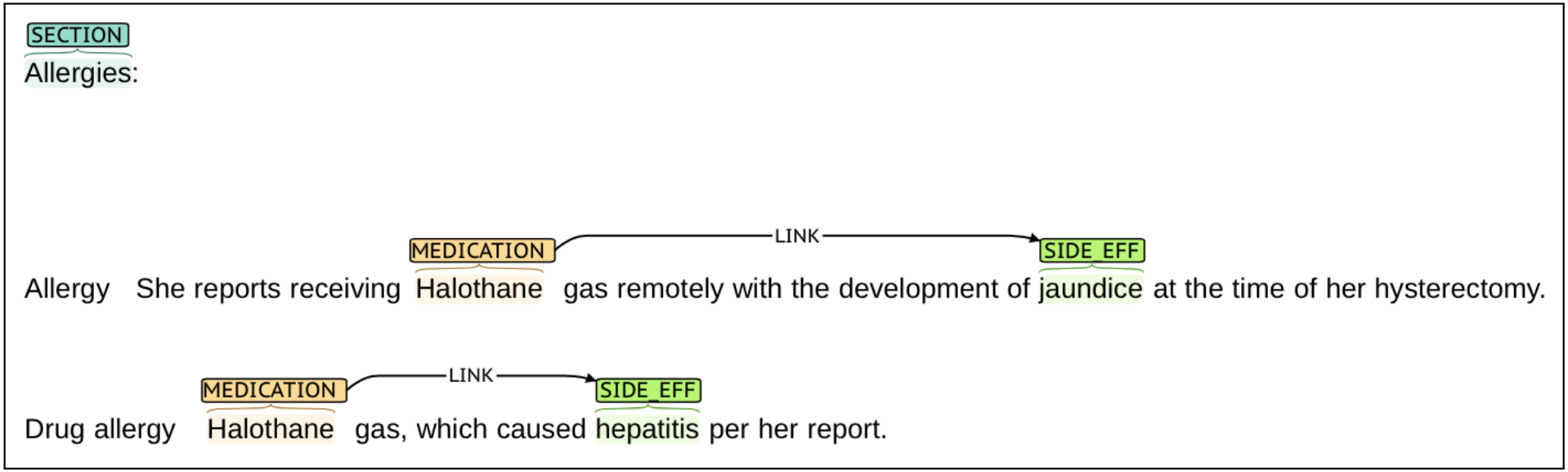


CAD | HYPERLIPIDEMIA | HYPERTENSION | MEDICATION | PHI : existing annotations

Medication side effects identification

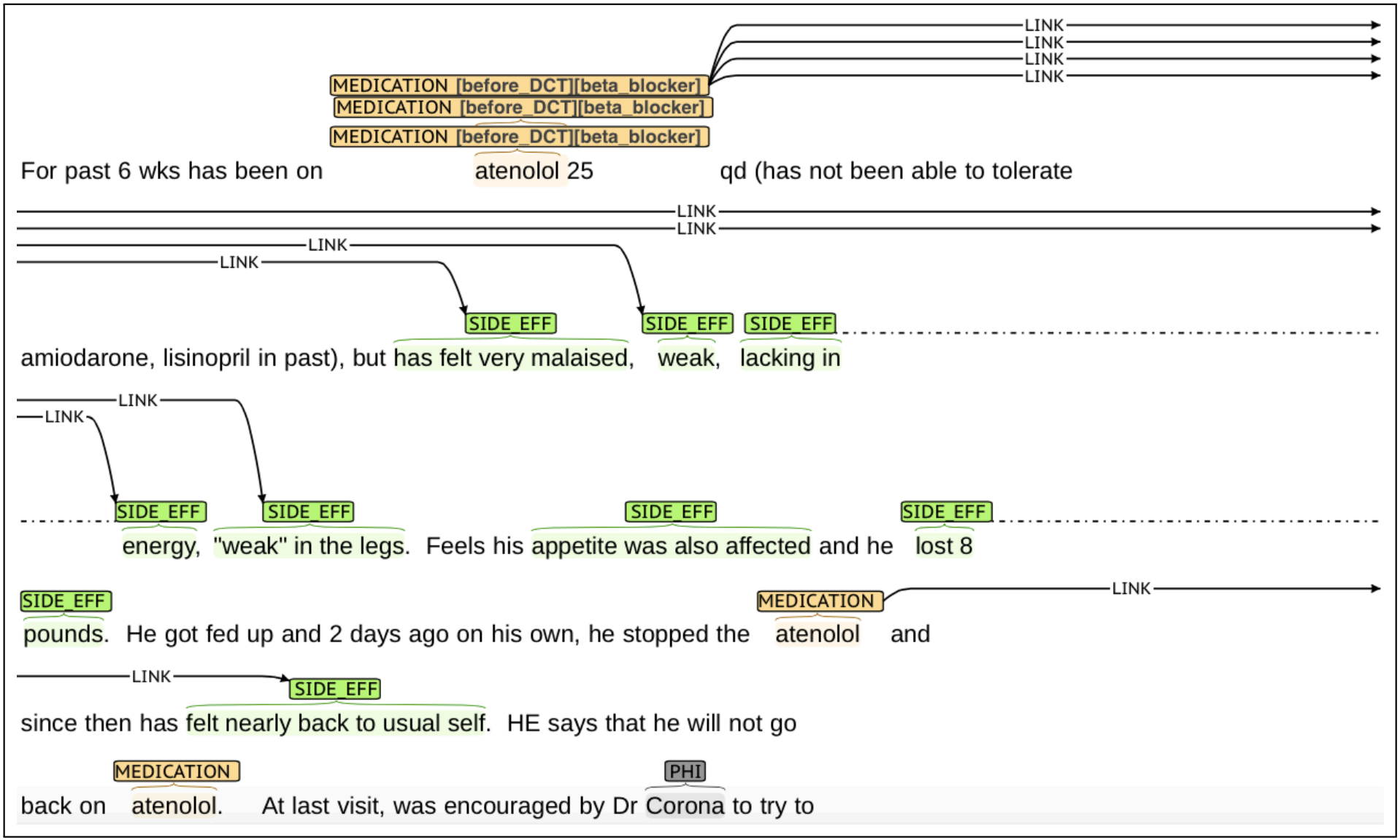
Corpus annotation: simple cases → *Useful annotations*





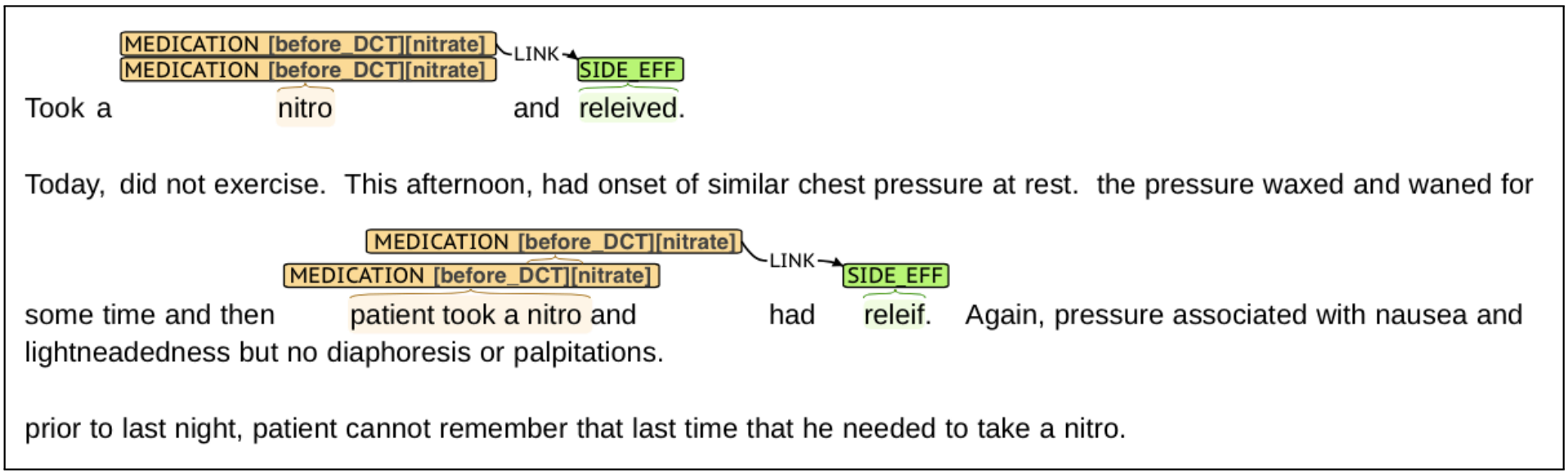
Medication side effects identification

Corpus annotation: variety of side effects → *Further normalization needed*



Medication side effects identification

Corpus annotation: expected effects → *Are annotations useful?*



Medication side effects identification

Corpus annotation

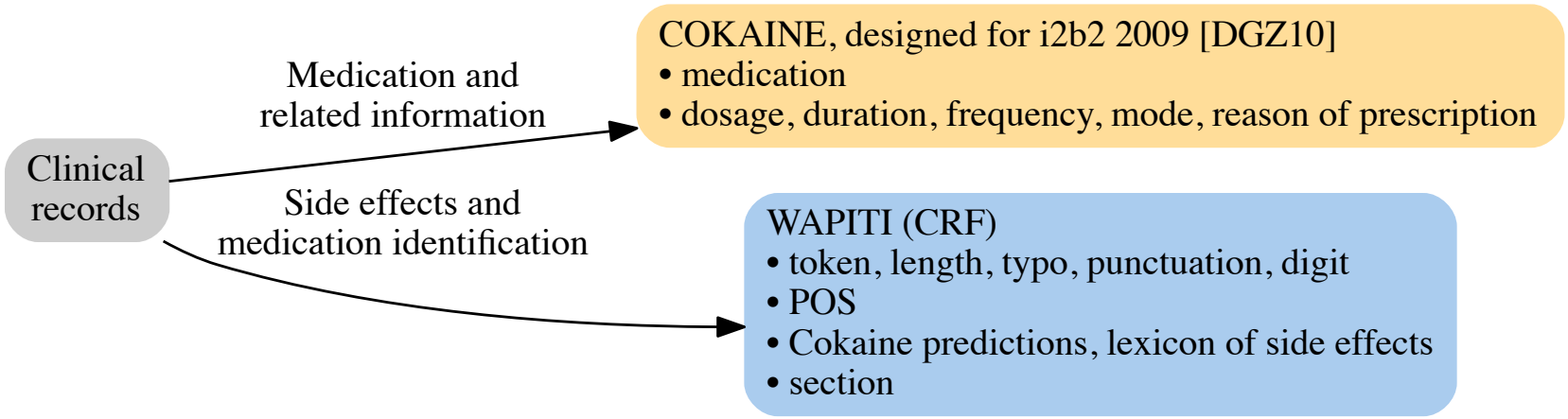
Corpus	Sections of Content		Side Effects	
	Number	Avg Nb/Doc	Number	Avg Nb/Doc
Training set	6,288 (59.4%)	8.0	371 (54.7%)	0.47
Test set	4,292 (40.6%)	8.4	307 (45.3%)	0.60
Overall	10,580 (100%)	8.1	678 (100%)	0.52

Medication side effects identification

Methods

Pipeline:

1. corpus annotation of medications and related information
2. identification of sections of content and side effects



Medication side effects identification

Methods

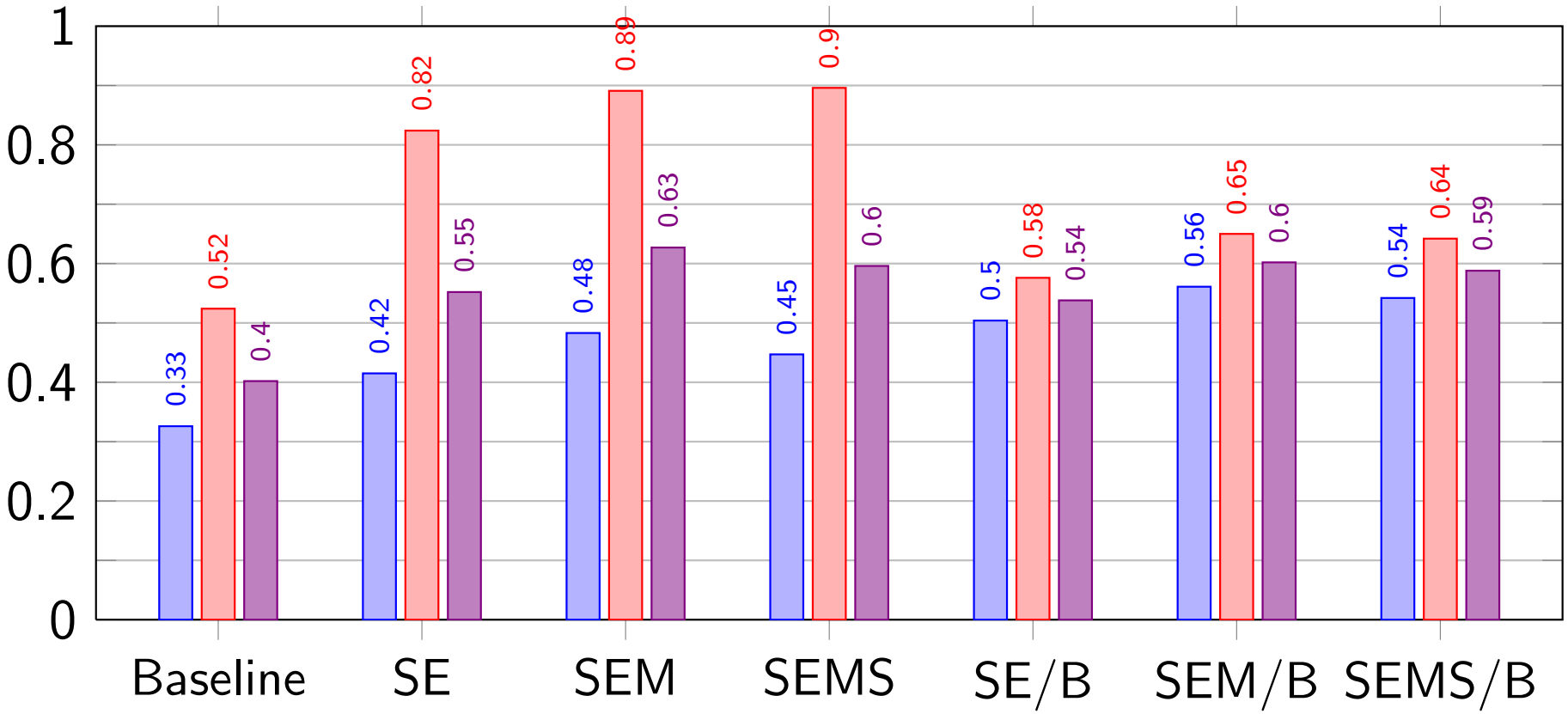
- **Baseline:** rule-based, identification of tokens from the *Allergies* section within the list of 132 common side effects
- **CRF-based:**
  - SE model: predictions of side effects only
  - SEM model: predictions of side effects and medications. As side effects are associated with medications, predicting both medications and side effects would be helpful
  - SEMS model: predictions of all categories: side effects, medications and whether a phrase is a section title or not. We hypothesized that a global model allows to make out all categories
- **Combination:** CRF models + baseline to refine the outputs

Medication side effects identification

Evaluation

Evaluation on the side effects category:

- Precision
- Recall
- F-measure



1 Introduction

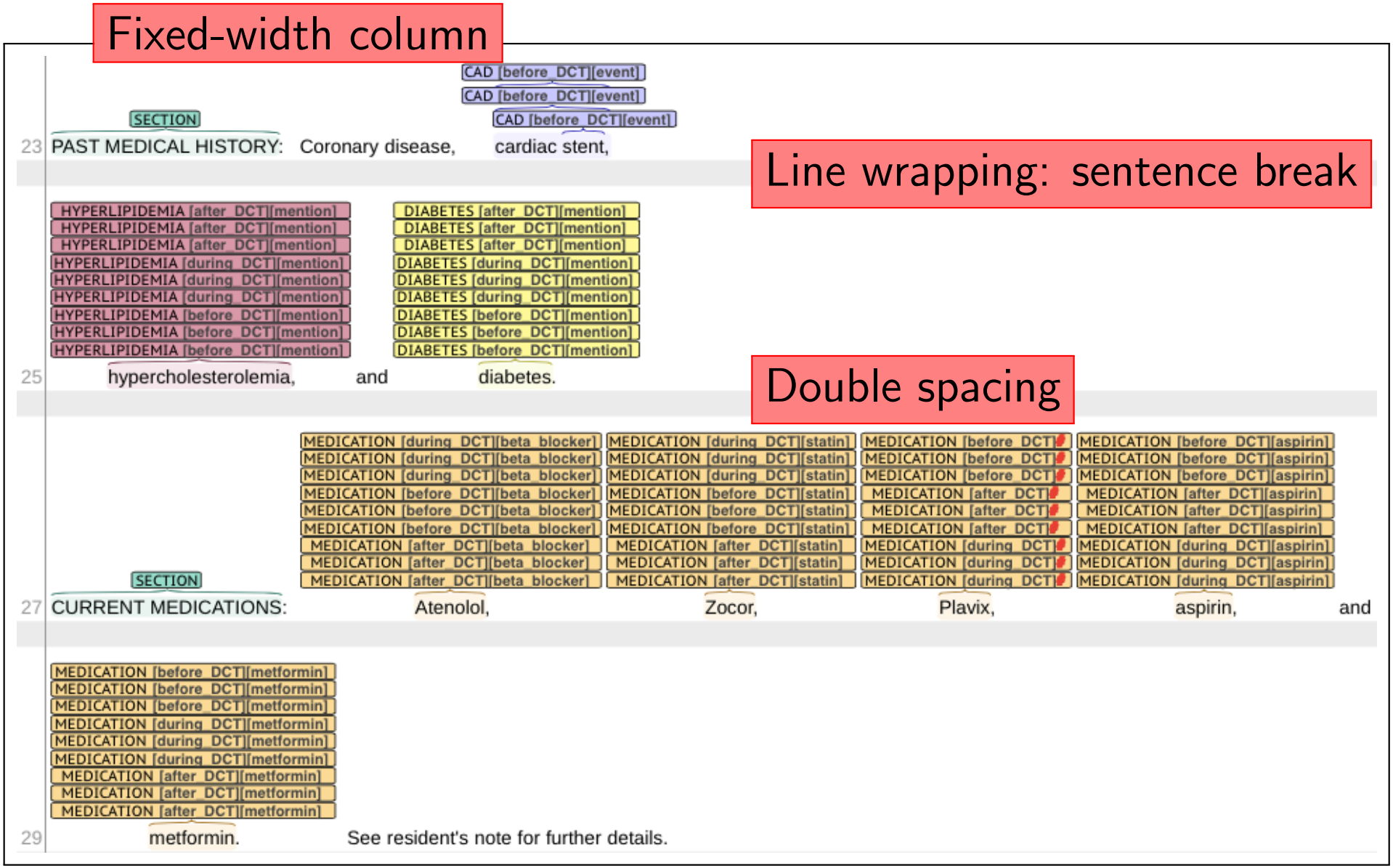
2 Medication side effects identification

3 Clinical records reformatting

4 Conclusion

Clinical records reformatting

Problem statement



Clinical records reformatting

Corpus analysis

- i2b2 corpora: four possible document structures:

Spacing	Double ✖	Single ✔	Single ✔	Single ✔
Fixed-width columns	Yes ✖	Yes ✖	No ✔	No ✔
Tabulations	Yes ✖	Yes ✖	Yes ✖	No ✔
Multi spaces	Yes ✖	Yes ✖	Yes ✖	No ✔
Word hyphenation	Yes ✖	Yes ✖	No ✔	No ✔

Clinical records reformatting

Method

1. **double spacing:** remove blank lines every other line
2. **indentation:** reduced, help processing of wrapped lines
3. **line wrapping:** text is not reformatted in its entirety
   → local constraints are checked to decide whether a line should be pasted to the following line
   - Section title → starts a new line
   - Numbered list → each item starts a new line
   - Tabulated list → each item starts a new line
   - Very short line → always ends current reformatted line
   - Moderately short line
     - ★ if ends w/ strong punctuation → ends current reformatted line
     - ★ if looks like a title → self-standing line
     - ★ otherwise → treated as normal line
   - Idiosyncrasies (initial identifiers, separation) → self-standing line
   - Otherwise: line is pasted to the preceding line

Clinical records reformatting

Corpus split into several parts

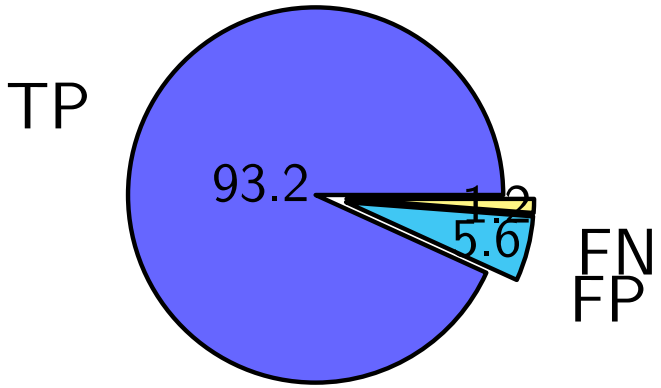
Part	#doc
Corpus	1,304
Training	790
Train (tune the system)	390
Dev (intrinsic evaluation)	131
Test	269
Testing (extrinsic evaluation)	514
Double spacing and line wrapping	125
Single spacing and line wrapping	98
No need for reformatting	291

Clinical records reformatting

Evaluation

Intrinsic evaluation: line wrapping

- 1,367 lines were correctly detected as being wrapped
- 82 were incorrectly detected as wrapped
- 17 wrapped lines were not detected



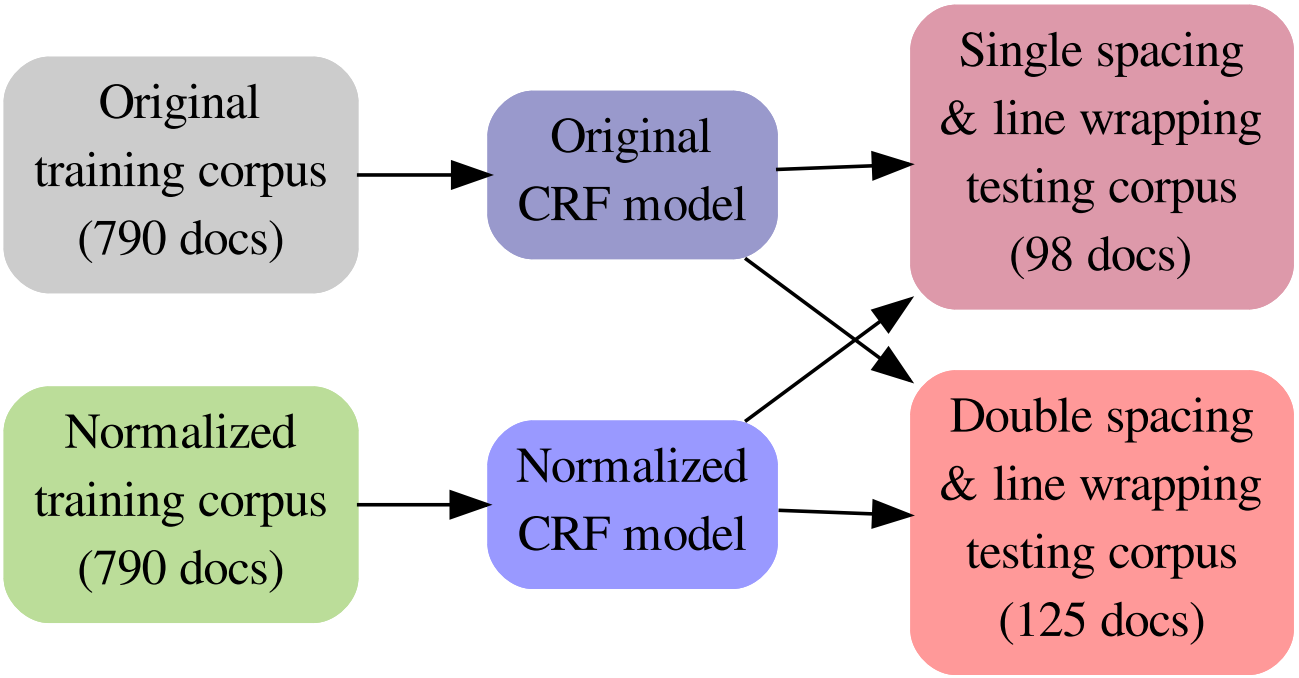
Sub-corpus	P	R	F
Training/dev	.9434	.9877	.9651

Clinical records reformatting

Evaluation

Extrinsic evaluation: application to i2b2/UTHealth challenge

- Design of experiments: 2 CRF models, 4 experiments



Clinical records reformatting

Extrinsic evaluation: Impact of reformatting on P, R, F

- Micro averaged measures on i2b2/UTHealth tasks depending on whether or not the texts have been reformatted

i2b2 task	Processing	Precision	Recall	F-measure
PHI (514 docs)	Original	.8937	.7332	.8055
	Normalized	.9015 ✔	.7314 ✖	.8076 ✔
Risk factor (514 docs)	Original	.9057	.7922	.8451
	Normalized	.9085 ✔	.7903 ✖	.8453 ✔

Clinical records reformatting

Extrinsic evaluation: Impact of reformatting on P, R, F

- Micro averaged measures on the risk factor identification task depending on:
  - the sub-corpus
  - and whether or not the texts have been reformatted

Sub-corpus	Processing	Precision	Recall	F-measure
Single (98 docs)	Original	.8761	.7755	.8227
	Normalized	.8779 ✔	.7705 ✖	.8207 ✖
Double (125 docs)	Original	.8887	.8174	.8516
	Normalized	.8984 ✔	.8222 ✔	.8586 ✔

1 Introduction

2 Medication side effects identification

3 Clinical records reformatting

4 Conclusion

Conclusion

- **medication side effects identification:**
  - best results using a CRF to detect both side effects and medication (F=0.63)
  - work in progress:
    - ★ are all annotations relevants?
    - ★ verification of annotations is needed
    - ★ application to French on messages from health forums
- **clinical records formatting:**
  - high results for wrapped lines detection (F=.9651)
  - but low improvement on both i2b2 tasks
    - ★ PHI: F=.8055→.8076
    - ★ Risk factor: F=.8451→.8453
  - Zweigenbaum, P. and Grouin, C. (2014). Reformatting clinical records based on global layout statistics. In *Proc of SMBM*, pages 53–60, Aveiro, Portugal.

Thank you!

Project Vigi4MED, grant ANSM-2013-S-060