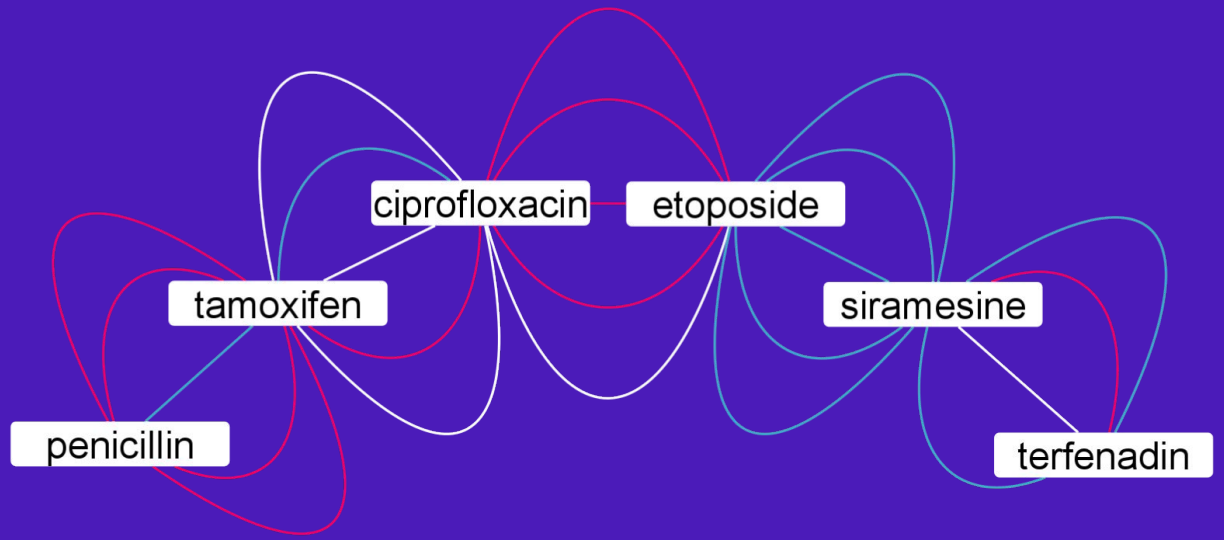
Relation Extraction

from medical literature
using SciBERT



Lykke Axlin & Klara Broman D17

Table of Contents

1 Aim

2 Model & Tools

3 Method

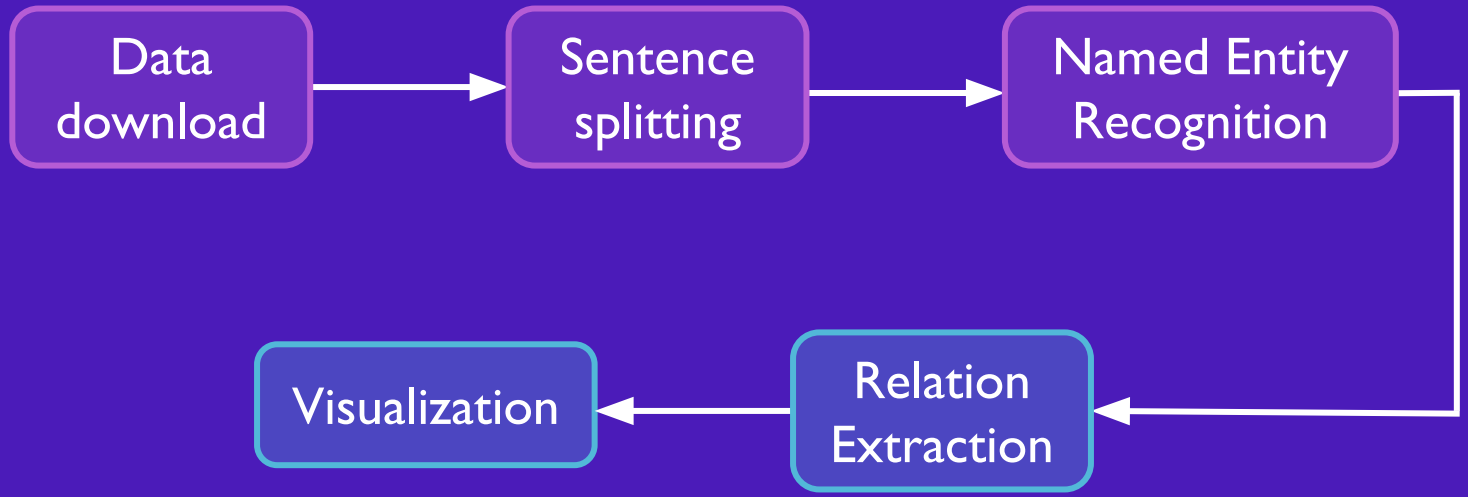
4 Result

5 Conclusion & Future Work

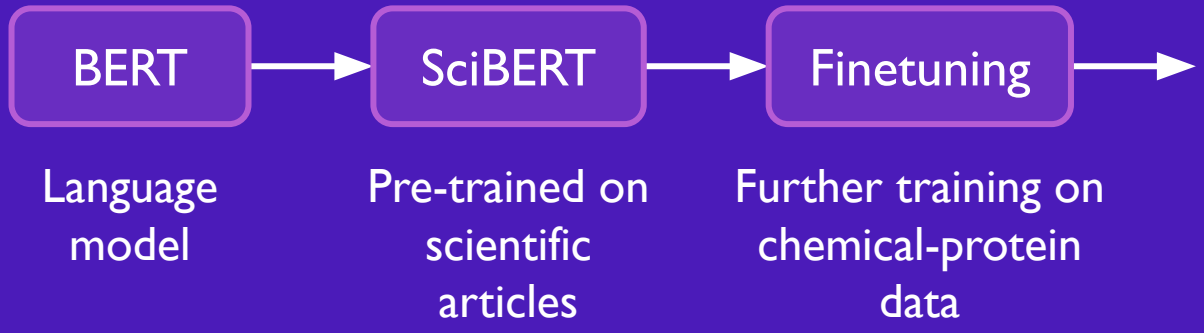
Aim & Overall goal

- Train models to extract relations in medical text
- Understand networks that control cellular functions and agents (drugs, pathogens, toxins) that can target them

Text mining pipeline



Model & Tools


BERT
SciBERT
Finetuning
Language model
Pre-trained on scientific articles
Further training on chemical-protein data

Chemprot corpus

Class	Train		Dev	
	Count	%	Count	%
INTERACTOR	2 583	40.13	1 350	37.96
NOT	241	3.74	175	4.92
PART-OF	308	4.79	153	4.30
REGULATOR-NEGATIVE	2 505	38.92	1 302	36.61
REGULATOR-POSITIVE	799	12.41	576	16.20
Total	6 436	100	3 556	100

Chemprot corpus

REGULATOR-NEGATIVE

"In conclusion, the renin inhibitor << Aliskiren >> dose-dependently decreases [[Ang II]] levels in humans following oral administration."

Method

Train → Evaluate:

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$\text{F1-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Result baseline model

Class	Train F1-score	Dev F1-score
INTERACTOR	0.95	0.66
NOT	0.70	0.25
PART-OF	0.93	0.47
REGULATOR-NEGATIVE	0.95	0.74
REGULATOR-POSITIVE	0.88	0.43

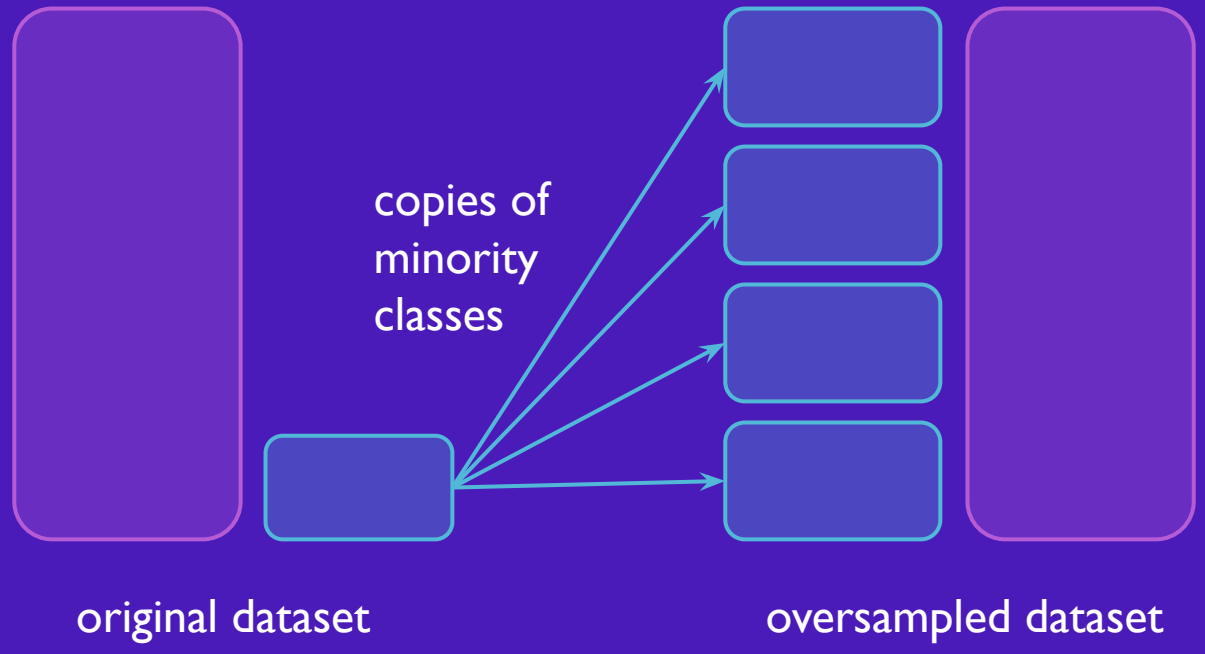
epoch 4

Result baseline model

Class	Train F1-score	Dev F1-score	Support (%) Train	Support (%) Dev
INTERACTOR	0.95	0.66	40.13	37.96
NOT	0.70	0.25	3.74	4.92
PART-OF	0.93	0.47	4.79	4.30
REGULATOR-NEGATIVE	0.95	0.74	38.92	36.61
REGULATOR-POSITIVE	0.88	0.43	12.41	16.20

epoch 4

Random oversampling



Chemprot oversampled corpus

Class	Train		Oversampled	
	Count	%	Count	%
INTERACTOR	2 583	40.13	2 583	26.87
NOT	241	3.74	1 205	12.53
PART-OF	308	4.79	924	9.61
REGULATOR-NEGATIVE	2 505	38.92	2 505	26.06
REGULATOR-POSITIVE	799	12.41	2 397	24.93
Total	6 436	100	9 614	100

Result oversampled model

Chemprot dev

Class	Original model epoch 4 F1-score	Oversampled model epoch 5 F1-score
INTERACTOR	0.66	0.70
NOT	0.25	0.36
PART-OF	0.47	0.58
REGULATOR-NEGATIVE	0.74	0.74
REGULATOR-POSITIVE	0.43	0.28

Creating artificial corpus

start phrase + entity 1 + interaction phrase +

entity 2 + end phrase

Artificial corpus

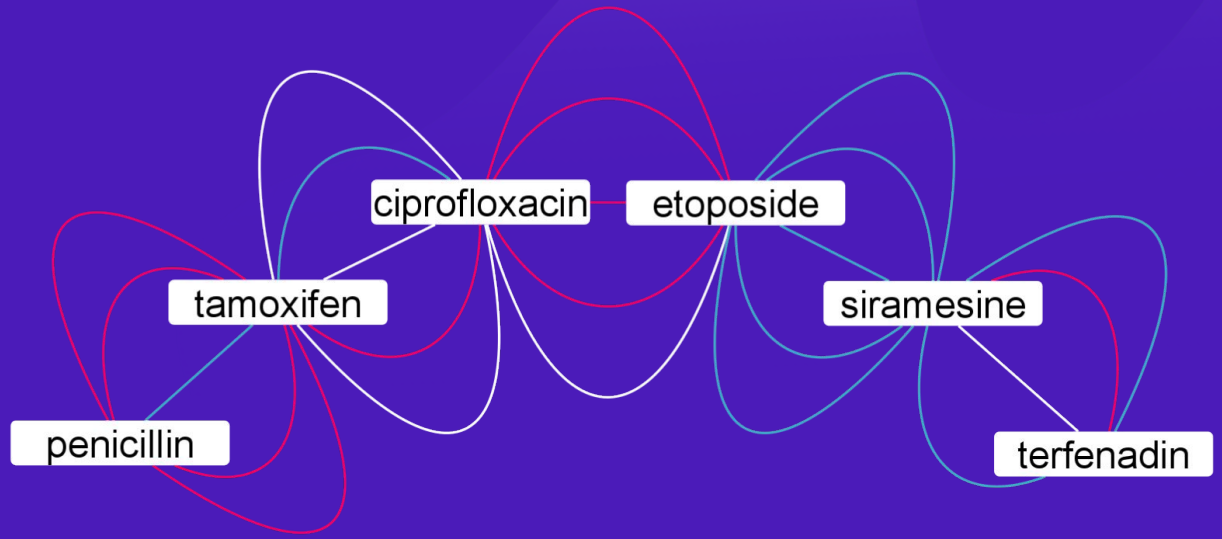
Class	Train		Dev	
	Count	%	Count	%
INTERACTOR	10 000	20	100	20
NOT	10 000	20	100	20
PART-OF	10 000	20	100	20
REGULATOR-NEGATIVE	10 000	20	100	20
REGULATOR-POSITIVE	10 000	20	100	20
Total	50 000	100	500	100

Result

F1-score

Model	Best epoch	Chemprot		Artificial	
		Train	Dev	Train	Dev
Original	4	0.88	0.51	-	-
Oversampled	5	0.97	0.65	-	-
Artificial	5	-	0.30	1.00	1.00
Artificial and original	5 + 1	-	0.38	0.57	0.56

Visualization



INTERACTOR

REGULATOR-POSITIVE

REGULATOR-NEGATIVE

Conclusion & Future work

- Best model: oversampling - 65% accuracy
- Not enough training data:
  - Create a more realistic artificial corpus
  - Combine Chemprot and artificial corpus

Questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**