The accuracy scores are token-level (not concept level).
Concepts predicted as N/As were considered as false predictions. In the newer vision, these predictions are processed a
present or absent.

Assertion Status Classification Using BiLSTM-CRF

With and without information on section types

Kexin Chen, Feb 27, 2024

Previous study

The bars in the graph show the distribution of assertion classes per section type. Concepts in these four sections are more likely to be "absent" compared to concepts in other sections: *Physical examination/Status*, *RoS*, *Allergies*, *Complications*. More than 50% concepts in these sections are absent.

(Note: Section headings are extracted using regular expressions.)

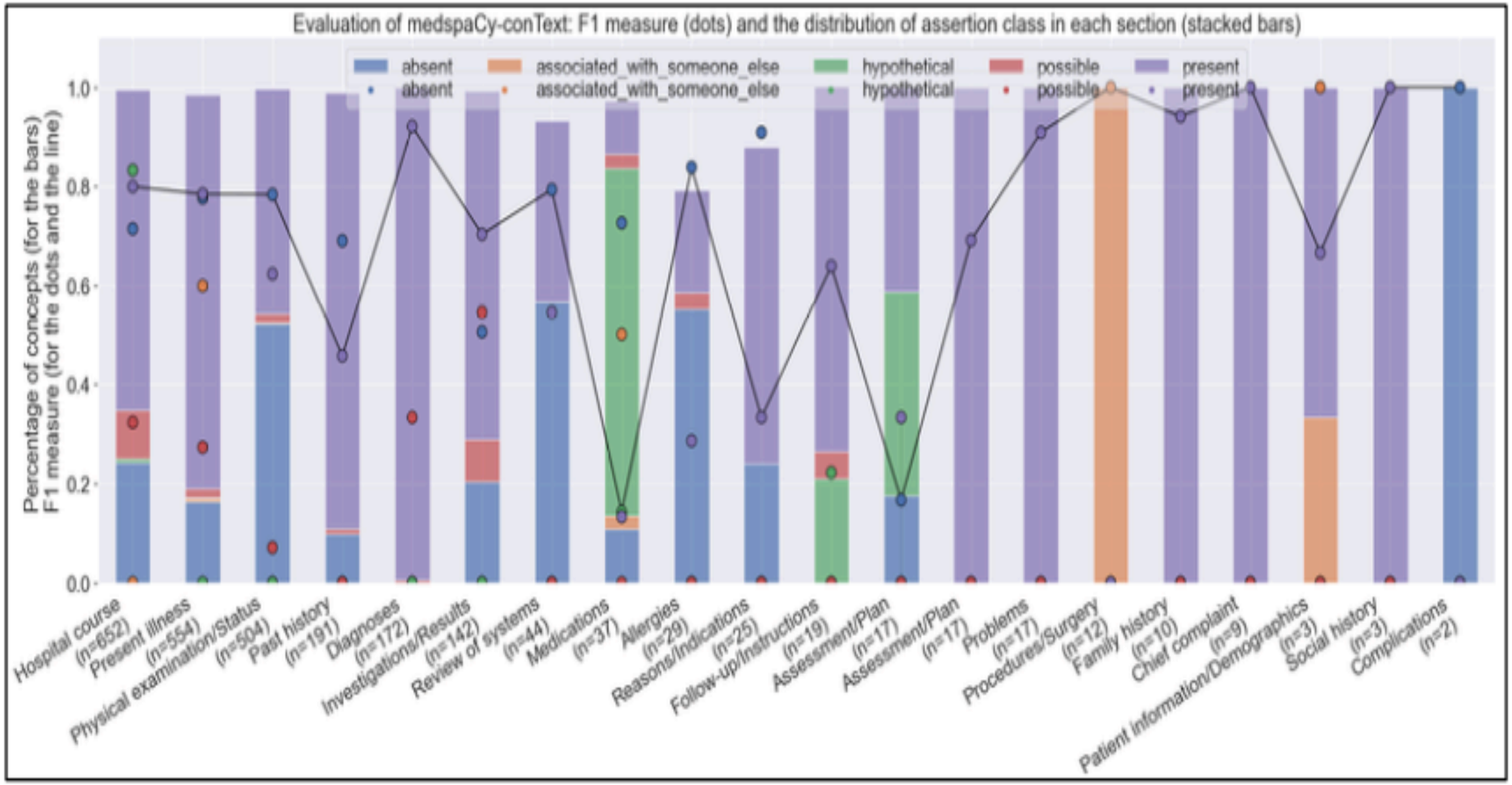


Figure 1. Stacked bars: Distribution of assertion classes per section. They do not add up to 1 because model's unability to determine "conditional" concepts. Dots: F1 measure of each assertion class. The black line connects the dots that represent the most frequently occurring assertion class in each section. This line serves as a reference point for comparison. Colors: Assertion classes.

2010 i2b2 Challenge

- Task: Assigning assertion types for medical problem concepts.
- Note type: Discharge summaries and progress notes
- A total of 826 clinical notes are split into 280, 69, and 477 for train, dev, and test, respectively.
- The dataset is annotated with medical problem concepts and their assertion status, i.e. present, absent, possible, conditional, hypothetical, and associated with someone else.

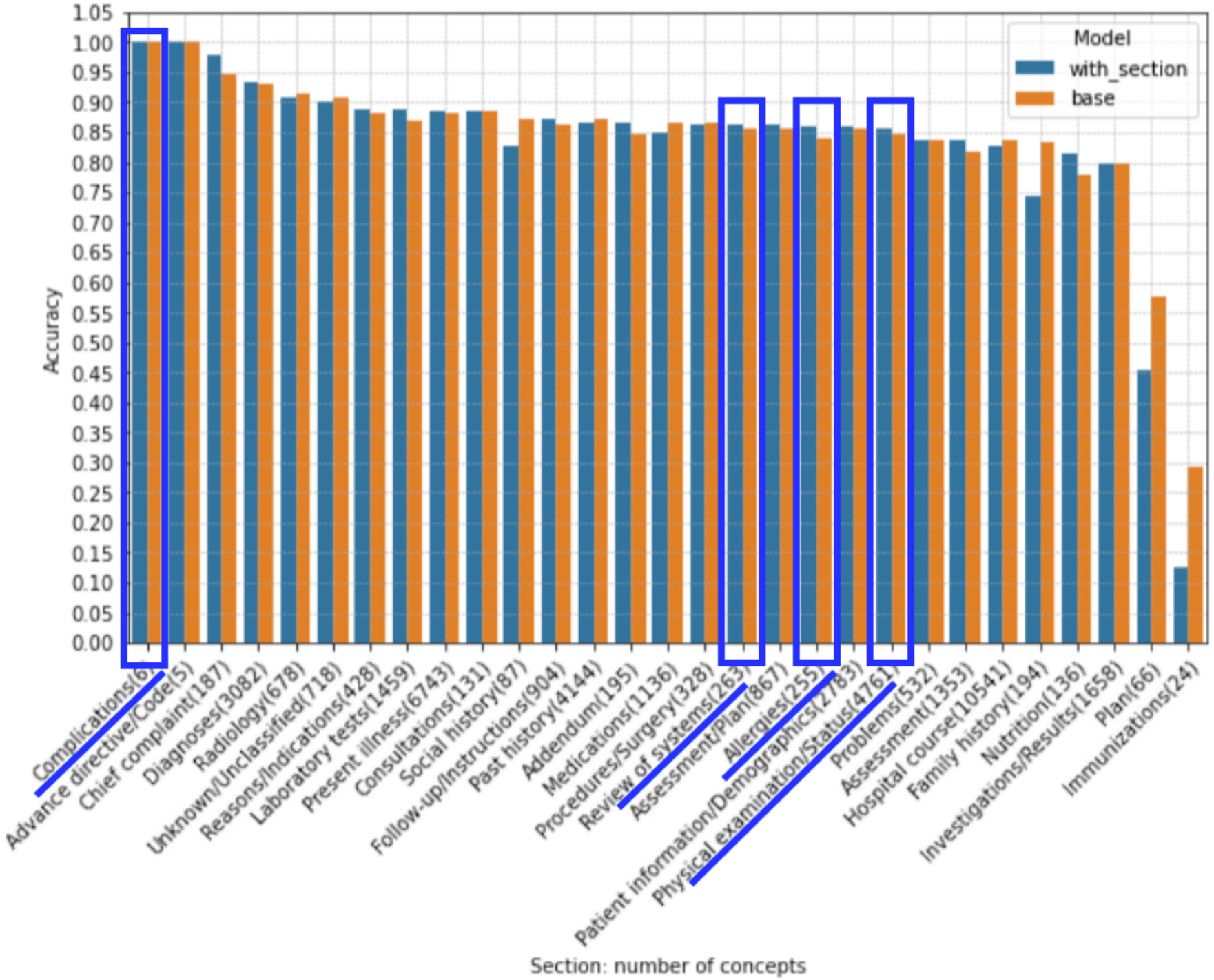
BiLSTM-CRF base model vs using section type as additional input

- Two models are trained:
  - BiLSTM-CRF (base)
  - BiLSTM-CRF using section type as additional input
- Incorporating section information does not seem to have an impact on the task.
- Therefore, I evaluate the models on just the sentences that belong to the highly (or lowly) negated sections. Using section information increases accuracy by 1.2% on concepts in the highly-negated sections.
- Incorporating sections seems to worsen model performance on concepts in the lowly-negated sections.

Table 1

	Evaluated on ...	Accuracy	Correct predictions	# of Concepts
BiLSTM-CRF Base	Full	0.8595	37529	43664
Use section type	Full	0.8599	37548	43664
BiLSTM-CRF Base	Concepts in the highly-negated sections	0.8458	4470	5285
Use section type	...highly-negated sections	**0.8570**	4529	5285
BiLSTM-CRF Base	Concepts in the lowly-negated sections	**0.8614**	33059	38379
Use section type	... lowly-negated sections	0.8603	33019	38379

- The graphs shows the two models' performance on concepts in different sections.
- It is not obvious whether using section information will increase accuracy for concepts in the lowly-negated (<40%) sections.
- Using section information slightly helps classifying the assertion status of concepts in the highly-negated sections.



Models trained on the full corpus vs trained on a subset (i.e. sentences in the highly/lowly-negated sections)

- Sentences are tagged as “highly-negated” if they belong to one of the four highly-negated sections. They form the “highly-negated” dataset. All other sentences form the “lowly-negated” dataset.
- BiLSTM-CRF models are trained on each subset with and without using section information as an additional input. Results are shown in Table 2.
- Only the base model trained on the lowly-negated subset has improved accuracy.
- Combining the outputs of the two models leads to lower accuracy than the models trained on the full set. A potential reason is that the disadvantages of training on a smaller dataset (especially, the highly-negated subset) outweighs the benefit of developing specific models for different sections.

Dataset	Description	Train	Dev	Test
Full	The original set	236,113 tokens	56,386 tokens	29,744 sents 441,950 tks
Highly-negated	Sentences that belong to the highly-negated sections	1,553 sents 21,919 tokens	286 sents 4,037 tokens	2,715 sents 36,316 tokens
Lowly-negated	Sentences that belong to all other sections	14,126 sents 214,192 tokens	3,518 sents 52,347 tokens	27,029 sents 405,629 tks

Table 2

Method	Trained on which dataset (or subset)	Accuracy (on the corresponding test set)	Correct predictions	# of Concepts
BiLSTM-CRF	Full	0.8595	37529	43664
BiLSTM-CRF + section type	Full	0.8599	37548	43664
BiLSTM-CRF	Concepts in the highly-negated sections	0.7941	4197	5285
BiLSTM-CRF + section type	...highly-negated sections	**0.8104**	4283	5285
BiLSTM-CRF	Concepts in the lowly-negated sections	**0.8677**	33300	38379
BiLSTM-CRF + section type	... lowly-negated sections	0.8478	32539	38379
Combine		0.8588	37494	43664
Combine (using section)		0.8433	36822	43664

Down-sample the datasets and train models on the smaller subsets for fair comparison - Dataset

- ~10% sentences are randomly selected to form the “downsample-full” set.
- ~10% sentences from the lowly-negated sentences are randomly selected to form the “downsample-lowly-negated” dataset.
- In this way, the size of training data is the same across all three datasets (i.e. highly-negated, downsample-full, downsample-low-negated)

Dataset	Description	Train	Dev	Test
Full	The original set	236,113 tokens	56,386 tokens	29,744 sents 441,950 tks
Highly-negated	Physical examination/ Status, RoS, Allergies, Complications	1,553 sents 21,919 tokens	286 sents 4,037 tokens	2,715 sents 36,316 tokens
Lowly-negated	Sentences that belong to other sections	14,126 sents 214,192 tokens	3,518 sents 52,347 tokens	27,029 sents 405,629 tks
Downsample-Full	~10% of all the sentences	1,553 sents 23,293 tokens	286 sents 4,439tokens	2,715 sents 38,899 tokens
Downsample-lowly-negated	~9% of the lowly-negated sentences	1,553 sents 23,198 tokens	286 sents 4,471 tokens	2,715 sents 38,824 tokens

Down-sample the datasets and train models on the smaller subsets for fair comparison - Results

- When training data size is small, adding section type as an additional input improves accuracy by >1%.
- Compared to models trained on the full size of data, models trained on the small subsets shows obviously worse performance. Apart from the impact of decreasing data size, another potential reason is that randomly selected sentences lack consistency in meaning. Sentences in the highly-negated sections remain relatively complete and consistent, and thus have better performance.

Table 2

Method	Trained on which dataset (or subset)	Accuracy (on the corresponding test set)	Correct predictions	# of Concepts
BiLSTM-CRF	Downsample full	0.5480	2210	4033
BiLSTM-CRF + section type	Downsample full	0.7441	3001	4033
BiLSTM-CRF	Highly-negated	0.7941	4197	5285
BiLSTM-CRF + section type	Highly-negated	0.8104	4283	5285
BiLSTM-CRF	Downsample lowly-negated	0.6820	2743	4022
BiLSTM-CRF + section type	Downsample lowly-negated	0.6974	2805	4022

Summary

- Using section information as an additional input when training models slightly helps classifying assertion status of the concepts in the *highly-negated* sections.
- Using section information as an additional input does not help classifying assertion status of the concepts in the *lowly-negated* sections.
- Training different models on different sections leads to worse performance, possibly because disadvantages of training on a smaller dataset (especially, the highly-negated subset) outweighs the benefit of developing specific models for different sections. My experiment design cannot lead to conclusions that when training data is sufficient, whether training different models on different sections improves accuracy or not.