FarsTail: A Persian Natural Language Inference Dataset

Fifth IPM Advanced School on Computing: Artificial Intelligence

Soroush Faridan

September 2021

Contents

Natural Language Inference

The goal of NLI is to determine the inference relationship between a premise (p) and a hypothesis (h).

Premise	Hypothesis	Label
It's raining	The ground is wet	Entailment
It's raining	The sky is sunny	Contradiction
It's raining	The wind is blowing	Neutral

Applications of NLI

1. Question answering
2. Semantic search
3. Automatic summarization
4. Evaluation of machine translation systems

FarsTail: A Persian Natural Language Inference Dataset

Why is this dataset needed?

1. The dramatic growth of Natural Language Processing
2. Little attention to the inference task in Persian language
3. Existence of special complications in Persian language

FarsTail is the first large textual dataset in Persian for the Natural Language Inference task.

It was prepared by a team of six for 22 months.

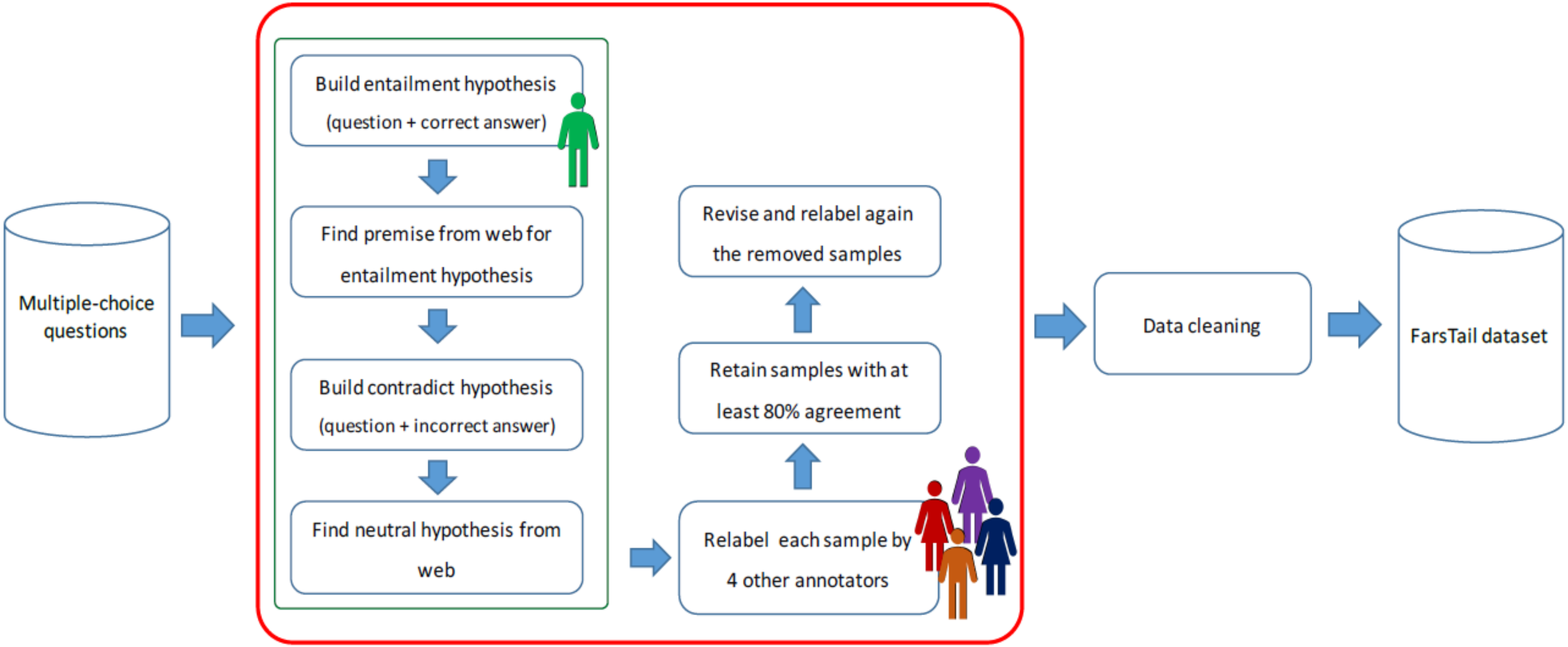
The data generation scenario is a bit like the SciTail [1] dataset.

[1] Khot, T., Sabharwal, A., & Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. In Thirty-Second AAAI Conference on Artificial Intelligence.

Textual datasets for the NLI task

Dataset Name	Year of publication	Number of data
SICK	2014	10,000
SNLI	2015	570,000
MultiNLI	2018	433,000
SciTail	2018	27,000
XNLI	2018	112,500

Development steps of FarsTail dataset



Example of generating instances

Multiple-choice question:

دبیر کل سازمان ملل متحد قبل از آنتونیو گوترش چه کسی بود؟

- خاویر سولانا
- بان کی مون (جواب صحیح)
- کوفی عنان
- یوشیرو موری

Who was the Secretary-General of the United Nations before António Guterres?

- Javier Solana
- Ban Ki-moon (correct answer)
- Kofi Annan
- Yoshirō Mori

Entailment hypothesis (question + correct answer):

دبیر کل سازمان ملل متحد قبل از آنتونیو گوترش، بان کی مون بود.

Ban Ki-moon was the Secretary-General of the United Nations before António Guterres.

Premise (from web):

مجمع عمومی سازمان ملل متحد رسماً آنتونیو گوترش را بعنوان دبیرکل بعدی سازمان ملل متحد و جانشین بان کی مون انتخاب کرد.

The United Nations General Assembly formally elected António Guterres as the next UN Secretary-General and Ban Ki-moon's successor.

Contradiction hypothesis (question + incorrect answer):

کوفی عنان پیش از آنتونیو گوترش بعنوان دبیر کل سازمان ملل متحد انتخاب شده بود.

Before António Guterres, Kofi Annan had been selected as the United Nations Secretary-General.

Neutral hypothesis (from web):

اعضای سازمان ملل متحد به اتفاق آرا آنتونیو گوترش را بعنوان نامزد دبیر کلی سازمان ملل متحد معرفی کردند.

The United Nations members unanimously nominated António Guterres as UN Secretary-General.

Statistics of the FarsTail dataset

subset	class	samples	prem. tokens	hyp. tokens	prem. proc. tokens	hyp. proc. tokens	overlap	proc. overlap
Train	E	2,429	40.50	15.53	19.35	8.42	0.67	0.68
	C	2,389	40.23	15.61	19.20	8.30	0.57	0.54
	N	2,448	40.52	15.62	19.31	8.26	0.40	0.30
	Total	7,266	40.42	15.59	19.29	8.33	0.55	0.51
Val	E	515	39.70	14.85	19.13	8.27	0.67	0.66
	C	499	39.58	15.09	19.17	8.11	0.58	0.54
	N	523	39.71	14.95	19.16	8.06	0.39	0.29
	Total	1,537	39.67	14.96	19.15	8.14	0.54	0.50
Test	E	519	39.57	15.48	18.84	8.39	0.68	0.68
	C	510	39.44	15.81	18.86	8.38	0.57	0.52
	N	535	39.23	16.02	18.73	8.36	0.38	0.27
	Total	1,564	39.41	15.78	18.81	8.38	0.54	0.49

Results

Model	Representation	Val Accuracy	Test Accuracy
SVM	tf-idf	0.5303	0.5301
	LASER	0.5459	0.5198
	word2vec	0.5120	0.5448
	fastText	0.5296	0.5371
	ELMo	0.5621	0.5710
LSTM	word2vec	0.5172	0.5243
	fastText	0.5205	0.5192
	ELMo	0.5478	0.5505
BiGRU	word2vec	0.5192	0.5224
	fastText	0.5211	0.5243
	ELMo	0.5582	0.5428
DecompAtt	word2vec	0.6597	0.6662
ESIM	fastText	0.7033	0.7116
HBMP	word2vec	0.6617	0.6604
ULMFiT	Learned	0.7281	0.7244
BERT	ParsBERT	0.8081	0.8299
	mBERT	**0.8263**	**0.8338**

Dataset bias by pointwise mutual information

$$\mathrm{PMI}(word, class) = \log \frac{p(word, class)}{p(word, .)p(., class)}.$$

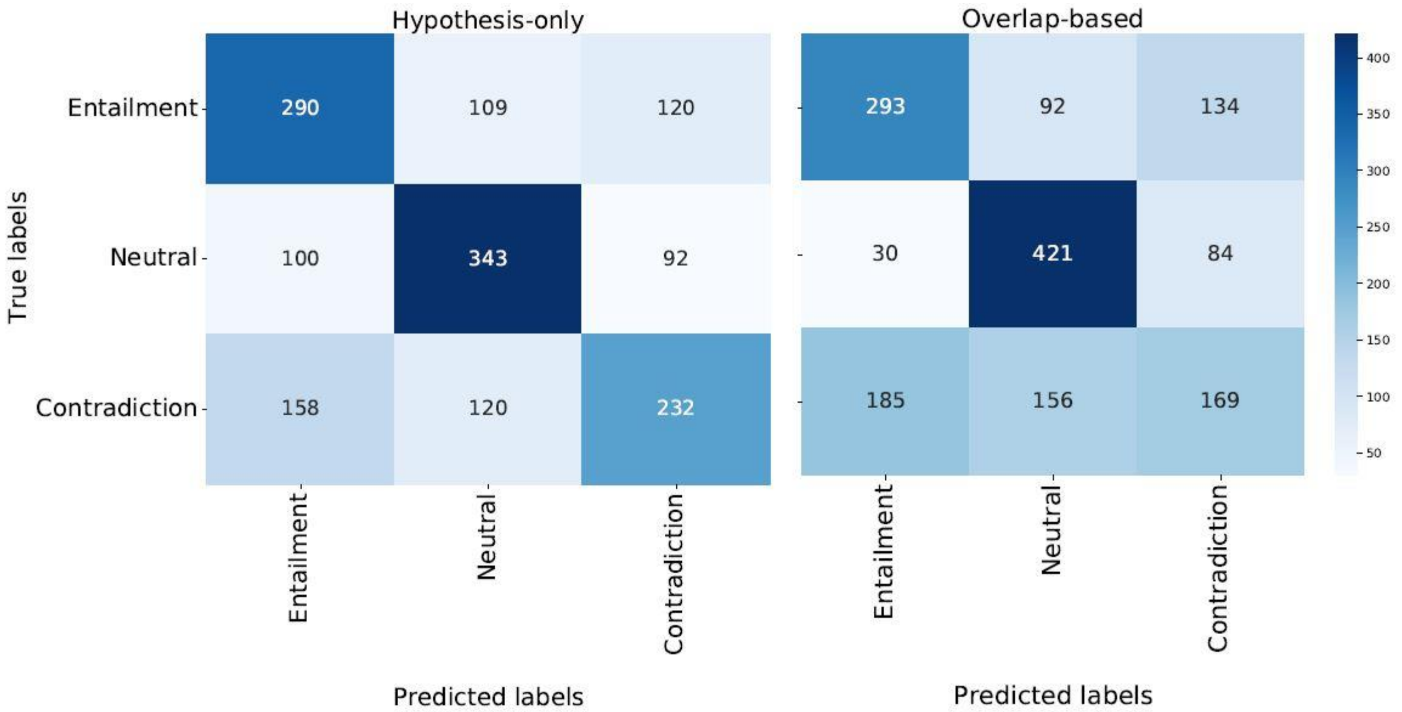
	Word	Class	PMI	Counts
MultiNLI	never	Contradiction	0.852	6599/8363
	no	Contradiction	0.820	12499/16515
	nothing	Contradiction	0.775	2090/2758
	any	Contradiction	0.735	5430/7739
	none	Contradiction	0.681	553/702
	anything	Contradiction	0.668	2239/3336
	completely	Contradiction	0.664	855/1190
	also	Neutral	0.644	1845/2726
	refused	Contradiction	0.644	401/498
	nobody	Contradiction	0.603	612/881
SciTail	to	Neutral	0.488	3541/5266
	have	Neutral	0.481	845/1155
	the	Neutral	0.479	14194/21758
	definite	Entailment	0.478	144/146
	because	Neutral	0.466	571/749
	system	Neutral	0.461	654/885
	.	Neutral	0.454	14790/23261
	a	Neutral	0.451	6086/9514
	off	Neutral	0.437	7644/12162
	and	Neutral	0.430	2771/4352
FarsTail	:	Neutral	0.244	95/158
	"	Entailment	0.227	466/1053
	"	Contradiction	0.222	463/1053
	تنها (only)	Contradiction	0.221	61/87
	باشد (be)	Contradiction	0.202	202/440
	نیز (also)	Neutral	0.179	50/76
	فقط (only)	Contradiction	0.168	38/50
	خود (self)	Neutral	0.163	143/319
	بعد (after)	Contradiction	0.162	74/144
	اثر (work,effect)	Entailment	0.159	70/135

Another approach for investigating dataset biases

1. Hypothesis-only model: Fine-tuning the mBERT model on the hypotheses to predict the entailment labels without seeing the premises. The model obtained an accuracy of 55.31% on the test set.

2. Overlap-based model: Using the cosine similarity between the bag-of-word count vectors of the premise and hypothesis as the input feature to investigate the ability of a model in deciding about the inference relationship just exploiting the similarity between the premise and hypothesis. An SVM classifier trained on this input feature obtained an accuracy of 56.46% on the test set.

Confusion matrices of the biased models



Partitioning the test set into two subsets

We partitioned the FarsTail test set into two subsets (for each biased model): easy and hard.

% of the total data	Easy/Hard for each biased model
32%	Easy for Hypothesis-only and Overlap-based biased models
20%	Hard for Hypothesis-only and Overlap-based biased models
25%	Hard for Hypothesis-only biased model
23%	Hard for Overlap-based biased model

Accuracy of different models on different subsets

Model	Full	Hypothesis-only Easy	Hypothesis-only Hard	Overlap-based Easy	Overlap-based Hard
DecompAtt (word2vec)	0.6662	0.7341	0.5823	0.7633	0.5404
HBMP (word2vec)	0.6604	0.7618	0.5350	0.7565	0.5360
ESIM (fastText)	0.7116	0.7931	0.6109	0.8120	0.5815
mBERT	0.8338	0.8763	0.7811	0.8981	0.7504

These results shows that the models' accuracy on the hard subset obtained by the overlap-based biased model is usually lower than that of the hypothesis-only biased model. This reveals that the models exploit more of the overlap information between premises and hypotheses than the biased patterns in the hypotheses.

Conclusion

- We introduced, to the best of our knowledge, the first relatively large-scale NLI dataset for Persian language.
- We presented the details of the FarsTail development process, which is carefully designed to ensure the data quality.
- We also presented the dataset statistics as well as the results of some traditional and state-of-the-art methods on it. The best obtained result on the FarsTail test set, using the powerful BERT method, is 83.38%.
- We also investigated the dataset biases in FarsTail.

Thanks!

Any questions?

soroush.faridan@gmail.com
faridan.github.io