FII SMART at SemEval 2023 Task7: Multi-evidence Natural Language Inference for Clinical Trial Data

Mihai B. Vołoșincu[1], Cosmin Lupu[1], Daniela Gifu[1,2], Diana Trandabat[1]
[1]Faculty of Computer Science, Alexandru Ioan Cuza University of Iași, România
[2]Institute of Computer Science, Romanian Academy - Iasi Branch
{bogdan.volosincu, cosminlupu, diana.trandabat)@gmail.com daniela.gifu@iit.academiaromana-is.ro

Abstract

The "Multi-evidence Natural Language Inference for Clinical Trial Data" task at SemEval 2023 competition focuses on extracting essential information on clinical trial data, by posing two subtasks on textual entailment and evidence retrieval. In the context of SemEval, we present a comparison between a method based on the BioBERT model and a CNN model. The task is based on a collection of breast cancer Clinical Trial Reports (CTRs), statements, explanations, and labels annotated by domain expert annotators. We achieved F1 scores of 0.69 for determining the inference relation (entailment *vs* contradiction) between CTR - statement pairs. The implementation of our system is made available via Github[1].

1 Introduction

Nowadays, investigation of clinical natural language processing tasks play a major role in the advancement of medical research, especially for innovation in evidence-based medicine (EBM) (Shivade et al., 2015). As reports of publications of Clinical Trial Reports (CTRs) continue to amass at rapid pace, it has become infeasible for clinical practitioners to remain constantly updated with the literature in order to provide personalized care (DeYoung et al., 2020).

Based on recent studies (Sutton et al., 2020), the legitimate research question is *how we successfully connect the latest evidence to support personalized care*? One direction is using textual entailment (Sammons et al., 2010; Cabrio and Manini, 2013), also known as Natural Language Inference (NLI) (Ravichander, A. et al., 2019, Silva et al., 2020), adapted for clinical texts. The relevant information from CTRs is structured around the patient's problem format (Eriksen and Frandsen 2018): patient, intervention, comparison and outcome (PICO), formulation which makes clinical questions accessible and really helpful for further automated processing cause from human perspective the huge amount of data from randomized trials makes analysis endless and unavailing. Often, clinical results can be contradictory, with indeterminate findings due to variations in parameters (medication, age, location, duration, etc.) and creates an immense number of possible combinations between inclusion criteria, interventions, effects and outcomes, context in which the inference of evidence (Galashov, A. et al., 2019) was seen as a risky or too challenging area to tackle.

This perspective has known a great deal of progress in recent years due to developments of transformer models with fine-tuned variants like BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) and SciBERT (a BERT model trained on scientific text) trained on medical corpora that facilitate rapid prototyping and experiments (Lee et al. 2019).

The rest of the paper is organized as follows: section 2 briefly presents studies related to textual entailment in clinical text, section 3 provides information about the system designed to determine the inference relation (entailment *vs.* contradiction) between CTR - statement pairs, while section 4 describes the experimental setups. Section 5 resumes the results of the conducted experiments, with their interpretations, followed by section 6 with the conclusions and further directions.

2 Background

This topic has attracted large attention in recent years, evidenced by increasing number of

[1]https://github.com/volosincu/FII_Smart__Semeval2023

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 212–220

scientific events (e.g., Workshop on Curative Power of MEdical DAta (MEDA 2017; 2018; 2020) (Cohen, K.B et al., 2020., Gifu, D. et al., 2019), Workshop on on Clinical Natural Language Processing (ClinicalNLP 2016; 2019; 2020; 2022) (Naumann, T et al., 2022; Rumshisky, A et al., 2020; Rumshisky, A et al., 2019; Rumshisky, A et al., 2016) and competitions.

Competitions such as SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data are challenging, especially due to the data labelling problem, which is directly dependent on the automatic determination of the inference relation (entailment *vs.* contradiction) between CTR - statement pairs.

Different strategies are used in studies about clinical trial reports, from classical NLP techniques to transformer-based methods. The objective of this paper is to test if a transformer pretrained model on bio medical data can offer a competitive edge over the more classical approaches of NLP processing.

A consequent hypothesis, derived from initial analysis of the dataset, was that employing a traditional solution consisting of a CNN (convolutional neural network) model would lead to lesser or inconsistent results.

The dataset provided by organizers (Mael et al. 2023) for this competition consists of 999 breast cancer CTRs in English[2], with statements, explanations and labels annotated by domain expert annotators. Each CTR[3] may contain 1-2 patient groups, called cohorts or arms. These groups may receive different treatments or have different baseline characteristics.

For task 1, each training instance contains 1-2 CTRs, a statement, a section marker, and an entailment/contradiction label.

The annotations collected from reports of Randomized Control Trials (RCTs) represent the raw material and the input consumed by the system.

The task entries provided by organizers for dev and test datasets are following the ICO intervention, comparison and outcomes format (DeYoung et al., 2020) and may refer to the same study or to two different studies.. For ease of reference, a demonstrative and trivial example would be to infer entailment between sentences chosen from a scenario like *given a treatment A, a comparator B, and an outcome*, the inference of entailment is to be established based on the statement and validity of the premises.

There is an argument sustained by one or multiple premises with a "total of 2,400 statements split evenly across the different sections and classes" (Mael et al. 2023).

The NLI4CT instances from where the prompts have been extracted is a collection of clinical trial reports (CTRs) formatted following the PICO standards consisting of:

- Clinical Trial Id
- Eligibility criteria
- Intervention
- Results
- Adverse effects.

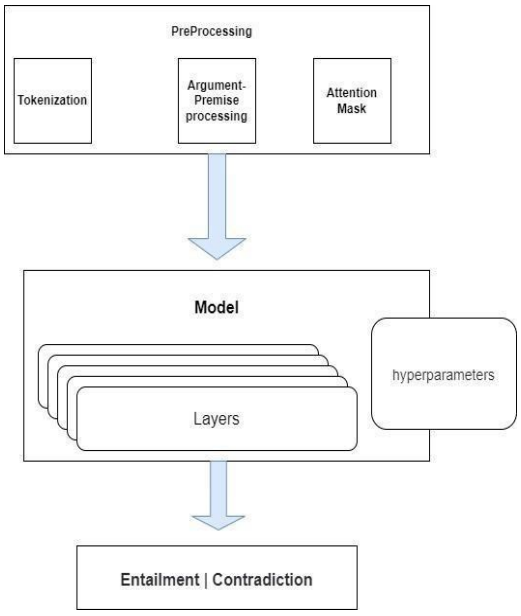


Figure 1: The FII-SMART system architecture

Each clinical trial contains one or two patient groups (cohorts). The records and ICOs for development and testing are provided in JSON format (see Appendix D1). In the implementation, there are modules that handle data formatting before and after the model training and predictions.

[2] https://clinicaltrials.gov/ct2/home

[3] Appendix D.1 A full display example of an NLI4CT document with annotation from a clinical trial record

3 System overview

The strategy we considered for this task represents a causal approach (see Figure 1) inspired from the architecture used in (Lee et al. 2019). The solution consists of a neural network model with multiple layers, the first of which is a pre-trained BioBert transformer extended with extra dense layers, detailed in Tables 1 and 2.

The given argument-premise pair forms a sparse input space, with medical terms, quantities and relations between them having a very subjective nature. The pretrain model is the first layer in the overall model architecture, having the role to reduce the dimensionality of the input argument-premises pairs[4].

Semantic representation is a well-known method in the industry (Lee et al. 2019), but with the transformer model incorporated into the solution, we needed to leverage the capabilities and fine tune the mask and type ids settings. However, the results did not show improvements or relevant differences between the variants of the model where adaptations have been applied on the input mask of the transformer.

Attention mask is normally used to differentiate between the information containing tokens and the padding ones to predict the proper words based on local context. *What if the attention would have been focused only on tokens with relevant information*? The process to determine the relevant tokens presents itself as a challenging quest where discrete and extrinsic context come into play. This makes it a crucial step where more advanced work has to be done in future but for current implementation two examples are stated below. The first is a favorable one, because there is a similarity between the two pairs and sentences are smaller and have a local context well defined *ECOG score smaller than 2*.

3.1 Example

Study: NCT01125566
Argument, ECOG score < 2 is necessary to be eligible for the primary trial"
Premise (Eligibility), „Eastern Cooperative Oncology Group (ECOG) score of 0 or 1"

3.2 Example

Study: NCT00328783
Argument „The adverse events section in the primary trial is empty"
Premise (Adverse Events) „Total: 0/0"

In the case of the second example, it is challenging to determine if *Total: 0/0* is strictly referring to all adverse effects or to some other aspects. Since the adverse section should be empty, then if it is not empty, it could mean that *Total: 0/0* is pointing to other type of measurement.
There are numerous cases that would require extra-processing and analysis before feeding a neural network and that can be added gradually as modules into the system. For a couple more examples that illustrate these kinds of situations see Appendix F1.

Layer	Nodes	Activation	Alpha
1	BioBert	N/A	N/A
2	256	LeakyRelu	0.9
3	1024	LeakyRelu	0.9
4	2048	LeakyRelu	0.9
5	256	LeakyRelu	0.9
6	128	LeakyRelu	0.7
7	2	softmax	N/A

Table 1: Solution 1. Layer configuration (nodes/layer)

After input processing is complete, the semantic representation obtained from the BioBert encoder is sent further to extra dense layers to learn more data patterns.
For the second solution we choose a standard CNN[5]. Although not a typical choice in text processing, since in the industry the convolutional networks are mostly known and employed in image processing there are instances and research where CNN are proven effective in causal inference problems (Ghasempour et al. 2023) reason why to be considered an adequate solution for textual inference task. For the preprocessing step, the same modules have been used, including

[4] The algorithm for processing the argument-premises pairs is presented in detail in as pseudocode in Appendix E.1

[5] CNN notebook

the input structure of the transformer model, but in a different configuration of layers.

Layer	Nodes	Activation	Alpha
1	512	LeakyRelu	0.9
2	2048	LeakyRelu	0.9
3	4096	LeakyRelu	0.9
4	18048	LeakyRelu	0.9
5	9128	LeakyRelu	0.9
6	1024	LeakyRelu	0.9
7	64	LeakyRelu	0.9
8	2	softmax	N/A

Table 2: Solution 2. Layer configuration (nodes/layer)

4 Experimental setups

Both solutions have been developed and trained in Google collab environment[6][7]. For the first experiment, using the transformer, in the medical domain we choose the BioBert pretrained model as best fit for the task and the dataset. The transformer model input tensor can accept up to 500 tokens, but this raises a resources challenge at memory allocation. In order to be able to train the system in reasonable time (minutes), the input has been pre-processed.

The solution was adapted so that the input is split into multiple instances, each with its own statement and premise pair, until we reached the parameter set for maximum sequence size for BioBert input. As extra processing, in case the size of premises was greater than max size, the premises were split in two or more entries. This helped us to mitigate the resource problem, leaving some questions open, especially regarding the possibility of improving accuracy by optimizing for larger inputs for models.

Some pairs argument - premises have lost context and the contradiction or entailment label did not match on all tested instances, but at the same time, having a more concise context might have helped in other regards. This is one area that needs further attention, especially in deciding if and how the data can be better organized.

For hyperparameters, the experiments have been run with multiple values from different ranges, empirically obtaining best results with size 16 batch and 15 epochs.

Batch Size	**16**	**32**	**64**	**128**
Epochs	**15**	**30**	**50**	**100**

Table 3: Batch sizes and epochs used in training models

The practical sequence length for the BioBert input was between 100 and 290 tokens, values at which we could train the model within the limits of RAM maximum allocation (89Gb). Beyond that length the machine would go out of GPU memory and the model would crash.

Name	**Value**
loss	sparse_categorical_cross entropy
optimizer	Adam
learning rate	2e-5, 5e-5
activation hidden	leaky relu
activation output	softmax
metric	accuracy

Table 4: Hyperparameters used in training and values selected from runs with top results

The token sequence is built dynamically at runtime depending on the maximum size we set (ex. 290, 100). For each statement we search the premises in the CTR file and then merge arguments and premises (using [CLS] + [SEP] BERT).

In the validation phase, after prediction, we iterate the results and search the entries that have been split and choose the final prediction based on majority voting.

5 Results

For the purposes of benchmarking and familiarizing participants with the task's

[6] Transformer notebook

[7] CNN notebook

challenges, the organizers provided a starting kit that uses a simple TF-IDF algorithm.
This baseline solution has achieved results of 48% F1 score, compared to our solution with the best result 69.9% F1 score for development data.

F1 score	precision	recall
0.502415	0.485981	0.520000

Table 5: Baseline score obtained with Tf*idfVectorizer and cosine distance

The complete history and settings (epochs, learning rate) are available in the notebook version history. It can be observed that training with chunked data with inputs size around 100 tokens issued predictions between 65-70%[8]. A plausible explanation could be that smaller sentences keep more concentrated semantic context, but this must be looked upon on a more detailed level since only some arguments could be validated by claims on semantic grounds.

Input size	120 tokens					200 tokens		
	CNN	BioBert	CNN	BioBert	CNN	BioBert	CNN	BioBert
Batch size	16	16	128	128	16	16	128	128
Mean F1	0.6043	0.6269	0.5143	0.6371	0.4757	0.6234	0.5889	0.6077
Best F1	0.6739	0.6690	0.6716	0.6811	0.6622	0.6753	0.6666	0.6666

Table 6: FII-SMART overall results.

Table 6 shows the results side by side for the 2 models with results. The choice of batch size and input size as references in table is based on the observation that these two parameters are most relevant for comparison and impact on results.
The complete list of results shown in Appendix A.1 and B.1 contains the values after each retrain process for dev labeled dataset.
One big and surprising advantage of the CNN model is that, in terms of training time and resources, it was much faster and required less computing power (RAM, CPU, GPU).

6 Conclusions

We demonstrated in this paper that a transformer pretrained model on bio medical for entailment relation task, does not necessarily give a competitive edge over a more classical approach, such as CNN. Compared with the baseline system, both solutions implemented by our team scored better results and give valuable insights into further investigations regarding how the architecture and model can evolve.
Another area that can bring a lot of gains in the overall score is to extract essential information or relevant expressions from the statement and premises in order to match the maximum BioBert input vector and get more accurate semantic representation.
It is possible that the inference task be more influenced by factors that are not necessarily dependent on medical context. We found that this is a challenging task (F-score = 0.69), with many open promising directions for further research.

References

Shivade, C., Hebert, C., Lopetegui, M., Marneffe, M.-C., Fosler-Lussier, E., Lai, A.. 2015. Textual Inference for Eligibility Criteria Resolution in Clinical Trials. *Journal of biomedical informatics.*58S. 10.1016/j.jbi.2015.09.008.

DeYoung, J., Lehman, E., Nye, B., Marshall, J. J., Wallace, B. C. 2020. Evidence Inference 2.0: More Data, Better Models. *arXiv:2005.04177*. https://doi.org/10.48550/arXiv.2005.04177

Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D., C., Fedorak, R. N., Kroeker, K. I. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 3(1):1–10.

[8] Detailed results and comparison tables in Appendix A.1 and Appendix B.1

Sammons, M., Vydiswaran, V. G. V., Roth, D. 2010. Ask Not What Textual Entailment Can Do for You... In ACL 2010, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, July 11-16, 2010, Uppsala, Sweden, pages 1199–1208.

Cabrio, E., Magnini, B. 2013. Decomposing Semantic Inferences. Linguistics Issues in Language Technology - LiLT. Special Issues on the Semantics of Entailment, 9(1), August.

Ravichander, A., Naik, A., Carolyn Penstein Rosé, and Hovy, E.H. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. *CoRR*, abs/1901.03735.

Silva, V. S., Freitas, A., Handschuh, S. 2020. XTE: Explainable Text Entailment. *arXiv:2009.12431v1* [cs.CL].

Eriksen, M. B., Frandsen., T., F. 2018. The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *Journal of the Medical Library Association* 106 (4): 420-430. https://doi.org/10.5195/jmla.2018.345

Galashov, A., Schwarz, J., Hyunjik K., Garnelo, M., Saxton, D., Kohli, P., Eslami, S. M., A., The, Y. W. 2019. Meta-learning surrogate models for sequential decision making. CoRR, abs/1903.11907.

Lee, J., Yoon, W., Sungdong K., Donghyeon K., Sunkyu K., Chan H. S., Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *CoRR* abs/1901.08746.

Cohen, K.B., Gîfu, D., Li, Y., Ripple, A. and Xia, J., 2020, August. MEDA 2020: The curative power of medical data. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (pp. 575-576).

Gîfu, D., Trandabăț, D., Cohen, K. and Xia, J., 2019. Special Issue on the Curative Power of Medical Data. *Data*, *4*(2), p.85.

Naumann, T., Bethard, S., Roberts, K., Rumshisky, A. (Eds.) 2022. Proceedings of the 4th Clinical Natural Language Processing Workshop. Association for Computational Linguistics, Seatle, US.

Rumshisky, A., Roberts, K., Bethard, S., Naumann, T. (Eds.) 2020. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Association for Computational Linguistics.

Rumshisky, A., Roberts, K., Bethard, S., Naumann, T. (Eds.) 2019. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Association for Computational Linguistics, Minneapolis, US.

Rumshisky, A., Roberts, K., Bethard, S., Naumann, T. (Eds.) 2016. Proceedings of the Clinical Natural Language Processing Workshop. Association for Computational Linguistics, Osaka, Japan.

Mael, J., Valentino, M., Frost, H., O'Regan, P., Landers, D., Freitas, A. 2023. SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Ghasempour, M., Moosavi, N., & de Luna, X. 2023. Convolutional neural networks for valid and efficient causal inference. arXiv preprint arXiv:2301.11732.

Appendices

A.1 Results for input size 200

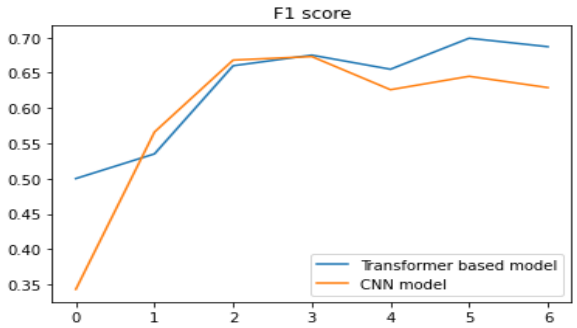
Batch	16						128					
Model	CNN			Biobert			CNN			Biobert		
Score	F1	precison	recall	F1	prec.	recall	F1	precision	recall	F1	prec.	recall
	0.5288	0.5092	0.55	0.5578	0.5888	0.53	0.6428	0.5328	0.81	0.6210	0.5714	0.68
	0.3932	0.4487	0.35	0.6238	0.5762	0.68	0.421	0.4444	0.40	0.5671	0.5643	0.57
	0.3034	0.4888	0.22	0.6407	0.6226	0.66	0.5081	0.5529	0.47	0.6224	0.5319	0.75
	0.4597	0.5405	0.4	0.6528	0.5563	0.79	0.6494	0.5146	0.88	0.6614	0.5414	0.85
	0.6539	0.5276	0.86	0.6046	0.7222	0.52	0.6431	0.4972	0.91	0.6518	0.5176	0.88
	0.2222	0.5384	0.14	0.6582	0.5693	0.78	0.6212	0.5000	0.82	0.6260	0.5538	0.72
	0.4949	0.5000	0.49	0.6226	0.5892	0.66	0.6222	0.4941	0.84	0.6638	0.5724	0.79
	0.3066	0.4600	0.23	0.6428	0.5806	0.72	0.6666	0.5000	1.00	0.4895	0.5108	0.47
	0.5726	0.5000	0.67	0.6549	0.5054	0.93	0.5258	0.4955	0.56	0.5636	0.5166	0.62
	0.5052	0.5333	0.48	0.6753	0.5954	0.78	0.6622	0.4974	0.99	0.6666	0.5820	0.78
	0.5511	0.4960	0.62	0.6108	0.6019	0.62	0.6166	0.5098	0.78	0.6374	0.5298	0.80
	0.3680	0.4761	0.3	0.5913	0.6395	0.55	0.5925	0.5034	0.72	0.6311	0.5347	0.77
	0.5937	0.4871	0.76	0.6637	0.5833	0.77	0.5106	0.5454	0.48	0.6470	0.5579	0.77
	0.6524	0.5054	0.92	0.6238	0.5762	0.68	0.5754	0.5446	0.61	0.5434	0.5952	0.50
	0.3172	0.5111	0.23	0.6297	0.5481	0.74	0.5485	0.4744	0.65	0.6523	0.5714	0.76
	0.6622	0.4974	0.99	0.6306	0.5737	0.7	0.5952	0.4934	0.75	0.6285	0.6000	0.66
	0.4456	0.4880	0.41	0.6718	0.5512	0.86	0.5952	0.4934	0.75	0.4402	0.5932	0.35
	0.2769	0.6000	0.18	0.4022	0.4729	0.35	0.5967	0.5000	0.74	0.6521	0.5769	0.75
	0.5560	0.5428	0.57	0.6428	0.5328	0.81	0.6503	0.5000	0.93	0.5418	0.5339	0.55
	0.6503	0.5000	0.93	0.6690	0.5257	0.92	0.5345	0.4957	0.58	0.6478	0.6106	0.69
Mean	0.4757			0.6234			0.5889			0.6077		
Best	0.6622			0.6753			0.6666			0.6666		

B.1 Results for input size 120

Batch	16						128					
Model	CNN			Biobert			CNN			Biobert		
Score	F1	precision	recall	F1	prec.	recall	F1	precision	recall	F1	prec.	recall
	0.6575	0.5000	0.96	0.6694	0.5755	0.80	0.5872	0.5111	0.69	0.6516	0.5209	0.87
	0.5000	0.5340	0.47	0.6118	0.5630	0.67	0.6716	0.5393	0.89	0.6028	0.5779	0.63
	0.5235	0.5494	0.50	0.4487	0.6250	0.35	0.4900	0.4900	0.49	0.6288	0.6489	0.61
	0.6332	0.5157	0.82	0.6839	0.6030	0.79	0.4939	0.6212	0.41	0.6443	0.5539	0.77
	0.6541	0.5240	0.87	0.6692	0.5398	0.88	0.5365	0.5238	0.55	0.6806	0.5869	0.81
	0.6254	0.5094	0.81	0.5483	0.5930	0.51	0.3194	0.5227	0.23	0.6355	0.5514	0.75
	0.6500	0.5055	0.91	0.5000	0.5972	0.43	0.5764	0.5116	0.66	0.6167	0.5511	0.70
	0.6071	0.5483	0.68	0.6812	0.6046	0.78	0.4494	0.5128	0.40	0.6533	0.5430	0.82
	0.5268	0.5142	0.54	0.6635	0.6228	0.71	0.4795	0.4895	0.47	0.6811	0.5340	0.94
	0.6048	0.5067	0.75	0.6139	0.5739	0.66	0.6070	0.4968	0.78	0.5656	0.5714	0.56
	0.6666	0.5025	0.99	0.6132	0.5803	0.65	0.6384	0.5187	0.83	0.5727	0.5398	0.61
	0.4949	0.5000	0.49	0.6820	0.6324	0.74	0.6120	0.5378	0.71	0.6533	0.5430	0.82
	0.5026	0.5274	0.48	0.6315	0.5306	0.78	0.3926	0.5079	0.32	0.6118	0.5630	0.67
	0.6453	0.5000	0.91	0.6666	0.5766	0.79	0.3142	0.5500	0.22	0.6416	0.5500	0.77
	0.6739	0.5317	0.92	0.6484	0.5966	0.71	0.5096	0.4907	0.53	0.6504	0.6320	0.67
	0.6716	0.5357	0.90	0.5786	0.5876	0.57	0.5253	0.4871	0.57	0.6666	0.6283	0.71

	0.6379	0.4972	0.89	0.5951	0.5809	0.61	0.4258	0.6000	0.33	0.644	0.5588	0.76
	0.5000	0.5340	0.47	0.6690	0.5257	0.92	0.6711	0.5050	1.00	0.6168	0.5789	0.66
	0.6595	0.5109	0.93	0.6532	0.5472	0.81	0.3797	0.5172	0.30	0.6622	0.4974	0.99
	0.6515	0.5243	0.86	0.6798	0.5620	0.86	0.6055	0.5033	0.76	0.6614	0.5414	0.85
Mean	0.6043			0.6269			0.5143			0.6371		
Best	0.6739			0.6690			0.6716			0.6811		

C.1 F1 score graphic



The training and tunnig results side by side of the 2 solutions

D.1 CTR Example

Below is a CTR file content from the dataset.

```
{
  "Clinical Trial ID": "NCT00181363",
  "Intervention": [
     "INTERVENTION 1: ",
     "Prone",
     "Prone position",
     "INTERVENTION 2: ",
     "Supine",
     "Supine position"
  ],
  "Eligibility": [
     "Inclusion Criteria:",
     "Patients should have had breast-conserving surgery for breast cancer or DCIS (Ductal Carcinoma in Situ)",
     (…)
  ],
  "Results": [
     "Outcome Measurement: ",
     "Dose Homogeneity 1: PTV",
     "Quantitatively compare the 3 D dose distribution in the PTV (Planning Target Volume) and normal tissues in prone position versus supine position",
     "Time frame: 1 day after treatment planning",
     (…)
  ],
  "Adverse Events": [
     "Adverse Events 1:",
     "Total: 0/10 (0.00%)",
     "Adverse Events 2:"
  ]
}
```

E.1 Argument-Premise algorithm

The below is the pseudocode of the input processing that is consumed by the systems.

```
algorithm argument-premise-map is
        input: ICO Q
        output:Mapm<argument, premise>
  argument := Q.argument
  PrimaryStudyPICO ←readFile(Q.id)
 if Q type is Single
   foreach premise index in Q.evidenceIndex do
          premise :=PrimaryStudyPICO[index]
          map := {(argument, premise)}
 if Q type is Comparison
   SecondaryStudyPICO ← readFile(Q.id)
   foreach premise index in Q.PrimaryEvidenceIndex do
        premise := PrimaryStudyPICO[index]
        map := {(argument, premise)}
   foreach premise index in Q.SecondaryEvidenceIndex do
        premise := SecondaryStudyPICO[index]
        map := {(argument, premise)}
 return map
```

F.1 Examples

"**3facad41-0221-42f8-834d-470e65c4aad5**": {
    "Type": "Single",
    "Section_id": "Results",
    "Primary_id": "NCT00428922",
    "Statement": "the outcome measurement of the primary trial is The length of time during and after the treatment that a patient survives with the disease ",
    "Label": "Contradiction",
    "Primary_evidence_index": [
      " Progression-free Survival (PFS) and to Evaluate Safety of the Trastuzumab, Bevacizumab and Docetaxel Regimen.",
      " The trial was designed as a single-stage phase II rather then usual two-stage design because of the progression free survival(…)",
    ]
  }

The premises (Primary_evidence_index) of the entry **3facad41-0221-42f8-834d-470e65c4aad5** fail to sustain the statement thus the contradiction label.

"**4a75574c-fa86-4e62-a210-81c7b98a3807**": {
    "Type": "Single",
    "Section_id": "Eligibility",
    "Primary_id": "NCT00022516",
    "Statement": "T4 N2 M4 patients are eligible for the primary trial",
    "Label": "Contradiction",
    "Primary_evidence_index": [
      "T1-3, N0-2, M0"
    ]
  }