Weryfikacja faktów w konkursie FEVER

Piotr Niewiński (Samsung R&D Polska), **Aleksander Wawer**, **Grzegorz Wojdyga** (Instytut Podstaw Informatyki PAN)

Dataset

Dataset consists of 185 445 claims generated by altering sentences extracted from Wikipedia.

Split	SUPPORTED	REFUTED	NEI
Training	80,035	29,775	35,639
Dev	3,333	3,333	3,333
Test	3,333	3,333	3,333
Reserved	6,666	6,666	6,666

Table 1: Dataset split sizes for SUPPORTED, REFUTED and NOTENOUGHINFO (NEI) classes

Dataset

Strict logical rules for annatations: Examples:

- claim: “Shakira is Canadian”
  evidence: “Shakira is a Colombian singer, songwriter, dancer, and record producer”.
- claim: “David Beckham was with United”,
  evidence: “David Beckham made his European League debut playing for Manchester United”

Excerpt from trainset


```
{"id": 150448, "verifiable": "VERIFIABLE", "label": "SUPPORTS", "claim": "Roman Atwood is a content creator.", "evidence": [[[174271, 187498, "Roman_Atwood", 1]], [[174271, 187499, "Roman_Atwood", 3]]]}

{"id": 214861, "verifiable": "VERIFIABLE", "label": "SUPPORTS", "claim": "History of art includes architecture, dance, sculpture, music, painting, poetry literature, theatre, narrative, film, photography and graphic arts.", "evidence": [[[255136, 254645, "History_of_art", 2]]]}

{"id": 156709, "verifiable": "VERIFIABLE", "label": "REFUTES", "claim": "Adrienne Bailon is an accountant.", "evidence": [[[180804, 193183, "Adrienne_Bailon", 0]]]}

{"id": 83235, "verifiable": "NOT VERIFIABLE", "label": "NOT ENOUGH INFO", "claim": "System of a Down briefly disbanded in limbo.", "evidence": [[[100277, null, null, null]]]}

{"id": 129629, "verifiable": "VERIFIABLE", "label": "SUPPORTS", "claim": "Homeland is an American television spy thriller based on the Israeli television series Prisoners of War.", "evidence": [[[151831, 166598, "Homeland_-LRB-TV_series-RRB-", 0], [151831, 166598, "Prisoners_of_War_-LRB-TV_series-RRB-", 0]]]}

{"id": 149579, "verifiable": "NOT VERIFIABLE", "label": "NOT ENOUGH INFO", "claim": "Beautiful reached number two on the Billboard Hot 100 in 2003.", "evidence": [[[173384, null, null, null]]]}

{"id": 229289, "verifiable": "NOT VERIFIABLE", "label": "NOT ENOUGH INFO", "claim": "Neal Schon was named in 1954.", "evidence": [[[273626, null, null, null]]]}
```

Task description

The task challenged participants to classify whether human-written factoid claims could be SUPPORTED or REFUTED using evidence retrieved from Wikipedia.

Given the Wikipedia and one claim, participants must verify whether it is true, false, or there are not enough info. Also, an answer should contain information about the wiki articles and sentences based on which the verification was performed.

Task description

Hence, there are in fact three tasks:

- document retrieval
- sentence selection
- claim verification

The single prediction is considered to be correct if and only if both the label is correct and the predicted evidence set (containing at most five sentences).

Baseline

Easy system provided by organizers:

- Document retrieval using TF-IDF
- Sentence selection using TF-IDF
- Claim verification using ESIM

Results

Rank	Team Name	Evidence (%)			Label	FEVER
		Precision	Recall	F1	Accuracy (%)	Score (%)
1	UNC-NLP	42.27	70.91	52.96	**68.21**	**64.21**
2	UCL Machine Reading Group	22.16	82.84	34.97	67.62	62.52
3	Athene UKP TU Darmstadt	23.61	**85.19**	36.97	65.46	61.58
4	Papelo	**92.18**	50.02	**64.85**	61.08	57.36
5	SWEEPer	18.48	75.39	29.69	59.72	49.94
6	Columbia NLP	23.02	75.89	35.33	57.45	49.06
7	Ohio State University	77.23	47.12	58.53	50.12	43.42
8	GESIS Cologne	12.09	51.69	19.60	54.15	40.77
9	FujiXerox	11.37	29.99	16.49	47.13	38.81
10	*withdrawn*	46.60	51.94	49.12	51.25	38.59
11	Uni-DuE Student Team	50.65	36.02	42.10	50.02	38.50
12	Directed Acyclic Graph	51.91	36.36	42.77	51.36	38.33
13	*withdrawn*	12.90	54.58	20.87	53.97	37.13
14	Py.ro	21.15	49.38	29.62	43.48	36.58
15	SIRIUS-LTG-UIO	19.19	70.82	30.19	48.87	36.55
16	*withdrawn*	0.00	0.01	0.00	33.45	30.20
17	BUPT-NLPer	45.18	35.45	39.73	45.37	29.22
18	*withdrawn*	23.75	86.07	37.22	33.33	28.67
19	*withdrawn*	7.69	32.11	12.41	50.80	28.40
20	FEVER Baseline	11.28	47.87	18.26	48.84	27.45

Figure 5: Results of FEVER 1.0

1st place article

The best system was one, described in ”Combining Fact Extraction and Verification with Neural Semantic Matching Networks” Nie et al.

1st place article

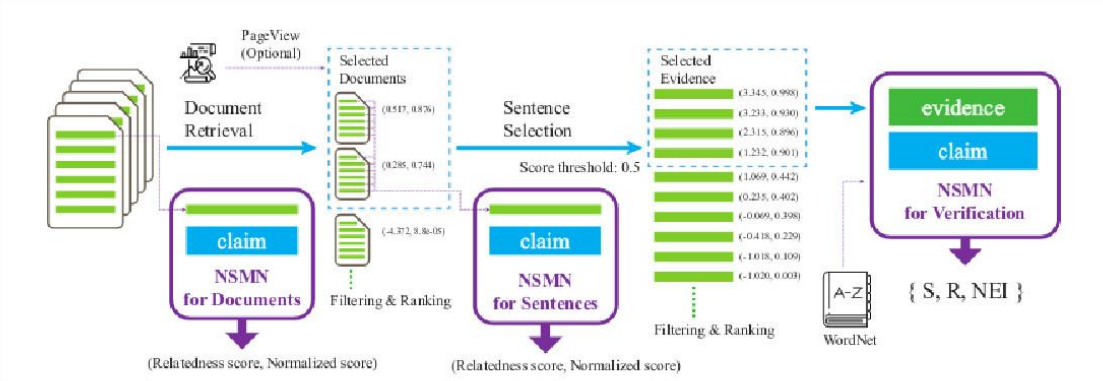

Figure 3: System used by Nie et al.

1st place article

Novelties:

- joint system consisting of three connected homogeneous networks for the 3-stage FEVER task
- use of external Pageview frequency information
- using additional semanticontological features from WordNet

1st place article

$$\mathbf{p} = \text{Maxpool}_{\text{row}}(\mathbf{P}) \in \mathbb{R}^{d_3},$$
$$\mathbf{q} = \text{Maxpool}_{\text{row}}(\mathbf{Q}) \in \mathbb{R}^{d_3},$$
$$h(\mathbf{p}, \mathbf{q}, |\mathbf{p} - \mathbf{q}|, \mathbf{p} \circ \mathbf{q}) = \mathbf{m},$$

$$\mathbf{P} = \text{BiLSTM}([\mathbf{S}, \mathbf{U}^*]) \in \mathbb{R}^{d_3 \times n},$$
$$\mathbf{Q} = \text{BiLSTM}([\mathbf{T}, \mathbf{V}^*]) \in \mathbb{R}^{d_3 \times m}.$$

$$\mathbf{S} = f([\bar{\mathbf{U}}, \tilde{\mathbf{U}}, \bar{\mathbf{U}} - \tilde{\mathbf{U}}, \bar{\mathbf{U}} \circ \tilde{\mathbf{U}}]) \in \mathbb{R}^{d_2 \times n},$$
$$\mathbf{T} = f([\bar{\mathbf{V}}, \tilde{\mathbf{V}}, \bar{\mathbf{V}} - \tilde{\mathbf{V}}, \bar{\mathbf{V}} \circ \tilde{\mathbf{V}}]) \in \mathbb{R}^{d_2 \times m}$$

$$\tilde{\mathbf{U}} = \bar{\mathbf{V}} \cdot \text{Softmax}_{\text{col}}(\mathbf{E}^\top) \in \mathbb{R}^{d_1 \times n},$$
$$\tilde{\mathbf{V}} = \bar{\mathbf{U}} \cdot \text{Softmax}_{\text{col}}(\mathbf{E}) \in \mathbb{R}^{d_1 \times m},$$

$$\mathbf{E} = \bar{\mathbf{U}}^\top \bar{\mathbf{V}} \in \mathbb{R}^{n \times m}.$$

$$\bar{\mathbf{U}} = \text{BiLSTM}(\mathbf{U}) \in \mathbb{R}^{d_1 \times n},$$
$$\bar{\mathbf{V}} = \text{BiLSTM}(\mathbf{V}) \in \mathbb{R}^{d_1 \times m},$$

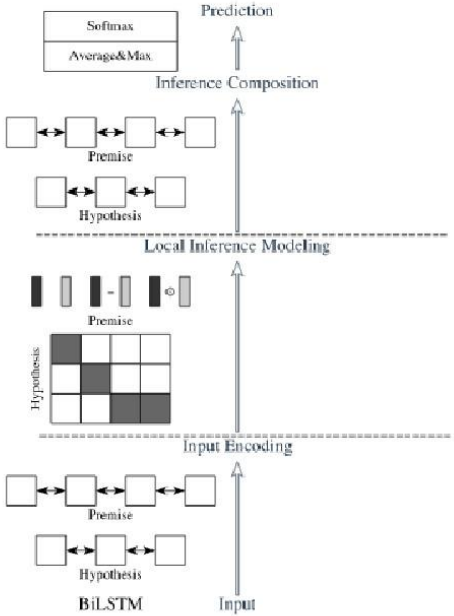

Figure 4: System used by Nie et al.

NSMN Claim Verification

During this task additional features are added:

1. WordNet: 30-dimension indicator features regarding onto-logical information from Wordnet. The 30 dimensions are divided into 10 embedding channels corresponding to 10 hypernyms / antonyms and edge-distance based phenomena,as shown in table 1 .
2. Number: We use 5-dimension real-value embeddings to encode any unique number token. This feature assists the model in identifying and differentiating numbers.
3. Normalized Semantic Relatedness Score: Two normalised relatedness scores, namely the two p(x= 1|ci, j) values produced by the document and sentence NSMN, respectively.

NSMN Claim Verification

Exact same lemma
Antonym
Hyponym
Hypernym
Hyponym with 1-edge distance in WN topological graph
Hypernym with 1-edge distance in WN topological graph
Hyponym with 2-edge distance in WN topological graph
Hypernym with 2-edge distance in WN topological graph
Hyponym with distance >2 edges in WN topological graph
Hypernym with distance >2 edges in WN topological graph

Table 1: Part of WordNet embedding

2nd place article

The second system was one, described in ”Ucl machine reading group: Four factor framework for fact finding” by Yoneda et al.

System description

This solution is a clever improvement of baseline. System is a four stage model:

1. document retrieval
2. sentence retrieval
3. NLI
4. aggregation

NLI

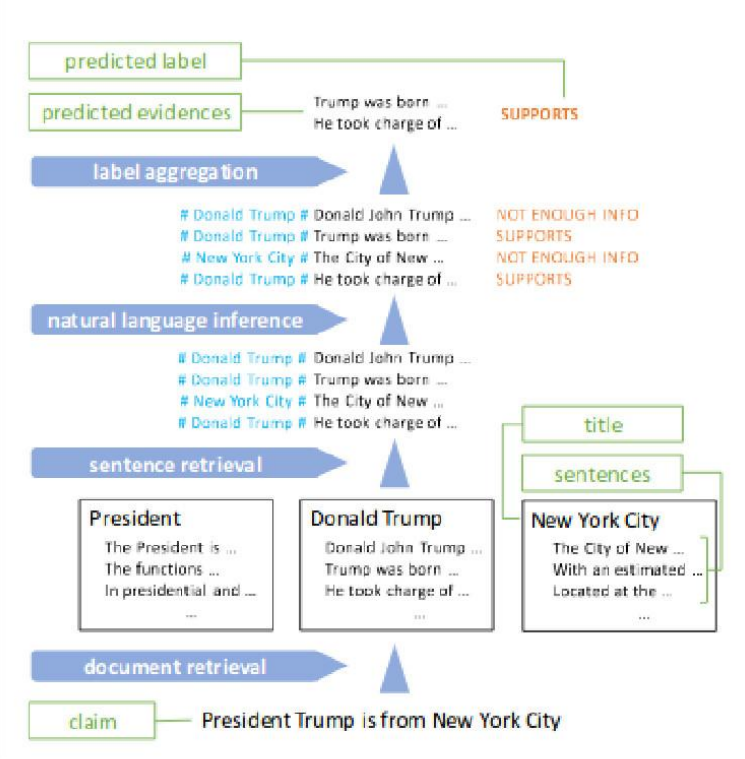

Figure 6: Model overview

Aggregation

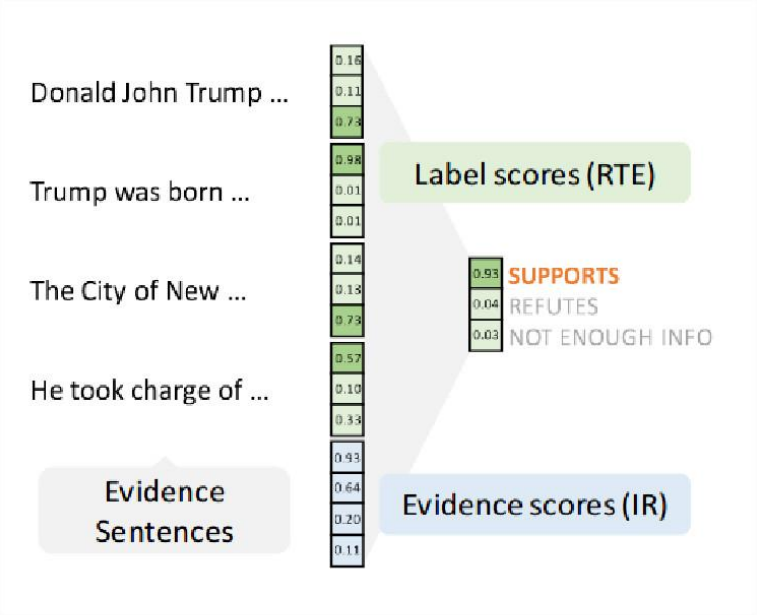

Figure 8: Aggregation Network [4]

3rd place article

The third system was one, described in ”UKP-Athene: Multi-Sentence Textual Entailment for Claim Verification” by Hanselowski et al. [1].

Document retrieval

Three steps:

1. Mention extraction – after parsing the claim, they considered every noun phrase as a potential entity mention (a heuristic that adds all words in the claim before the main verb)
2. Candidate article search – use the MediaWiki API to search through the titles of all Wikipedia articles for matches with the potential entity mentions found in the claim
3. Candidate filtering –remove results that are longer than the entity mention and do not overlap with the rest of the claim

4th place article

The fourth system was one, described in ”Team Papelo: Transformer Networks at FEVER” by Malon [2].

It used TF-IDF + some extensions for document retrieval and sentence selection. For claim verification it uses transformer network.

Transformer network

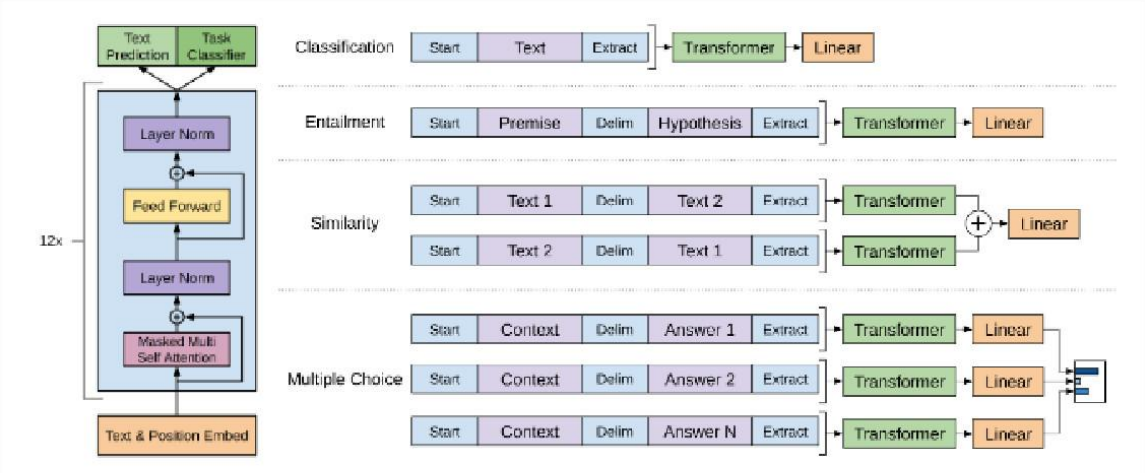

Figure 11: Transformer network for various tasks introduced by Radford et al. [3]

Fever 2.0

Novelties

- Systems must be provided with docker image to work in real-time.
- There were two additional phases: breakers and fixers.
- Very few new teams...

Technical details

Participants must submit a docker image (via dockerhub) on which it is possible to run a flask web server that accept requests and provide answers. Based on this mechanism, the evaluation is performed.

This time, the final score was resilience – the FEVER score over all the accepted instances generated by all the breakers.

Results

System	Resilience (%)	FEVER Score (%)
Papelo	37.31	57.36
UCLMR	35.83	62.52
DOMLIN	35.82	68.46
CUNLP	32.92	67.08
UNC	30.47	64.21
Athene	25.35	61.58
GPLSI	19.63	58.07
Baseline	11.06	27.45

Figure 12: Winners of builders phase

Results

System	Correct Rate (%)	Potency (%)
TMLab	84.81	66.83
CUNLP	81.44	55.79
NbAuzDrLqg	64.71	51.54
Rule-based Baseline	82.33	49.68
Papelo*	91.00	64.79

Figure 13: Winners of fixers phase

Domlin system

- two staged sentence selection strategy:
  claim: Ryan Gosling has been to a country in Africa.
  Evidence 1: He [...] has traveled to Chad , Uganda and eastern Congo [...].
  Evidence 2: Chad [...] is a landlocked country in Central Africa
- publicly available document retrieval module
- fine-tuned BERT checkpoints for sentence selection (x2) and as the entailment classifier