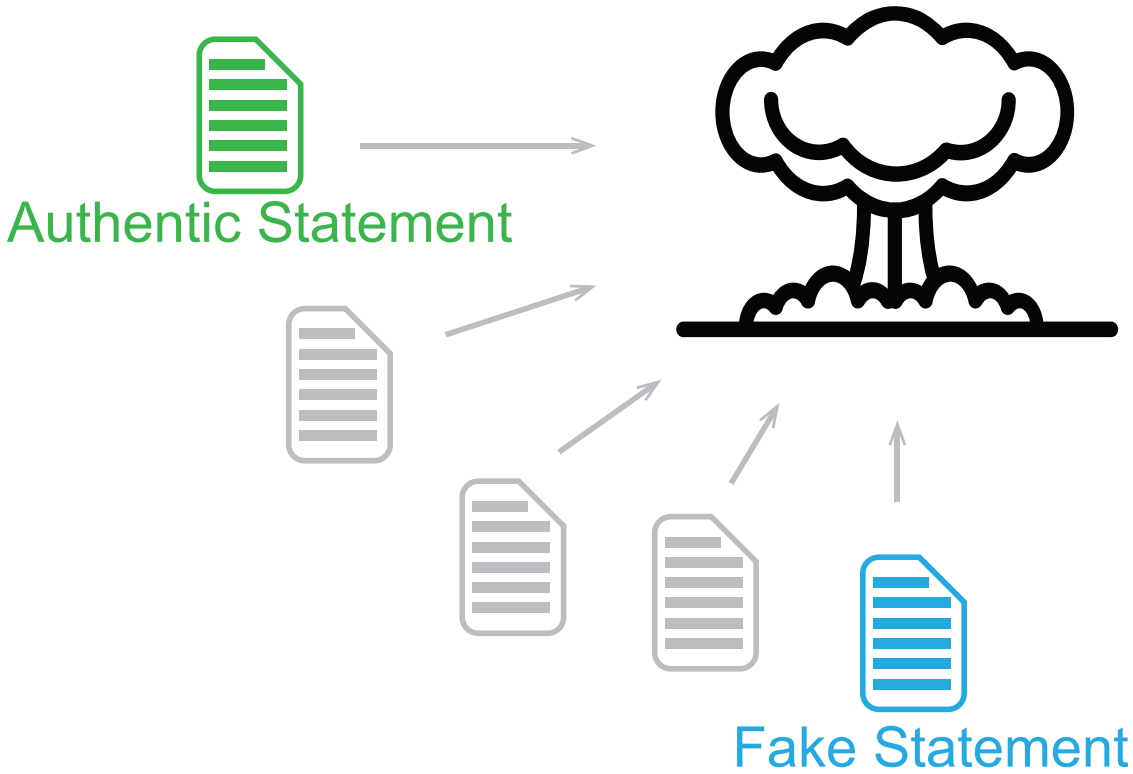
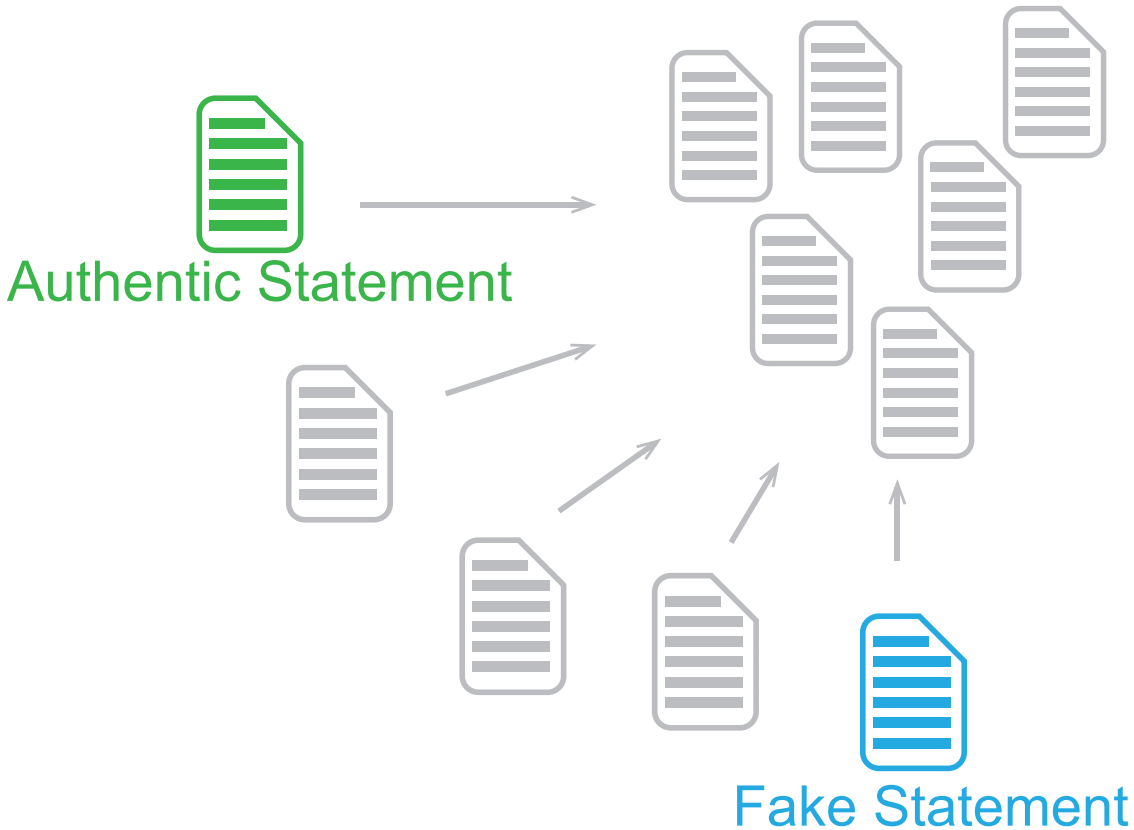
Combining Fact Extraction and Verification with Neural Semantic Matching Networks

Yixin Nie, Haonan Chen, Mohit Bansal

Background and Motivation



Background and Motivation



Background and Motivation

How to discriminate between truth and falsehood?

Task and Dataset

[Thorne et al, NAACL 2018]

Task Formalization:

Input: c (claim); P (evidence set)

Output: $\hat{y}$ (predicted label); $\hat{E}$ (predicted evidence set)

Evaluation:

$y = \hat{y}, E \subseteq \hat{E}$

Task and Dataset

[Thorne et al, NAACL 2018]

3 Subtasks:

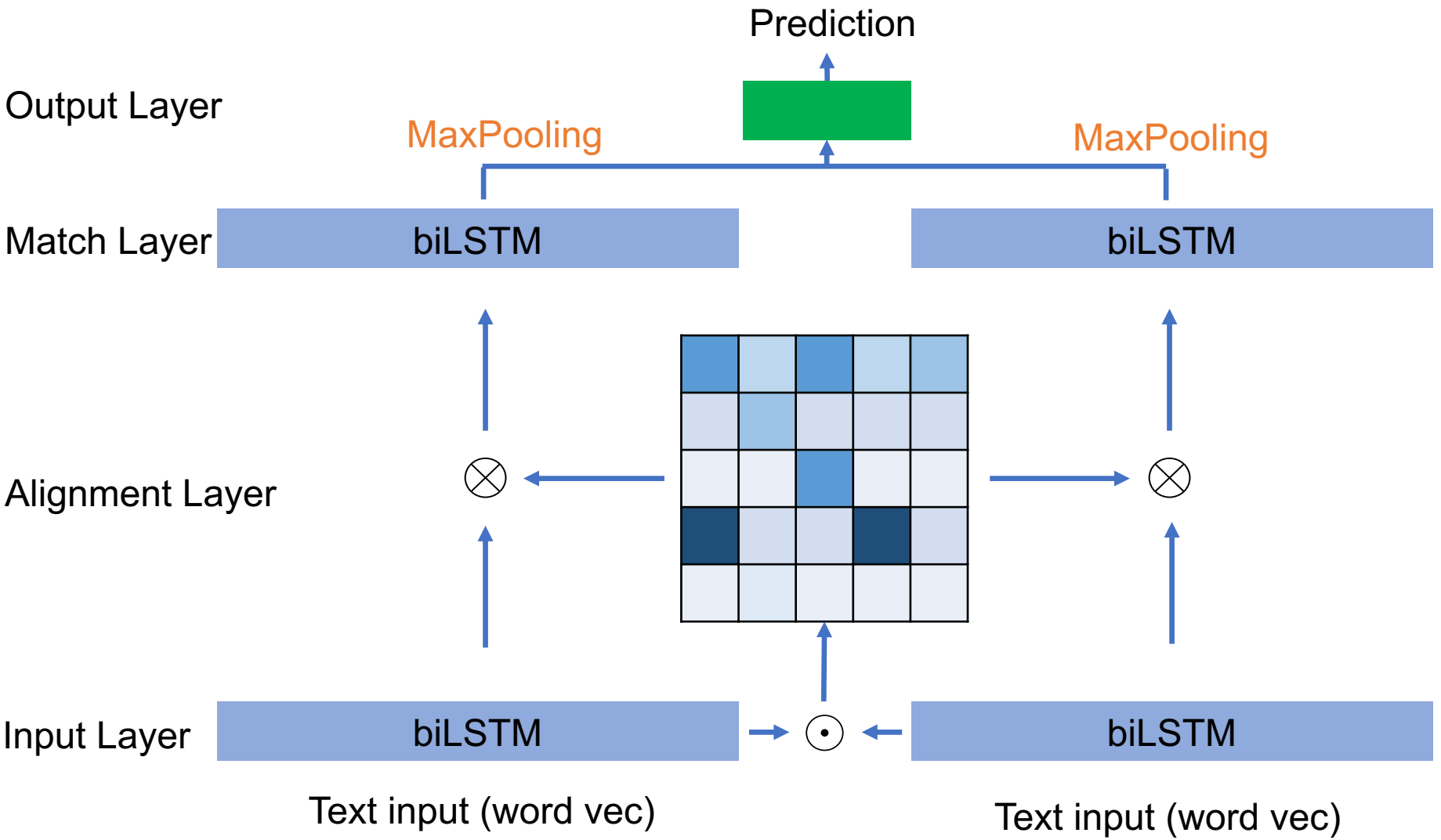
(1) Document Retrieval

(2) Sentence Selection

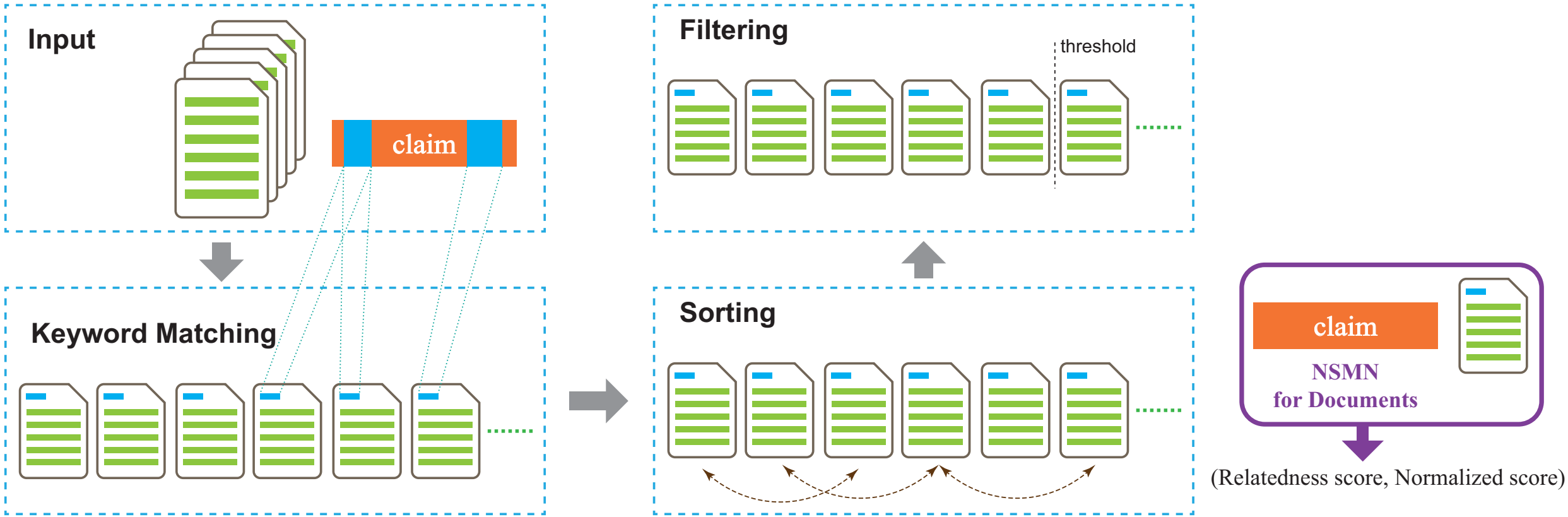
(3) Claim Verification

Neural Semantic Matching Network (NSMN)

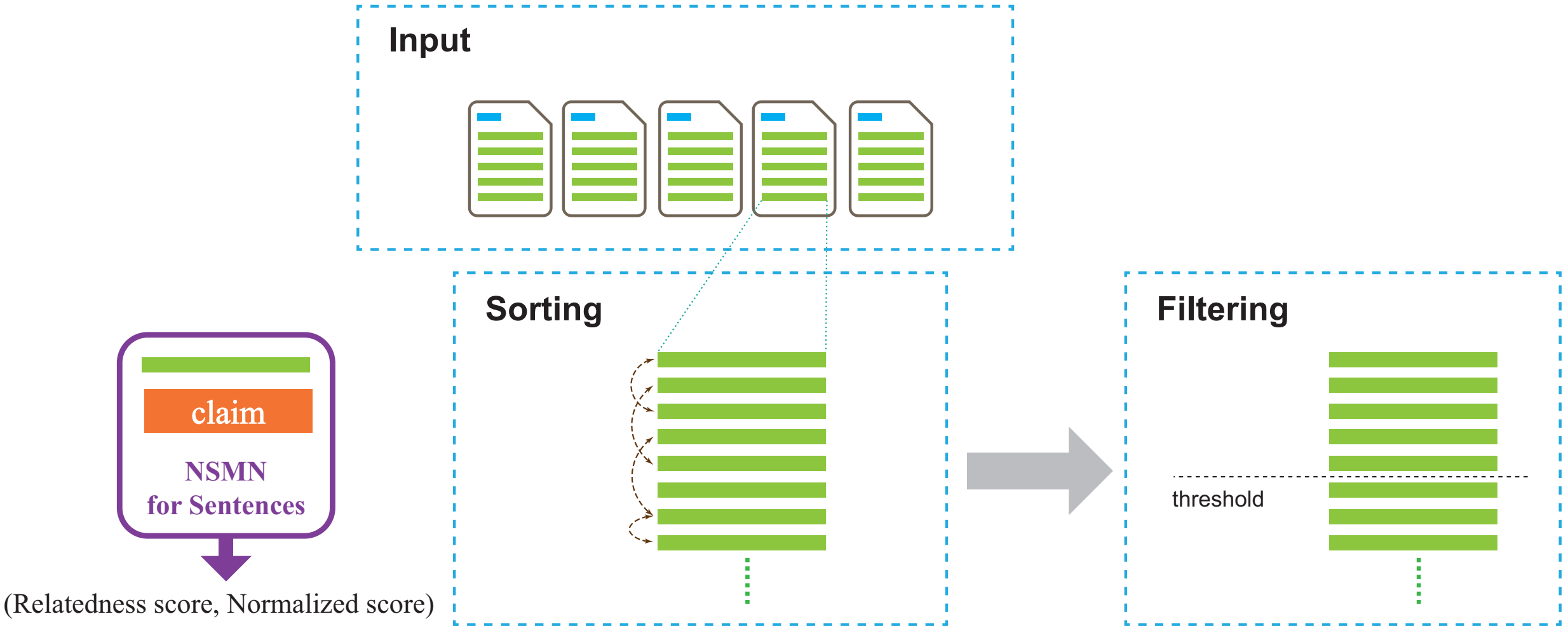
[Nie and Bansal, EMNLP RepEval 2017]



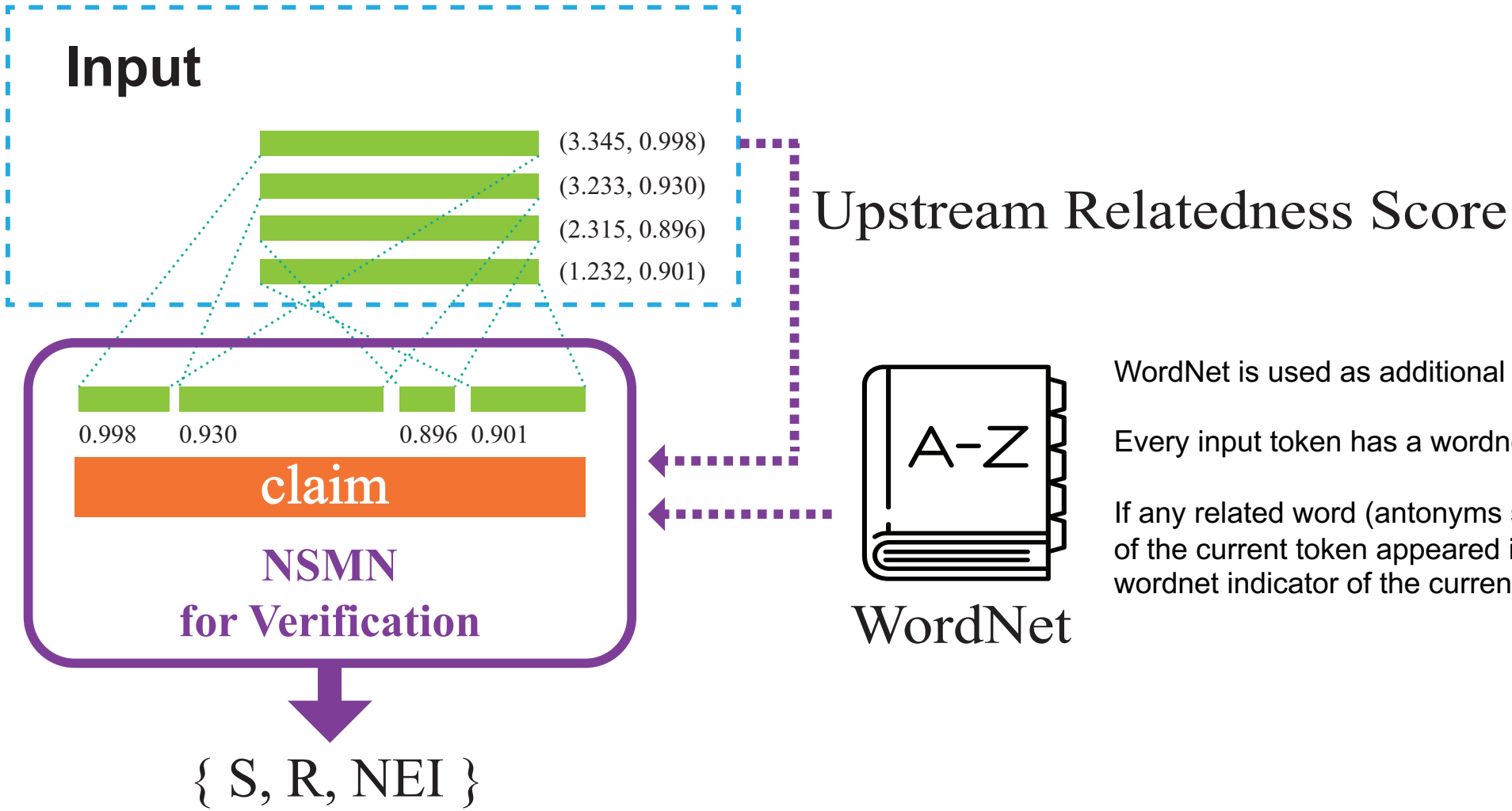
Document Retrieval



Sentence Selection



Claim Verification

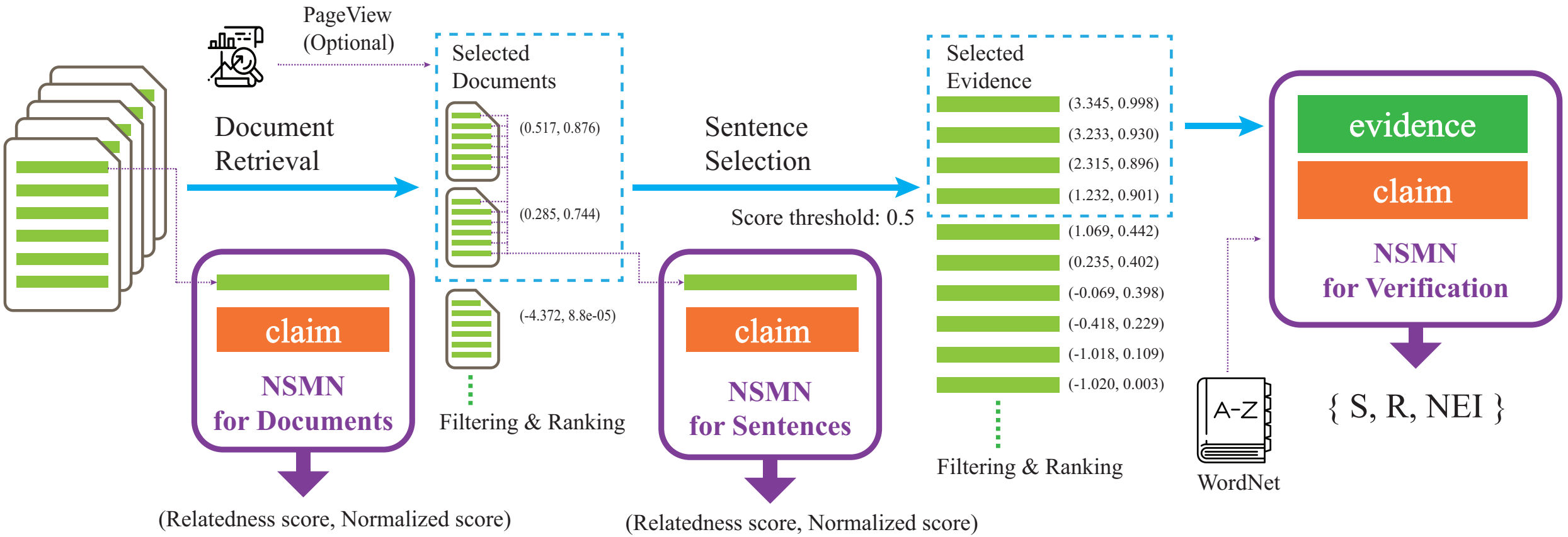


WordNet is used as additional token-level indicator features.

Every input token has a wordnet feature vector.

If any related word (antonyms synonyms, homonyms, etc.) of the current token appeared in the other sequence, the wordnet indicator of the current token will be fired.

System Overview



Results & Analysis (Document Retrieval)

[Thorne et al, NAACL 2018]

Model	Entire Dev Set				Difficult Subset (>10%)			
	OFEVER	Acc.	Recall	F1	**OFEVER**	Acc.	Recall	F1
FEVER Baseline	70.20	–	–	–	–	–	–	–
KM	88.86	44.90	83.30	58.35	60.15	23.89	60.15	34.20
KM + Pageview	91.98	45.90	87.98	60.32	85.61	29.32	85.61	43.68
KM + TF-IDF	91.63	42.83	87.45	57.50	85.60	28.66	85.60	42.94
KM + dNSMN	92.34	52.70	88.51	66.06	87.93	31.71	87.93	46.61
KM + Pageview + dNSMN	**92.42**	**52.73**	**88.63**	**66.12**	**88.73**	**31.90**	**88.72**	**46.93**
k = 5								
FEVER Baseline	77.24	–	–	–	–	–	–	–
KM	90.69	42.61	86.04	56.99	74.34	23.19	74.34	35.36
KM + Pageview	92.69	42.92	89.04	57.92	90.52	24.89	90.52	39.05
KM + TF-IDF	92.38	39.57	88.57	54.70	89.88	23.94	89.88	37.80
KM + dNSMN	**92.82**	51.04	**89.23**	**64.94**	91.33	28.30	91.33	43.21
KM + Pageview + dNSMN	92.75	**51.06**	89.13	64.93	**91.36**	**28.38**	**91.37**	**43.30**
k = 10								

Performance of different document retrieval methods.

K indicates the number of retrieved documents.
Difficult subset is built by choosing examples with least one evidence contained in the “disambiguative” document.

Results & Analysis (Document Retrieval)

[Thorne et al, NAACL 2018]

Model	Entire Dev Set				Difficult Subset (>10%)			
	OFEVER	Acc.	Recall	F1	**OFEVER**	Acc.	Recall	F1
FEVER Baseline	70.20	–	–	–	–	–	–	–
KM	88.86	44.90	83.30	58.35	60.15	23.89	60.15	34.20
KM + Pageview	91.98	45.90	87.98	60.32	85.61	29.32	85.61	43.68
KM + TF-IDF	91.63	42.83	87.45	57.50	85.60	28.66	85.60	42.94
KM + dNSMN	92.34	52.70	88.51	66.06	87.93	31.71	87.93	46.61
KM + Pageview + dNSMN	**92.42**	**52.73**	**88.63**	**66.12**	**88.73**	**31.90**	**88.72**	**46.93**
k = 5								
FEVER Baseline	77.24	–	–	–	–	–	–	–
KM	90.69	42.61	86.04	56.99	74.34	23.19	74.34	35.36
KM + Pageview	92.69	42.92	89.04	57.92	90.52	24.89	90.52	39.05
KM + TF-IDF	92.38	39.57	88.57	54.70	89.88	23.94	89.88	37.80
KM + dNSMN	**92.82**	51.04	**89.23**	**64.94**	91.33	28.30	91.33	43.21
KM + Pageview + dNSMN	92.75	**51.06**	89.13	64.93	**91.36**	**28.38**	**91.37**	**43.30**
k = 10								

Performance of different document retrieval methods.

K indicates the number of retrieved documents.
Difficult subset is built by choosing examples with least one evidence contained in the “disambiguative” document.
dNSMN gives the best and most discriminative sorting performance (better than Pageview).

Results & Analysis (Sentence Selection)

[Thorne et al, NAACL 2018] [Conneau et al, EMNLP 2017]

Method	Entire Dev Set				Difficult Subset (>12%)			
	OFEVER	Acc.	Recall	F1	**OFEVER**	Acc.	Recall	F1
FEVER Baseline	62.81	–	–	–	–	–	–	–
TF-IDF	83.77	34.16	75.65	47.07	53.01	38.54	51.01	44.63
Max-Pool Enc.	84.08	59.52	76.13	66.81	73.68	54.13	73.68	62.41
sNSMN w/o AS	86.65	**69.43**	79.98	**74.33**	68.34	**67.82**	68.34	**68.08**
sNSMN w. AS	**91.19**	36.49	**86.79**	51.38	**81.44**	34.56	**81.44**	48.53

Different methods for sentence selection on dev set.

Difficult subset for sentence selection is built by selecting examples in which the number of word-overlap between the claim and the ground truth evidence is below.

Results & Analysis (Sentence Selection)

[Thorne et al, NAACL 2018] [Conneau et al, EMNLP 2017]

Method	Entire Dev Set				Difficult Subset (>12%)			
	OFEVER	Acc.	Recall	F1	**OFEVER**	Acc.	Recall	F1
FEVER Baseline	62.81	–	–	–	–	–	–	–
TF-IDF	83.77	34.16	75.65	47.07	53.01	38.54	51.01	44.63
Max-Pool Enc.	84.08	59.52	76.13	66.81	73.68	54.13	73.68	62.41
sNSMN w/o AS	86.65	**69.43**	79.98	**74.33**	68.34	**67.82**	68.34	**68.08**
sNSMN w. AS	**91.19**	36.49	**86.79**	51.38	**81.44**	34.56	**81.44**	48.53

Different methods for sentence selection on dev set.

Difficult subset for sentence selection is built by selecting examples in which the number of word-overlap between the claim and the ground truth evidence is below.

Results & Analysis (Sentence Selection)

[Thorne et al, NAACL 2018] [Conneau et al, EMNLP 2017]

Method	Entire Dev Set				Difficult Subset (>12%)			
	OFEVER	Acc.	Recall	F1	**OFEVER**	Acc.	Recall	F1
FEVER Baseline	62.81	–	–	–	–	–	–	–
TF-IDF	83.77	34.16	75.65	47.07	53.01	38.54	51.01	44.63
Max-Pool Enc.	84.08	59.52	76.13	66.81	73.68	54.13	73.68	62.41
sNSMN w/o AS	86.65	**69.43**	79.98	**74.33**	68.34	**67.82**	68.34	**68.08**
sNSMN w. AS	**91.19**	36.49	**86.79**	51.38	**81.44**	34.56	**81.44**	48.53

Different methods for sentence selection on dev set.

Difficult subset for sentence selection is built by selecting examples in which the number of word-overlap between the claim and the ground truth evidence is below.

Results & Analysis (Claim Verification)

[Chen et al, ACL 2017]

Model	FEVER	LA	F1
			S/R/NEI
Final Model	**66.14**	**69.60**	**75.7**/69.4/**63.3**
w/o WN and Num	65.37	68.97	74.7/68.0/63.3
w/o SRS (sent)	64.90	69.07	74.5/**70.7**/60.7
w. SRS (doc)	66.05	69.69	75.6/70.0/62.8
Vanilla ESIM	65.07	68.63	73.9/68.1/63.0
Data from sNSMN			
Final Model	62.48	67.23	72.6/70.4/56.3
Data from TF-IDF			

Final Model:
The vNSMN with semantic relatedness score feature only from sentence selection.

Observations:

- WordNet and Number Embedding Feature improve F1 on `Support' and `Refute'.
- Upstream Semantic Relatedness Score Feature improves F1 on `Not Enough Info'.
- Performance is also sensitive to training data.

Results & Analysis (Claim Verification)

[Chen et al, ACL 2017]

Model	FEVER	LA	F1
			S/R/NEI
Final Model	**66.14**	**69.60**	**75.7**/69.4/**63.3**
w/o WN and Num	65.37	68.97	74.7/68.0/63.3
w/o SRS (sent)	64.90	69.07	74.5/**70.7**/60.7
w. SRS (doc)	66.05	69.69	75.6/70.0/62.8
Vanilla ESIM	65.07	68.63	73.9/68.1/63.0
Data from sNSMN			
Final Model	62.48	67.23	72.6/70.4/56.3
Data from TF-IDF			

Final Model:
The vNSMN with semantic relatedness score feature only from sentence selection.

Observations:

- WordNet and Number Embedding Feature improve F1 on `Support' and `Refute'.
- Upstream Semantic Relatedness Score Feature improves F1 on `Not Enough Info'.
- Performance is also sensitive to training data.

Results & Analysis (Claim Verification)

[Chen et al, ACL 2017]

Model	FEVER	LA	F1
			S/R/NEI
Final Model	**66.14**	**69.60**	**75.7**/69.4/**63.3**
w/o WN and Num	65.37	68.97	74.7/68.0/63.3
w/o SRS (sent)	64.90	69.07	74.5/**70.7**/60.7
w. SRS (doc)	66.05	69.69	75.6/70.0/62.8
Vanilla ESIM	65.07	68.63	73.9/68.1/63.0
Data from sNSMN			
Final Model	62.48	67.23	72.6/70.4/56.3
Data from TF-IDF			

Final Model:
The vNSMN with semantic relatedness score feature only from sentence selection.

Observations:

- WordNet and Number Embedding Feature improve F1 on `Support' and `Refute'.
- Upstream Semantic Relatedness Score Feature improves F1 on `Not Enough Info'.
- Performance is also sensitive to training data.

Results & Analysis (Claim Verification)

[Chen et al, ACL 2017]

Model	FEVER	LA	F1
			S/R/NEI
Final Model	**66.14**	**69.60**	**75.7**/69.4/**63.3**
w/o WN and Num	65.37	68.97	74.7/68.0/63.3
w/o SRS (sent)	64.90	69.07	74.5/**70.7**/60.7
w. SRS (doc)	66.05	69.69	75.6/70.0/62.8
Vanilla ESIM	65.07	68.63	73.9/68.1/63.0
Data from sNSMN			
Final Model	62.48	67.23	72.6/70.4/56.3
Data from TF-IDF			

Final Model:
The vNSMN with semantic relatedness score feature only from sentence selection.

Observations:

- WordNet and Number Embedding Feature improve F1 on `Support' and `Refute'.
- Upstream Semantic Relatedness Score Feature improves F1 on `Not Enough Info'.
- Performance is also sensitive to training data.

Results & Analysis (Noise Tolerance)

Threshold	**FEVER**	LA	Acc.	Recall	F1
0.5	66.15	69.64	36.50	86.69	51.37
0.3	66.42	69.76	33.17	86.90	48.01
0.1	66.43	69.67	29.83	86.97	44.42
0.05	**66.49**	**69.72**	28.64	87.00	43.10

Dev set results for claim verification on data with different degrees of noise.

The findings encourage our usage of **annealed sampling** during sentence selection training and providing high evidence **recall** for the final fact verification model.

Results & Analysis (Final Combination)

Combination	FEVER
Pageview + dNSMN + sNSMN + vNSMN	**66.59**
dNSMN + sNSMN + vNSMN	66.50
Pageview + sNSMN + vNSMN	66.43

We choose our final model as the combination of Pageview and NSMN for blind test evaluation (though the non-Pageview neural-only model is still comparable).

Leaderboard

UNC NLP

Rank	Δ	Team	Evidence F1	Δ	Accuracy	Δ	FEVER Score	Δ
1		UNC-NLP	0.5322	+0.0026	0.6798	-0.0023	0.6398	-0.0023
2		UCL Machine Reading Group	0.3521	+0.0024	0.6744	-0.0018	0.6234	-0.0019
3		Athene UKP TU Darmstadt	0.3733	+0.0036	0.6522	-0.0024	0.6132	-0.0026
4		Papelo	0.6471	-0.0013	0.6074	-0.0034	0.5704	-0.0032
5		SWEEPer	0.2994	+0.0025	0.5964	-0.0009	0.4986	-0.0009
6		ColumbiaNLP	0.3547	+0.0014	0.5728	-0.0018	0.4888	-0.0018
7		The Ohio State University	0.5854	+0.0001	0.4989	-0.0022	0.4322	-0.0020
8		GESIS Cologne	0.1981	+0.0021	0.5395	-0.0021	0.4058	-0.0019
9	+1	nayeon7lee	0.4929	+0.0017	0.5125	-0.0001	0.3858	-0.0002
10	-1	FujiXerox	0.1657	+0.0008	0.4677	-0.0037	0.3850	-0.0032
11		JanK	0.4218	+0.0008	0.4978	-0.0023	0.3831	-0.0020
12		Directed Acyclic Graph	0.4295	+0.0018	0.5122	-0.0014	0.3824	-0.0009
13		jg	0.2117	+0.0030	0.5404	+0.0007	0.3721	+0.0009
14	+1	SIRIUS-LTG-UIO	0.3037	+0.0018	0.4898	+0.0012	0.3664	+0.0010
15	-1	Py.ro	0.2977	+0.0015	0.4318	-0.0030	0.3630	-0.0028
16		hanshan	0.0000	+0.0000	0.3307	-0.0038	0.2982	-0.0038
17		lisizhen	0.3971	-0.0001	0.4517	-0.0021	0.2898	-0.0024
18		HZ	0.3722	+0.0000	0.3333	+0.0000	0.2867	+0.0000
19		UCSB	0.1255	+0.0014	0.5070	-0.0010	0.2835	-0.0005
20		FEVER Baseline	0.1866	+0.0040	0.4892	+0.0008	0.2771	+0.0026
21		ankur-umbc	0.3699	+0.0003	0.4489	+0.0000	0.2369	-0.0007
22		m6.ub.6m.bu	0.1673	+0.0008	0.5722	-0.0010	0.2275	-0.0015
23		ubub.bubu.61	0.1678	+0.0010	0.5528	-0.0014	0.2154	-0.0017
24		mithunpaul08	0.1866	+0.0040	0.3715	+0.0022	0.1928	+0.0028

Final Results

Model	F1	LA	FEVER
UNC-NLP (our shared task model)	52.96	68.21	64.21
UCL Machine Reading Group	34.97	67.62	62.52
Athene UKP TU Darmstadt	36.97	65.46	61.58
UNC-NLP (our final model)	52.81	68.16	**64.23**

Performance of systems on blind test results.

Example

Claim: The ruins of the ancient roman town of Herculaneum lie near Naples .

Herculaneum

From Wikipedia, the free encyclopedia

For the modern Italian commune, see Ercolano. For other uses, see Herculaneum (disambiguation).

Located in the shadow of Mount Vesuvius, **Herculaneum** (Italian: *Ercolano*) was an ancient Roman town destroyed by volcanic pyroclastic flows in 79 AD. Its ruins are located in the *comune* of Ercolano, Campania, Italy.

As a UNESCO World Heritage Site, it is famous as one of the few ancient cities that can now be seen in much of its original splendour, as well as for having been lost, along with Pompeii, Stabiae, Oplontis and Boscoreale, in the eruption of Mount Vesuvius in AD 79 that buried it. Unlike Pompeii, the deep pyroclastic material which covered it preserved wooden and other organic-based objects such as roofs, beds, doors, food and even some 300 skeletons which were discovered in recent years along the seashore. It had been thought until then that the town had been evacuated by the inhabitants.

Herculaneum was a wealthier town than Pompeii, possessing an extraordinary density of fine houses with, for example, far more lavish use of coloured marble cladding.

Naples

From Wikipedia, the free encyclopedia

"Napoli" redirects here. For other uses, see Napoli (disambiguation) and Naples (disambiguation).

Naples (/ˈneɪpəlz/; Italian: *Napoli* [ˈnaːpoli] (listen), Neapolitan: *Napule* [ˈnɑːpələ] or [ˈnɑːpulə]; Latin: *Neapolis*; Ancient Greek: Νεάπολις, lit. 'new city') is the regional capital of Campania and the third-largest municipality in Italy after Rome and Milan. In 2017, around 967,069 people lived within the city's administrative limits while its province-level municipality has a population of 3,115,320 residents. Its continuously built-up metropolitan area (that stretches beyond the boundaries of the Metropolitan City of Naples) is the second or third largest metropolitan area in Italy.

First settled by Greeks in the second millennium BC, Naples is one of the oldest continuously inhabited urban areas in the world.[3] In the ninth century BC, a colony known as Parthenope or Παρθενόπη was established on the Island of Megaride,[4] later refounded as Neápolis in the sixth century BC.[5] The city was an important part of Magna Graecia, played a major role in the merging of Greek and Roman society and a significant cultural centre under the Romans.[6] It was capital of the Duchy of Naples (661-1139), then the Kingdom of Naples (1282 and 1816) and finally the Two Sicilies until the unification of Italy in 1861.

Between 1925 and 1936, Naples was expanded and upgraded by Benito Mussolini's government but severely damaged by Allied bombing during World War II, leading to extensive post-1945 reconstruction work.[7] Naples has experienced significant economic growth in recent decades, helped by the construction of the Centro Direzionale business district and an advanced transportation network, which includes the Alta Velocità high-speed rail link to Rome and Salerno and an expanded subway network. Naples is the third-largest urban economy in Italy, after Milan and Rome[8]. The Port of Naples is one of the most important in Europe and home of the Allied Joint Force Command Naples, the NATO body that oversees North Africa, the Sahel and Middle East.[9]

Naples' historic city centre is the largest in Europe and a UNESCO World Heritage Site, with a wide range of culturally and historically significant sites nearby, including the Palace of Caserta and the Roman ruins of Pompeii and Herculaneum. Naples is also known for its natural beauties such as Posillipo, Phlegraean Fields, Nisida, and Vesuvius.[10]

Neapolitan cuisine is synonymous with pizza, which originated in the city but it includes many other less well-known dishes and is the Italian city with the highest number of accredited stars from the Michelin Guide.[11]

The best known sports team in Naples is the Serie A club S.S.C. Napoli, two-time Italian champions who play at the San Paolo Stadium in the southwest of the city.

Example

Claim: The ruins of the ancient roman town of Herculaneum lie near Naples .
(Multiple evidences extracted from different sources)

Herculaneum

From Wikipedia, the free encyclopedia

For the modern Italian commune, see Ercolano. For other uses, see Herculaneum (disambiguation).

Located in the shadow of Mount Vesuvius, **Herculaneum** (Italian: *Ercolano*) was an ancient Roman town destroyed by volcanic pyroclastic flows in 79 AD. Its ruins are located in the *comune* of Ercolano, Campania, Italy.

As a UNESCO World Heritage Site, it is famous as one of the few ancient cities that can now be seen in much of its original splendour, as well as for having been lost, along with Pompeii, Stabiae, Oplontis and Boscoreale, in the eruption of Mount Vesuvius in AD 79 that buried it. Unlike Pompeii, the deep pyroclastic material which covered it preserved wooden and other organic-based objects such as roofs, beds, doors, food and even some 300 skeletons which were discovered in recent years along the seashore. It had been thought until then that the town had been evacuated by the inhabitants.

Herculaneum was a wealthier town than Pompeii, possessing an extraordinary density of fine houses with, for example, far more lavish use of coloured marble cladding.

Naples

From Wikipedia, the free encyclopedia

"Napoli" redirects here. For other uses, see Napoli (disambiguation) and Naples (disambiguation).

Naples (/ˈneɪpəlz/; Italian: *Napoli* [ˈnaːpoli] (listen), Neapolitan: *Napule* [ˈnɑːpələ] or [ˈnɑːpulə]; Latin: *Neapolis*; Ancient Greek: Νεάπολις, lit. 'new city') is the regional capital of Campania and the third-largest municipality in Italy after Rome and Milan. In 2017, around 967,069 people lived within the city's administrative limits while its province-level municipality has a population of 3,115,320 residents. Its continuously built-up metropolitan area (that stretches beyond the boundaries of the Metropolitan City of Naples) is the second or third largest metropolitan area in Italy.

First settled by Greeks in the second millennium BC, Naples is one of the oldest continuously inhabited urban areas in the world.[3] In the ninth century BC, a colony known as Parthenope or Παρθενόπη was established on the Island of Megaride,[4] later refounded as Neápolis in the sixth century BC.[5] The city was an important part of Magna Graecia, played a major role in the merging of Greek and Roman society and a significant cultural centre under the Romans.[6] It was capital of the Duchy of Naples (661-1139), then the Kingdom of Naples (1282 and 1816) and finally the Two Sicilies until the unification of Italy in 1861.

Between 1925 and 1936, Naples was expanded and upgraded by Benito Mussolini's government but severely damaged by Allied bombing during World War II, leading to extensive post-1945 reconstruction work.[7] Naples has experienced significant economic growth in recent decades, helped by the construction of the Centro Direzionale business district and an advanced transportation network, which includes the Alta Velocità high-speed rail link to Rome and Salerno and an expanded subway network. Naples is the third-largest urban economy in Italy, after Milan and Rome[8]. The Port of Naples is one of the most important in Europe and home of the Allied Joint Force Command Naples, the NATO body that oversees North Africa, the Sahel and Middle East.[9]

Naples' historic city centre is the largest in Europe and a UNESCO World Heritage Site, with a wide range of culturally and historically significant sites nearby, including the Palace of Caserta and the Roman ruins of Pompeii and Herculaneum. Naples is also known for its natural beauties such as Posillipo, Phlegraean Fields, Nisida, and Vesuvius.[10]

Neapolitan cuisine is synonymous with pizza, which originated in the city but it includes many other less well-known dishes and is the Italian city with the highest number of accredited stars from the Michelin Guide.[11]

The best known sports team in Naples is the Serie A club S.S.C. Napoli, two-time Italian champions who play at the San Paolo Stadium in the southwest of the city.

Example

Claim: The ruins of the ancient roman town of Herculaneum lie near Naples.

Evidence:

Located in the shadow of Mount Vesuvius, **Herculaneum** (Italian: *Ercolano*) was an ancient Roman town destroyed by volcanic pyroclastic flows in 79 AD.

Naples' historic city centre is the largest in Europe and a UNESCO World Heritage Site, with a wide range of culturally and historically significant sites nearby, including the Palace of Caserta and the Roman ruins of Pompeii and Herculaneum.

Prediction: Support

Thanks

Yixin Nie
yixin1@cs.unc.edu
www.cs.unc.edu/~yixin1

Haonan Chen
chaonan99@cs.unc.edu
chaonan99.github.io

Mohit Bansal
mbansal@cs.unc.edu
www.cs.unc.edu/~mbansal

Acknowledgment: Verisk, Google, Facebook