Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis

Stefanos Angelidis and Mirella Lapata
Institute for Language, Cognition, and Computation
School of Informatics, University of Edinburgh

In *Transactions of the Association for Computational Linguistics (TACL)*, 2018.

http://stangelid.github.io
✉ s.angelidis@ed.ac.uk

Edinburgh
University of Edinburgh
Natural Language Processing
NLP

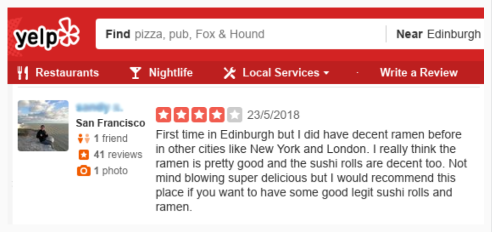
Online User Reviews

Heavily influence customer decisions:

- Travel booking (Ye et al., 2009)
- Box Office success (Duan et al., 2008)
- Shopping (TurnTo.com report, 2018)

Incredibly rich data source:

- 6.3 million Yelp reviews written in 2010
- 27.3 million in 2017



Document-level Sentiment Analysis

Rating: ★★

I had a very mixed experience at The Stand. The burger and fries were good. The chocolate shake was divine! The drive-thru was horrible. It took us at least 30 minutes to order. We complained about the wait and got no apology. I would go back because the food is good, but my only hesitation is the wait.

Document-level Sentiment Analysis

Rating: ★★

I had a very mixed experience at The Stand. The burger and fries were good. The chocolate shake was divine! The drive-thru was horrible. It took us at least 30 minutes to order. We complained about the wait and got no apology. I would go back because the food is good, but my only hesitation is the wait.

[insert favourite neural net here]

$\Downarrow$

Predicted rating: ★★

Document-level Sentiment Analysis

Rating: ⋆⋆

I had a very mixed experience at The Stand. The burger and fries were good. The chocolate shake was divine! The drive-thru was horrible. It took us at least 30 minutes to order. We complained about the wait and got no apology. I would go back because the food is good, but my only hesitation is the wait.

[Johnson and Zhang (2015); Yang et al. (2016); Liu and Lapata (2018)]

$\Downarrow$

Predicted rating: ⋆⋆

Fine-grained Sentiment Analysis

Rating: ★★

I had a very mixed experience at The Stand. The burger and fries were good. The chocolate shake was divine! The drive-thru was horrible. It took us at least 30 minutes to order. We complained about the wait and got no apology. I would go back because the food is good, but my only hesitation is the wait.

Positive:

- The burger and fries were good.
- The chocolate shake was divine!
- I would go back because the food is good.

Negative:

- The drive-thru was horrible.
- It took us at least 30 minutes to order.
- We complained about the wait and got no apology.
- My only hesitation is the wait.

Our Work

Large collections of rated reviews (Diao et al. 2014; Tang et al. 2015)

Detect and summarize fine-grained sentiment

- with Multiple Instance Learning and neural machinery
- w/o expert knowledge
- w/o expensive annotations

Unsupervised: Lexicon-based Methods

The starters were quite bland.

Unsupervised: Lexicon-based Methods

The starters were quite bland.

0 0 0 $1.10 \times -2 \rightarrow$ **-2.2**

Adjective:	
disgusting	−5
terrible	−4
bland	**−2**
so-so	−1
okay	1
great	2
amazing	4
divine	5

Intensifier:	
slightly	0.50
somewhat	0.70
pretty	0.90
quite	**1.10**
really	1.15
very	1.25
extraordinarily	1.50
(the) most	2.00

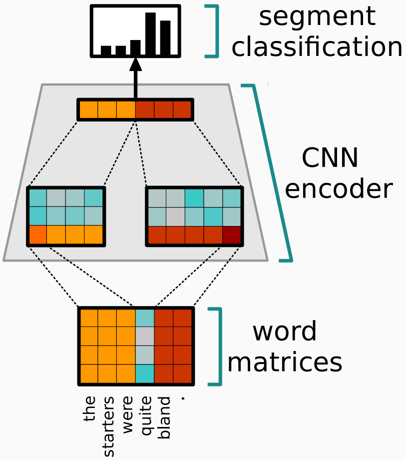
SO-CAL: Semantic Orientation CALculator (Taboada et al., 2011)

Fully Supervised: Segment-level CNNs

The starters were quite bland.	$\mapsto$	very negative
I didn't enjoy most of them,	$\mapsto$	negative
but the burger was brilliant!	$\mapsto$	very positive

Fully Supervised: Segment-level CNNs

The starters were quite bland.	→ very negative
I didn't enjoy most of them,	→ negative
but the burger was brilliant!	→ very positive



Segment CNN (Kim, 2014)

- multiple conv. filters of varying length
- max-over-time pooling

Successful for sentence classification ☺

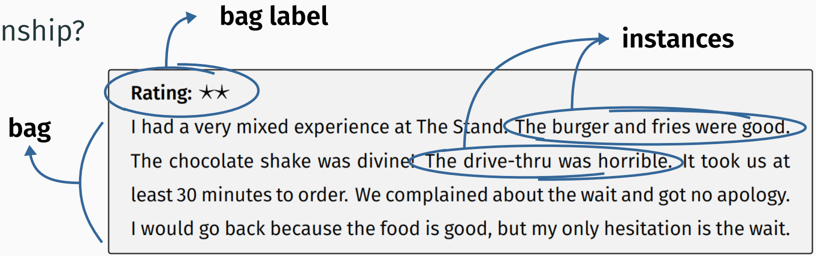
Segment encoder in larger networks ☺

Requires expensive annotations ☹

Our Approach

Multiple Instance Learning (MIL; Keeler and Rumelhart, 1992) *

- Training examples → *bags* of *instances*
- Bag labels → supervision
- Instance labels → latent
- Bag-instance relationship?



* **MIL In NLP:** Surdeanu et al. (2012); Kotzias et al. (2015); Pappas and Popescu-Belis (2017)

Model Assumptions

Sentiment aggregation:

- Segment s_i conveys sentiment polarity: $pol_i \in [-1, +1]$
- Segments have varying degrees of importance: $a_i \in [0, 1]$, $\sum_i a_i = 1$
- Overall polarity of review: average of polarities, weighted by importance

Review segmentation:

- words, phrases
- sentences
- clauses*

* *Elementary Discourse Units* (EDUs) from discourse parser (Feng and Hirst, 2012)

Multiple Instance Learning Network (MilNet)

Inputs:

Word Matrices $\mathbf{X}_i$

Segment encoding:

$\mathbf{v}_i = \mathrm{CNN}(\mathbf{X}_i)$

Segment classification:

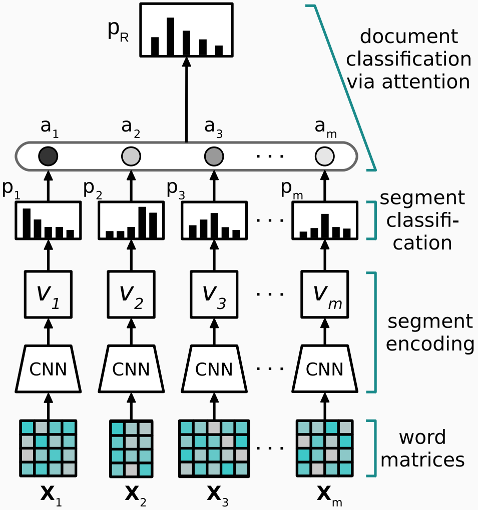
$\mathbf{p}_i = \mathrm{softmax}(\mathbf{W}_c \mathbf{v}_i + \mathbf{b}_c)$

Document classification:

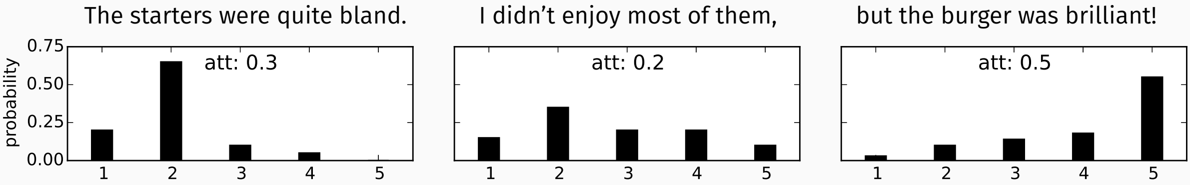
$p_R^{(c)} = \sum_i a_i p_i^{(c)}, c \in \{1, C\}$

Objective:

NLL of document predictions



Polarity Scoring via Gating



Polarity Scoring via Gating

Polarity of segment:

$$pol_i = \sum_c p_i^{(c)} w^{(c)} \;, \; \mathbf{w} = \langle -1, -0.5, 0, +0.5, +1 \rangle$$

Gated polarity $\rightarrow$ **accounts for segment importance:**

$$gpol_i = a_i \cdot pol_i$$

Polarity-based Opinion Extraction

Rating: ⋆⋆

I had a very mixed experience at The Stand. The burger and fries were good. The chocolate shake was divine! The drive-thru was horrible. It took us at least 30 minutes to order. We complained about the wait and got no apology. I would go back because the food is good, but my only hesitation is the wait.

Very negative ↕ Very positive

[+1.00]	**The chocolate shake was divine**
[+0.86]	**I would go back because the food is good**
[+0.50]	**The burger and fries were good**
[-0.05]	I had a very mixed experience at The Stand.
[-0.10]	but my only hesitation is the wait
[-0.10]	and got no apology
[-0.25]	We complained about the wait
[-0.43]	**It took us at least 30 minutes to order**
[-0.89]	**The drive-thru was horrible**

Discrete predictions → Gated polarities → Rankings → Opinion extraction

Experimental Setup: Datasets

Document-level	Yelp'13	IMDB
Documents	335K	348K
Avg # Sentences	8.90	14.02
Avg # EDUs	19.11	37.38
Avg # Words	152	325
Vocabulary Size	129K	97K
Classes	1–5	1–10

$\succ$

Segment-level	Yelp'13	IMDB
Documents	100	100
Sentences	1,065	1,029
EDUs	2,110	2,398
Classes	{– , 0 , +}	

Review collections:

- Yelp'13 and IMDB rated reviews
- Used for training MilNet

Sentiment Polarity (SpoT) dataset:

- Sampled from test splits
- Sentence- and EDU-level
- 3 annotations per segment (Majority Vote; *kappa* $\approx$ 0.8)

Experimental Setup

Segment-level Classification:

- Gated polarities $\rightarrow$ Positive/Neutral/Negative
- For Sentences & EDUs

Comparison Systems:

- **Unsupervised:** SO-CAL (Taboada et al. 2011)
- **Fully-Supervised:** SEG-CNN (Kim, 2014)
- **Document-level:** Hierarchical Attention Network (HIERNET; Yang et al. 2016)

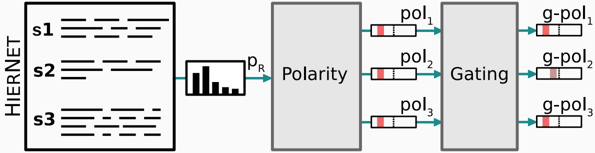
Experimental Setup

Segment-level Classification:

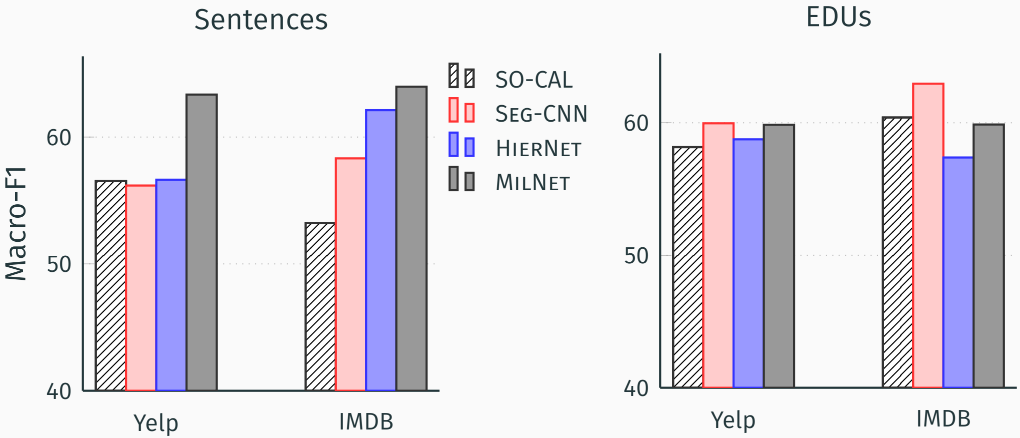
- Gated polarities $\rightarrow$ Positive/Neutral/Negative
- For Sentences & EDUs

Comparison Systems:

- **Unsupervised:** SO-CAL (Taboada et al. 2011)
- **Fully-Supervised:** SEG-CNN (Kim, 2014)
- **Document-level:** Hierarchical Attention Network (HIERNET; Yang et al. 2016)



Results: Segment-level Sentiment



Human Evaluation of Opinion Summaries

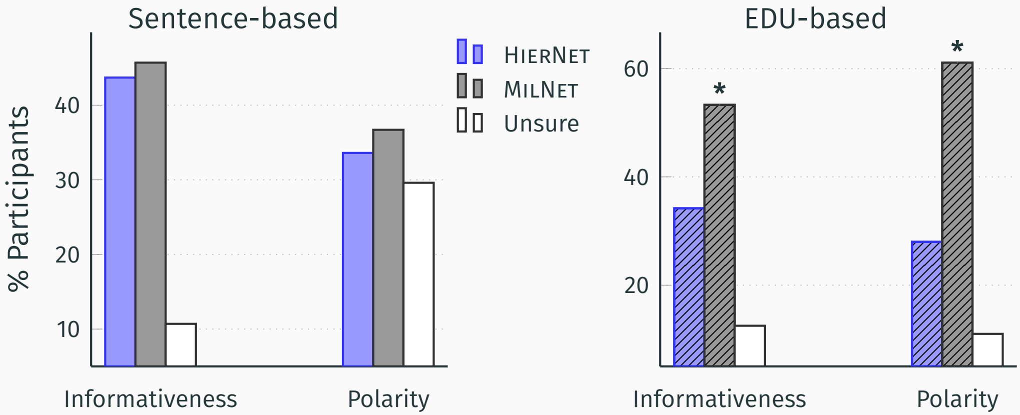
Compare the quality of opinion summaries

- On Yelp & IMDB reviews from SPOT
- Produce extractive summaries from competing models
- Show original review + summaries to 3 human judges

Participants asked to select best summary according to:

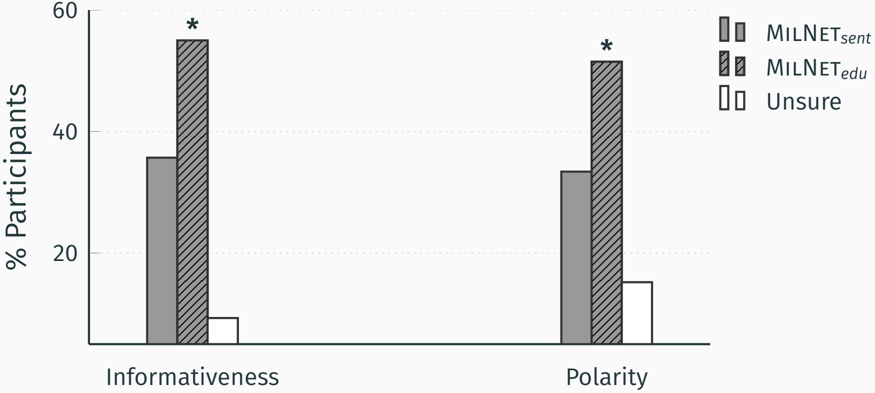
- Informativeness (*Best captures the salient points of the review?*)
- Polarity (*Best highlights positive and negative comments?*)
- *’Not sure’* option available

Is MilNet better than HierNet?



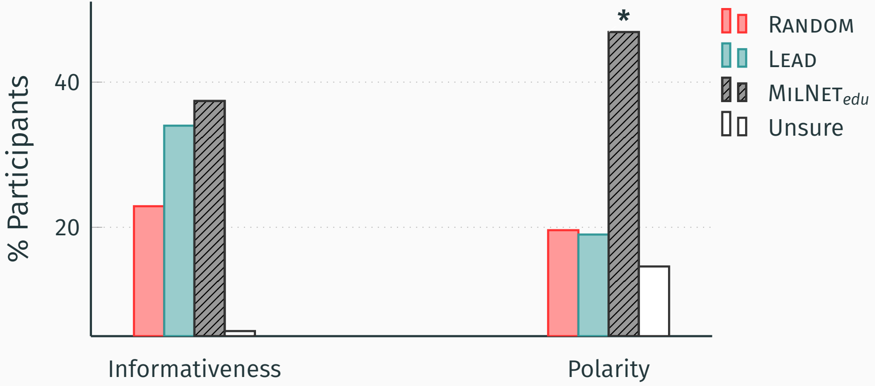
* significant difference (sign-test; $p < 0.01$)

Are EDUs better than Sentences?



* significant difference (sign-test; $p < 0.01$)

How does MilNet compare to Summarization Baselines?



* significant difference (sign-test; $p < 0.01$)

Conclusions

- A MIL neural model for fine-grained sentiment analysis
- Attention-based polarity scoring method facilitates opinion extraction
- Experiments on new test dataset (SPOT)
- Ongoing work: Extends to opinion extraction from multiple-reviews

Thank you

Code + Data:
`stangelid.github.io`

...and some MilNet summaries!

Very tasty and fresh,
I really enjoyed it.
Our server was a bit aloof!
Very sweet girl though.
Haha!

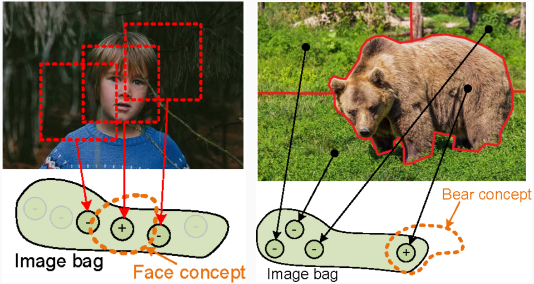
The good things are the acting.
Mostly brilliant, and believable.
On the negative side is, well every-thing else.
I bet even the catering was bad on this film.

I would give zero stars.
it was ice cold.
This was torture!
The staff is clueless.
Horrible service!

Multiple Instance Learning

MIL for object recognition:

- Bags $\longrightarrow$ images
- Instances $\longrightarrow$ image patches
- Bag is positive if at least 1 instance is positive
- OR-style label aggregation

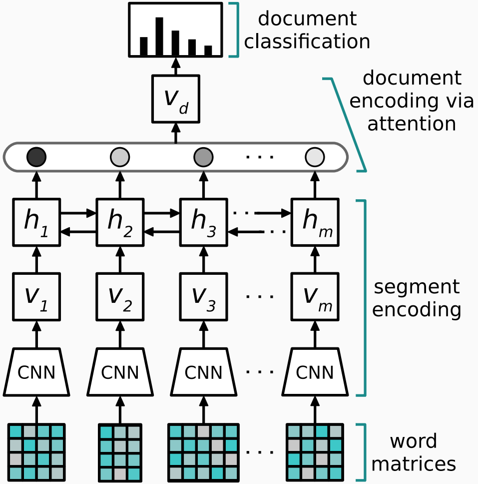


[5] Carbonneau et al. (2016)

Document-level Classification with Hierarchical Networks

Hierarchical Network (HIERNET)

- Based on Yang et al. (2016)
- Attention models segment importance
- Produces fixed-size document-vector
- No natural way to predict segment sentiment

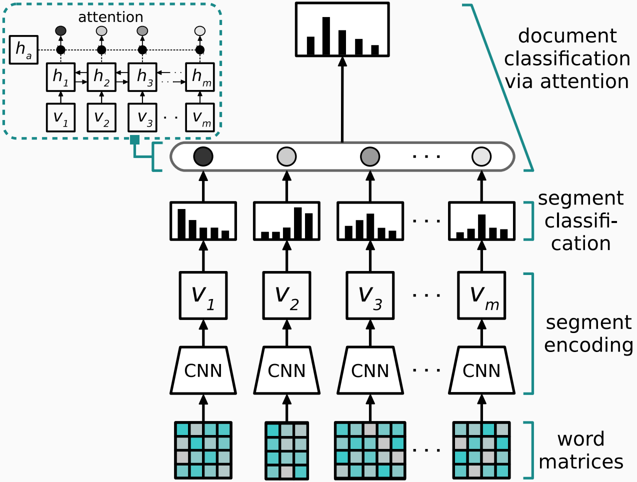


Multiple Instance Learning Network

- **Attention Mechanism:**

$$h_i = \overleftrightarrow{\mathrm{GRU}}(v_i)$$
$$h'_i = tanh(W_a h_i + b_a)$$
$$a_i = \frac{exp(h_i'^{\mathsf{T}} h_a)}{\sum_i exp(h_i'^{\mathsf{T}} h_a)}$$

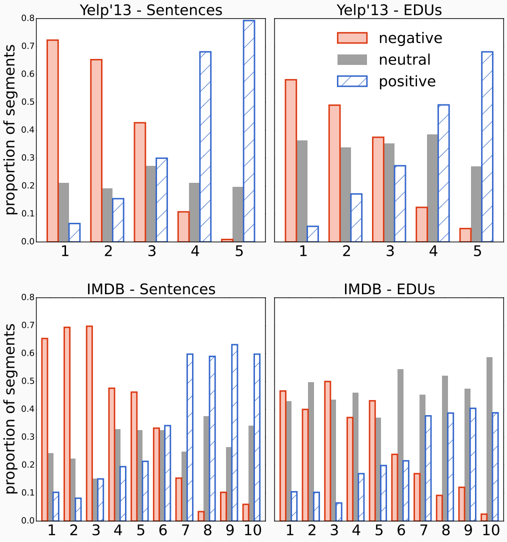
- **Intuition:**
  - GRU encodes segment interrelations
  - Vector h_a is a trained key
  - learns to recognize sentiment-heavy segments



SPOT: Segment-level Polarity Annotations

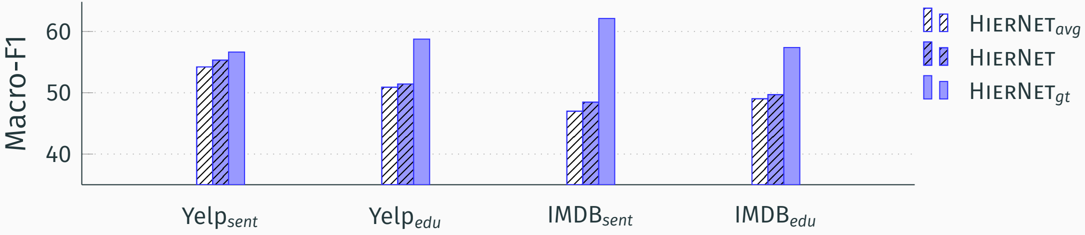
Document segments:	Sentiment conveyed:
I had no particular desire	Positive Neutral **Negative**
to see hulk at the cinema ,	Positive **Neutral** Negative
but seeing	Positive **Neutral** Negative
it has at least convinced me it 's not worth another watch .	Positive Neutral **Negative**
Yet , it all started so well -	**Positive** Neutral Negative
the first 5 minutes of the film	Positive **Neutral** Negative
(the intro sequence) was truly inspiring .	**Positive** Neutral Negative
I was just about to think ' perhaps this will be pretty good , '	**Positive** Neutral Negative

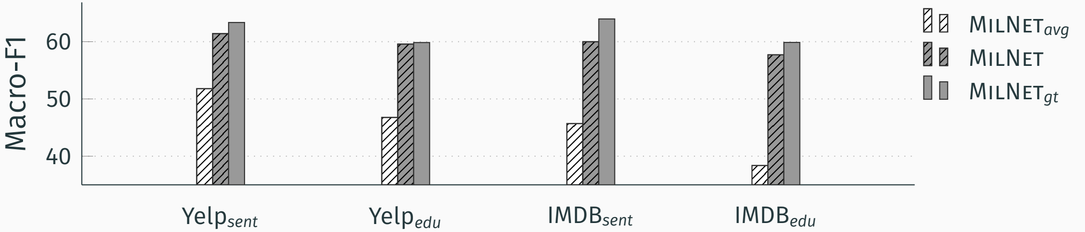
SPOT: Segment-level Polarity Annotations



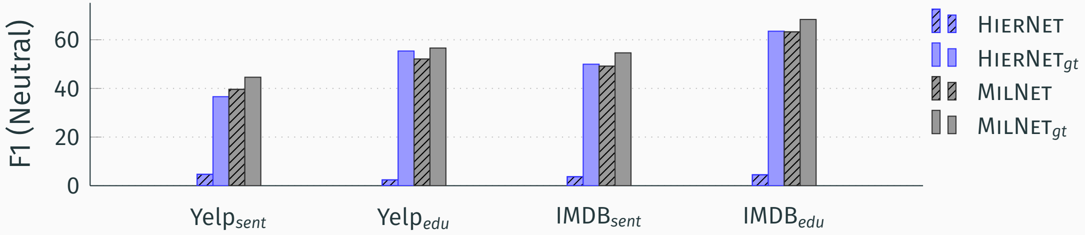
Distribution of segment-level labels

Segment Classification – Effect of Gating

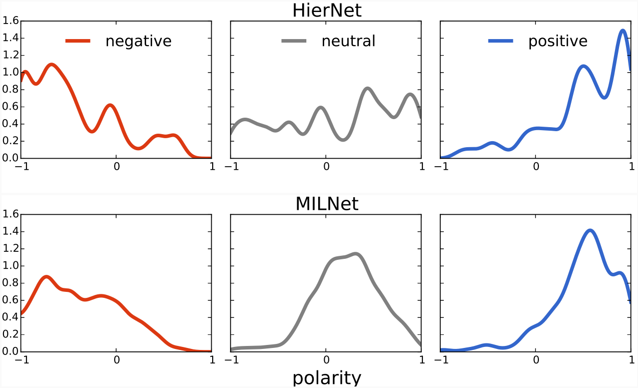




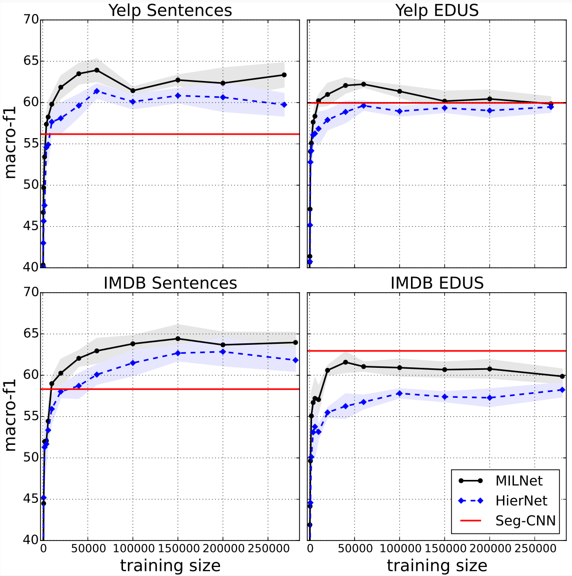
Segment Classification – Effect of Gating (Neutral Class)



Segment Classification - Distribution of Polarities



Segment Classification – Effect of Training Size



Human Evaluation of Opinion Summaries

Original customer review:

This is one of those places that gives you massive portions to allot for their somewhat higher pricing. However, overall i felt it was worth it. We dined on the patio outside, along the golf course, in the evening when it was cooler out. I had a salad, which i mistakenly did not order the half size! I was brought a regular full size which could definitely feed a small family. Haha. Very tasty and fresh, i really enjoyed it. Our server was a bit aloof. She just did n't seem to be there and maybe was a little stressed out or overwhelmed. Very sweet girl though.

Summary 1:

+ However, overall i felt it was worth it.
+ Very tasty and fresh, i really enjoyed it.
+ Very sweet girl though.

– Haha.
– Our server was a bit aloof.

Summary 2:

+ Very tasty and fresh, i really enjoyed it.
+ Very sweet girl though.

– to allot for their somewhat higher pricing.
– when it was cooler out.
– Haha.
– Our server was a bit aloof.

Informativeness:

Summary 1 | Not sure | Summary 2

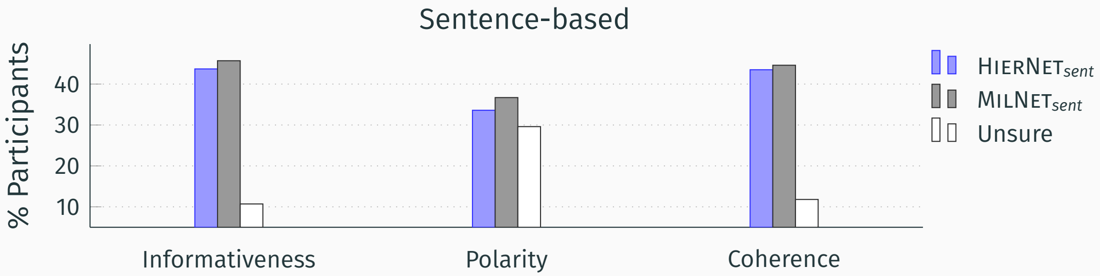
Polarity:

Summary 1 | Not sure | Summary 2

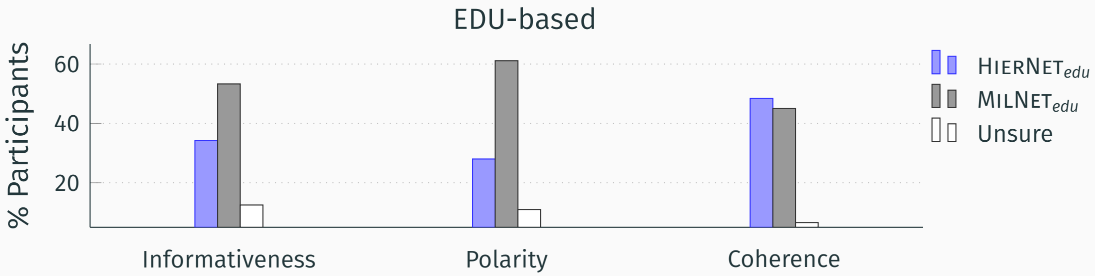
Coherence:

Summary 1 | Not sure | Summary 2

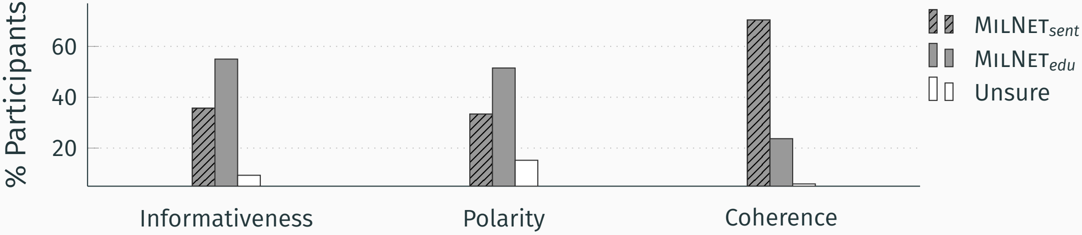
Is MilNet better than HierNet?



Is MilNet better than HierNet?



Are EDUs better than Sentences?



How does MilNet compare to Summarization Baselines