Embarrassingly Simple Unsupervised Aspect Extraction

Tolkens & Cranenburgh (2020)

HS Reproducibility in Natural Language Processing - SS 2021

Luisa Ribeiro da Silva

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Outline

- Aspect extraction
- Seq2Seq vs. Attention Mechanism
- Contrastive Attention based on RBF kernels
- Dataset and results
- Reproduction checklist
- Reproduction plan

Aspect Extraction

- An opinion always has a target. The target is often the aspect or topic to be extracted from a sentence.
- An aspect is the dimension on which an entity is evaluated.
- This task extracts aspects that have been evaluated.

The two things that really drew me to *vinyl* were the expense and the inconvenience.

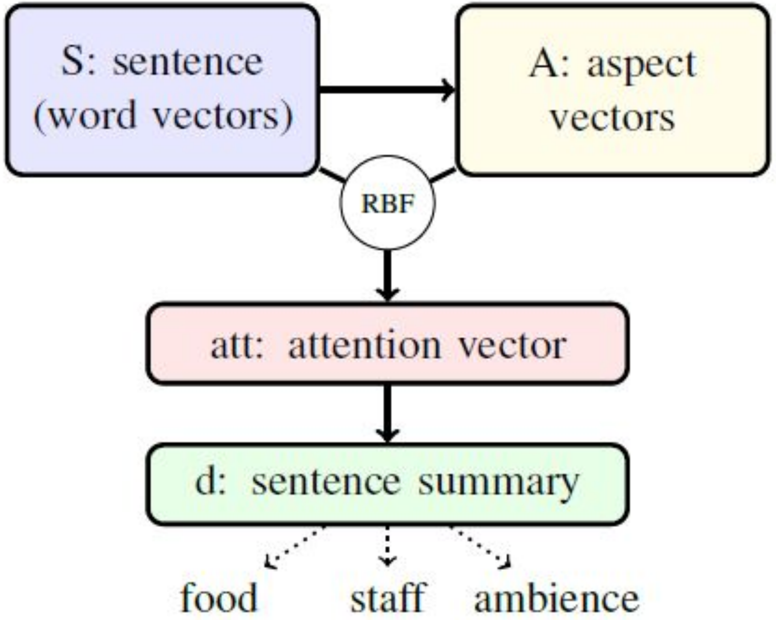
Aspect Extraction

- Aspects are domain-specific → Low transferability, lack of training data
- Most existing systems are supervised
- Autoencoders using attention mechanisms for aspect extraction have reached state of the art performance on a variety of datasets
  - They rely on unlabeled data to learn relevant patterns
  - Complex neural models with a large amount of parameters

The main idea of the paper

- A single-head attention mechanism, Contrastive Attention (CAt), based on Radial Basis Function (RBF) kernels is introduced.
- A simple unsupervised method for aspect extraction which only requires a POS tagger and in-domain word embeddings.
- Automatic assignment of aspect labels, while in previous work, labels are manually assigned to aspect clusters.

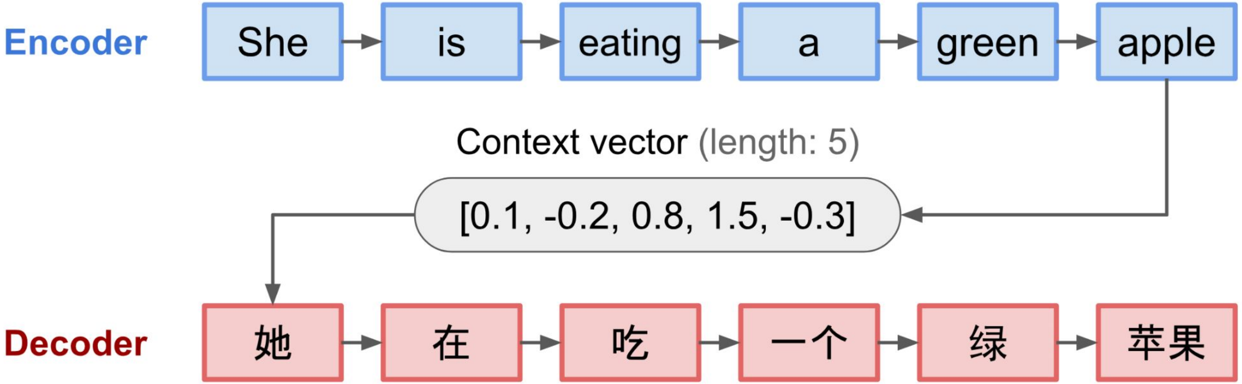
The main idea of the paper



Seq2Seq

- Seq2Seq aims at transforming an input sequence (source) into a new one (target) and both sequences can be of arbitrary lengths.
- The fixed-length context vector design is incapable of remembering long sentences.
- The attention mechanism was created (Bahdanau et al., 2015) to resolve this problem.

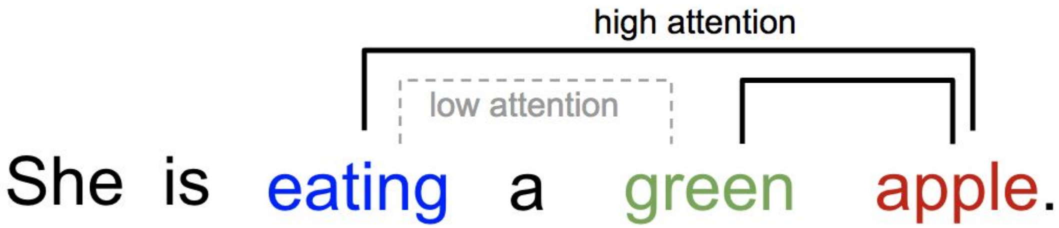
Seq2Seq



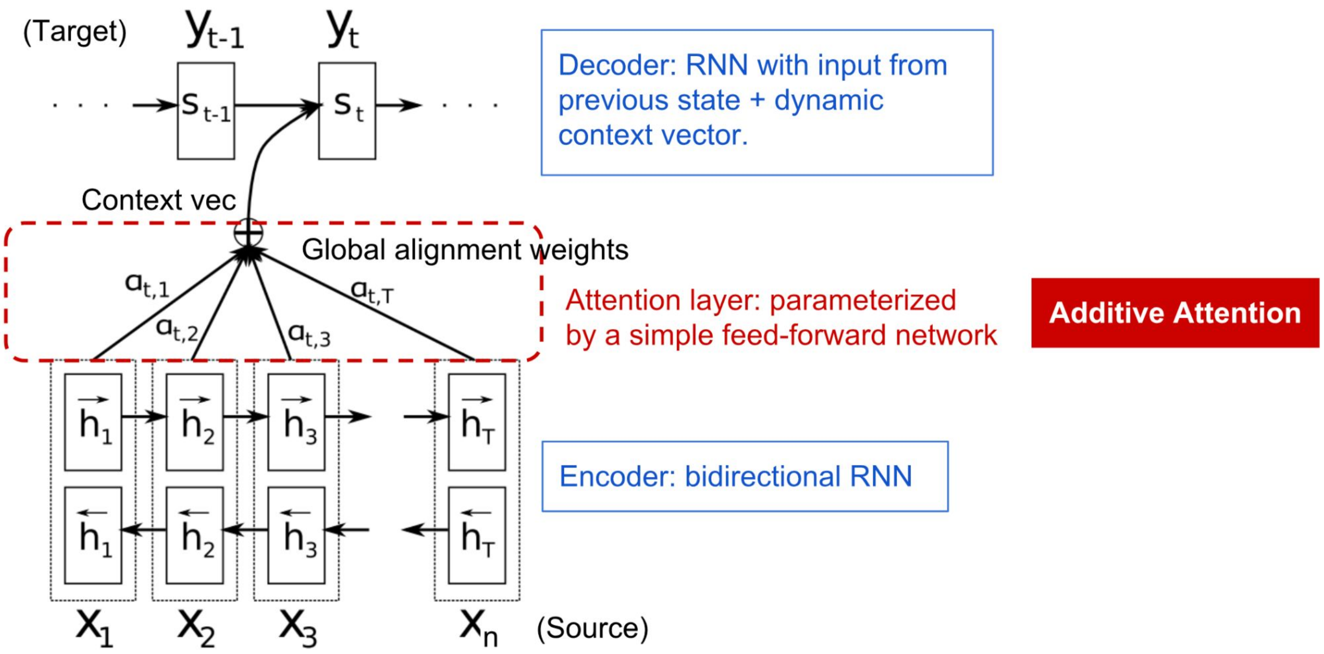
Encoder
She
is
eating
a
green
apple
Context vector (length: 5)
[0.1, -0.2, 0.8, 1.5, -0.3]
Decoder
她
在
吃
一个
绿
苹果

Attention Mechanism

- The attention mechanism was created to help memorize long source sentences in neural machine translation (NMT).
- Its main idea is to create weighted shortcuts between the context vector and the entire source input.



Attention Mechanism



Attention Mechanism for aspect extraction

(1)

$$\mathrm{att} = \mathrm{softmax}(aS)$$

probability distribution   aspect   matrix with sequence

(2)

$$d = \sum_i \mathrm{att}_i \, S_i$$

Contrastive Attention (Cat 🐈)

- Cat is a way of calculating attention that integrates a set of query vectors into a single attention distribution.
- The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other.

(3)
$$\mathrm{rbf}(x, y, \gamma) = \exp(-\gamma ||x - y||_2^2)$$

Contrastive Attention (Cat 🐈)

(4)

$$\text{att} = \frac{\sum_{a \in A} \text{rbf}(w, a, \gamma)}{\sum_{w \in S} \sum_{a \in A} \text{rbf}(w, a, \gamma)}$$

(5)

$$\hat{y} = \underset{c \in C}{\text{argmax}}(\cos(d, \vec{c}))$$

Datasets

- Restaurant reviews: Citysearch, SemEval 2014 and SemEval 2015
- All datasets have been annotated with one or more sentence-level labels, indicating the aspect expressed in that sentence
- Evaluation → Citysearch dataset
- SemEval 2014 and SemEval 2015 → development data
- Three labels: FOOD, SERVICE, and AMBIENCE

Evaluation and results

Method	P	R	F
Aspect: FOOD			
SERBM (2015)	89.1	85.4	87.2
ABAE (2017)	95.3	74.1	82.8
W2VLDA (2018)	**96.0**	69.0	81.0
AE-CSA (2019)	90.3	**92.6**	91.4
Mean	92.4	73.5	85.6
Attention	86.7	89.5	88.1
CAt 🐈	91.8	92.4	**92.1**

Evaluation and results

Aspect: STAFF			
SERBM (2015)	81.9	58.2	68.0
ABAE (2017)	80.2	72.8	75.7
W2VLDA (2018)	61.0	**86.0**	71.0
AE-CSA (2019)	92.6	75.6	77.3
Mean	55.8	85.7	67.5
Attention	74.4	69.3	71.8
CAt 🐈	**82.4**	75.6	**78.8**

Evaluation and results

Aspect: AMBIENCE			
SERBM (2015	80.5	59.2	68.2
ABAE (2017)	81.5	69.8	74.0
W2VLDA (2018)	55.0	75.0	64.0
AE-CSA (2019)	**91.4**	77.9	**77.0**
Mean	58.7	56.1	57.4
Attention	67.1	65.7	66.4
CAt 🐈	76.6	**80.1**	76.6

Evaluation and results

Method	**P**	**R**	**F**
SERBM (2015)	86.0	74.6	79.5
ABAE (2017)	**89.4**	73.0	79.6
W2VLDA (2018)	80.8	70.0	75.8
AE-CSA (2019)	85.6	86.0	85.8
Mean	78.9	76.9	77.2
Attention	80.5	80.7	80.6
CAt 🐈	86.5	**86.4**	**86.4**

Table 2: Weighted macro averages across all aspects on the test set of the Citysearch dataset.

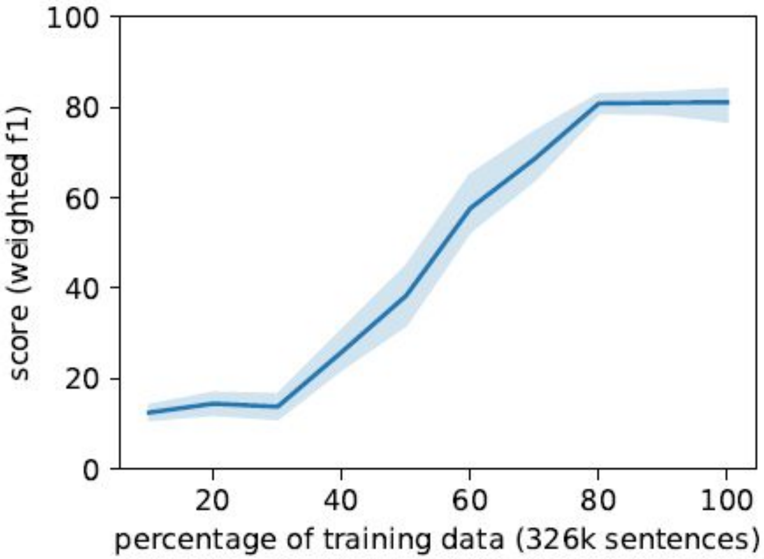
Analysis

An ablation study on the influence of POS tagging, in-domain word embeddings, and the amount of data on performance.

- Only selecting the *most frequent words* as aspects, regardless of their POS tag → F-score of **64.5** (Δ -21.9)
- Selecting nouns based on *adjective-noun co-occurrence* → F-score of **84.4** (Δ-2.2)
- Replacing the *in-domain word embeddings* trained on the training set with pretrained GloVe embeddings → F-score to **54.4** (Δ-32);

Analysis

How much in-domain data is required to achieve good performance?



Analysis

Manual categorization of error types

Phenomenon	Example
OOV	"I like the Somosas"
Data Sparsity	"great Dhal"
Homonymy	"Of course"
Verb > Noun	"Waited for food"
Discourse	"She didn't offer dessert"
Implicature	"No free drink"

Reproduction checklist

- Is the data and code used in the study available (publicly)?

Yes.

- If so, is it easy to obtain and run the code?

Yes.

https://github.com/clips/cat

Reproduction checklist

- Is there any indication of statistical assurances, e.g., significance tests,

presentation of (confidence) intervals?

Yes.

- Does the paper clearly describe the significance test applied?

Yes.

Reproduction checklist

- Is the dataset and potential splits clearly described?

Yes.

- Are there any pre-processing steps documented? Are they justified?

Yes.

- Does the paper include more general claims than data would allow?

No.

Reproduction checklist

- Are the baselines reported appropriate?

Yes.

- Were the baselines used (trained/tested) properly?

Looks like.

Reproduction checklist

- Can one ‘replicate’ the results and/or conclusions based on the descriptions in the paper?

Difficult, but yes.

- If code / data are provided, does it match the descriptions?

Yes.

Reproduction checklist

- Are the claims of practical/theoretical significance, novelty, being state of the art inflated?

Maybe.

- Does the paper look like an ‘unconditional’ supporter of a theory / hypothesis/method?

Maybe?

- Does the paper bias the presentation of the results (e.g., boldface numbers in tables)? If so, are they really relevant / justified?

Yes. Yes.

Reproduction plan

Step	Progress
Download data	✓
Inspect code	In progress
Attempt to achieve same results	TBD

Thank you!

References

Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Tulkens, S., & van Cranenburgh, A. (2020). Embarrassingly simple unsupervised aspect extraction. arXiv preprint arXiv:2004.13580.

Weng, L. (2018). Attention? Attention!. LilLog. https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html