Deep Learning Approaches for Opinion Mining on Conversational Social Media Texts

Master's Thesis

Sudeshna Dasgupta

Supervised By:
Prof. Dr. Georg Groh
Gerhard Hagerer, M.Sc.

Technische Universität München
Fakultät für Informatik
Arbeitsgruppe "Social Computing"

Munich, 29 October 2020

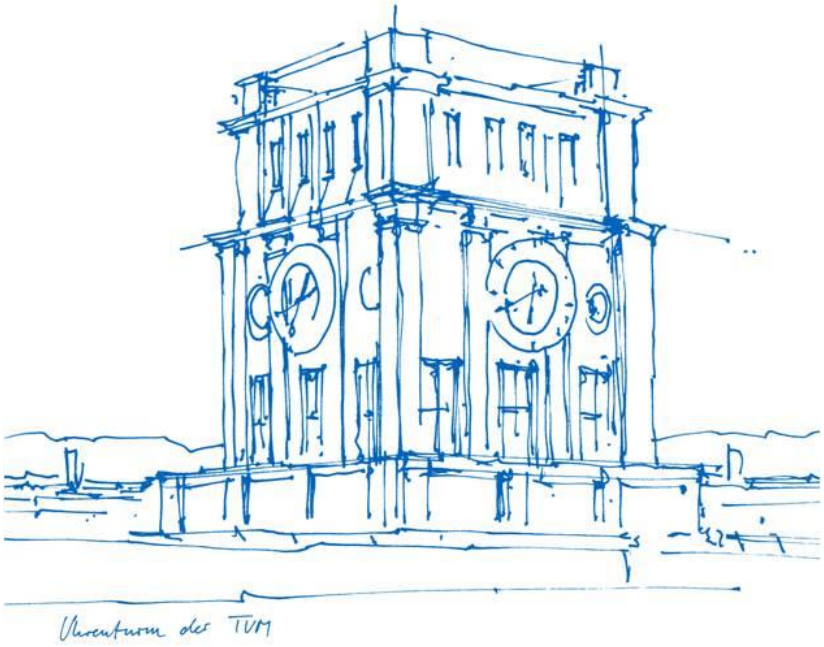


Table of Contents

Motivation

- Social networking sites are a medium to influence the views and decisions of their audience. Textual conversations on social media are essential resources for opinion mining.

"All have their share of Bars and Night life etc but for this I think LOCATION1 would be my favourite"

Opinion target : *LOCATION1*	Aspect: *Nightlife*

- In a document, features or attributes of the target expression are referred as aspects.

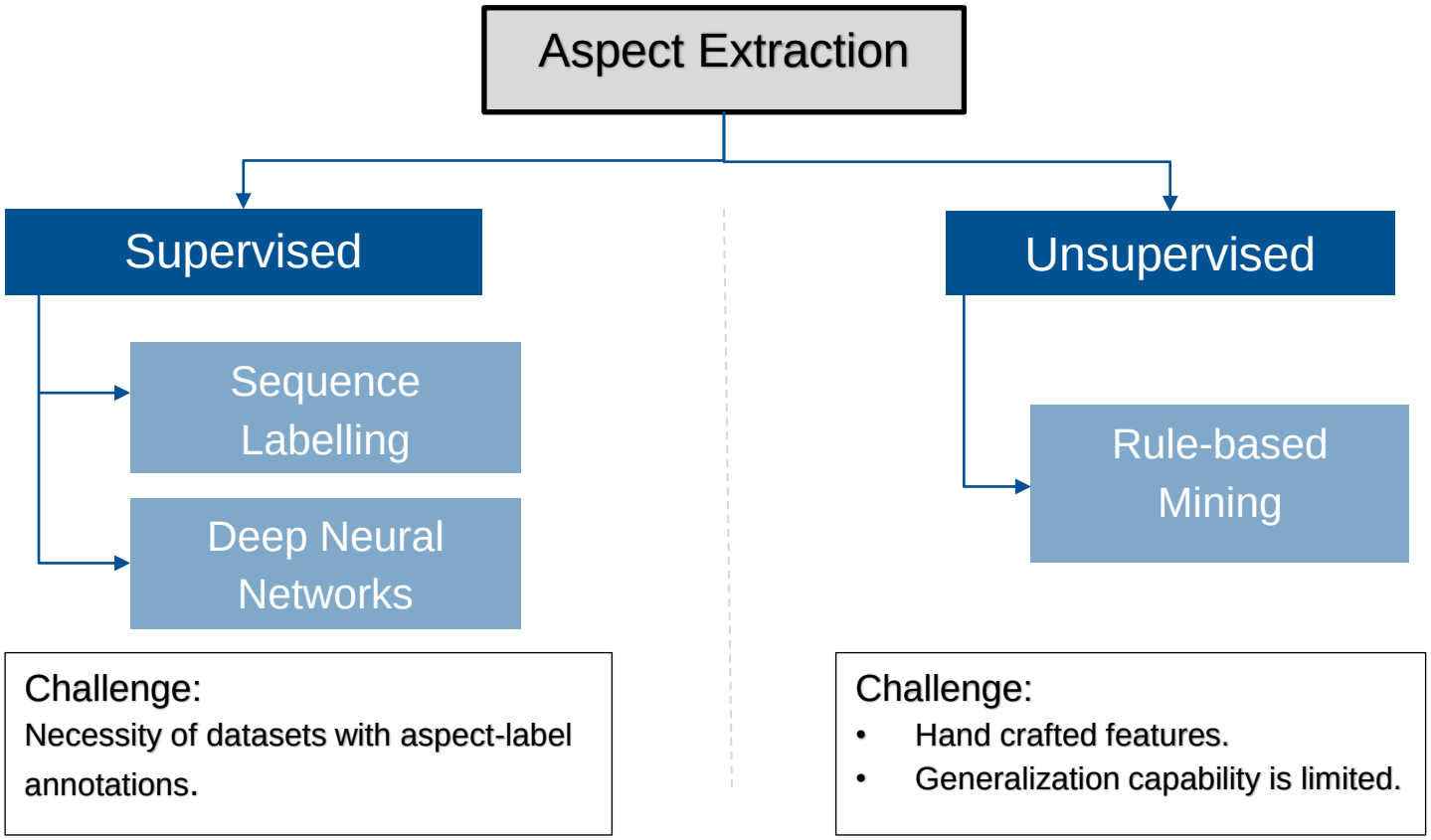
- Opinion mining extracts overall available sentiment in a document. Aspect-based sentiment analysis is a finer scrutiny to identify aspect-specific opinions.

All in all LOCATION1 one of the better suburbs of London though can be expensive

Opinion target : LOCATION1
Aspect: *general*, Sentiment: *positive*
Aspect: *price*, Sentiment: *negative*

- Aspect extraction is a key task in information retrieval.
  - Applications: aspect based sentiment analysis, sentence summarization.

Motivation



A more suitable approach: *Unsupervised deep neural networks*

Why?
- Automatic learning of semantic and syntactic information from model input.
- Scalable across varying datasets.

Research Objective

Objective:

Empirical study of unsupervised aspect extraction on Sentihood dataset.

Sentihood dataset:

- Reviews about the urban neighborhoods in London.
- 12 predefined aspect labels: *live', 'safety', 'price', 'quiet', 'dining', 'nightlife', 'transit-location', 'touristy', 'shopping', 'green-culture', 'multicultural', 'general'*

Data	Sentences	Labelled	Unlabelled
Train+Dev	3724	2526	1198
Test	1491	1003	488

Dataset	Aspect Labels	Reviews	Labelled Sentences
Sentihood	12	5215	3529
Restaurant	6	52574	3400
Beer	5	1,586,259	9245

Comparison of Sentihood and product review datasets

Research Objective

We want to know:

- Underlying challenges in extracting aspects from a dataset with
  - large number of aspect labels
  - explicit as well as implicit aspect labels
- If domain of dataset influences the choice of approach for aspect extraction
- if transfer learning can resolve the inadequacy of training data.

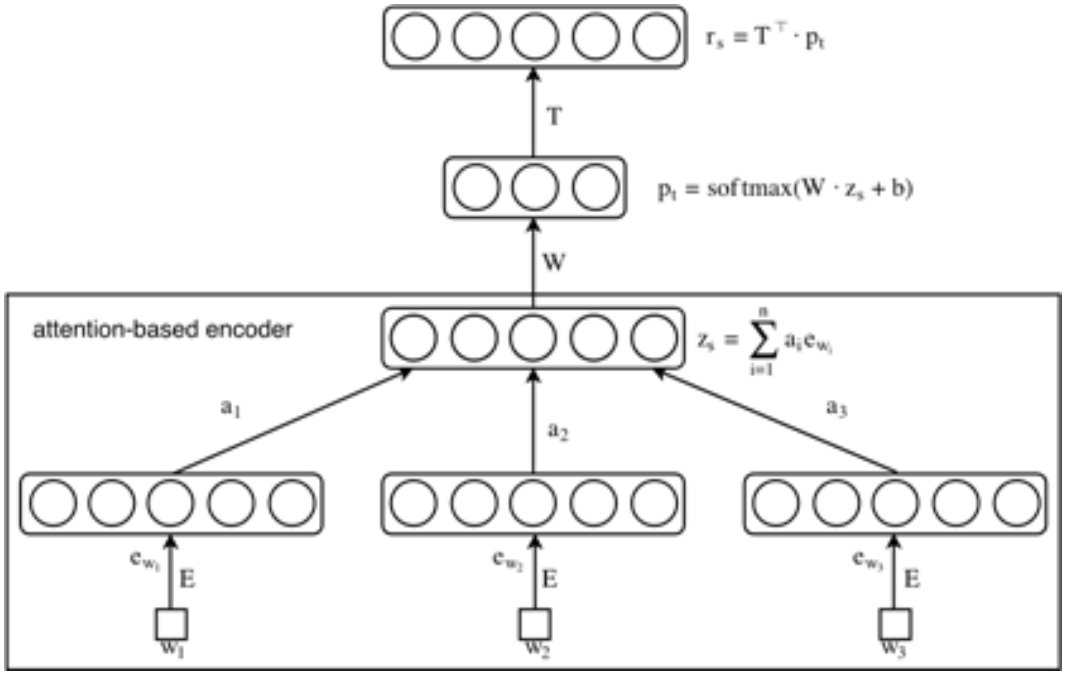
Approaches applied in this thesis:

- Attention Based Aspect Extraction (ABAE)
- Aspect Extraction with Sememe Attentions (AE-SA)
- Aspect Extraction via Context-enhanced Sememe Attentions (AE-CSA)

Methodology

Attention-based Aspect Extraction (ABAE)

- Model learns a set of aspect embeddings.
- Attention-based encoder extracts aspect-relevant words.
- Decoder reconstructs input sentence by a linear combination of aspect embeddings, to obtain reconstruction error or loss.
- Model output is a set of inferred aspects clusters.

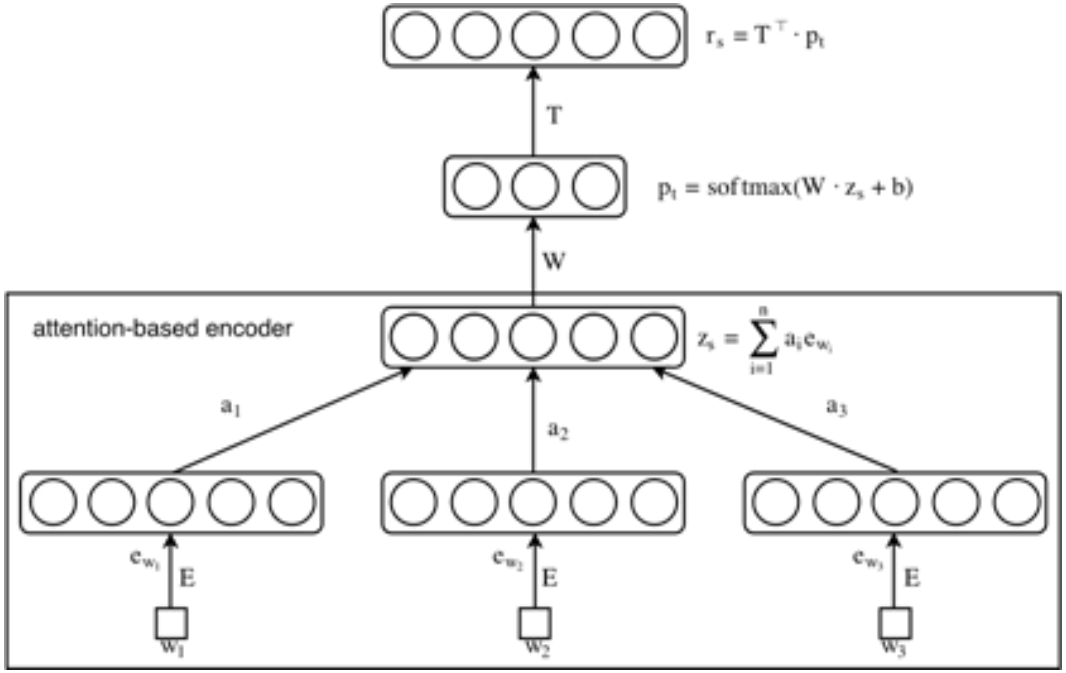


Input represented using word embeddings

Methodology

Attention-based Aspect Extraction (ABAE)

- Model learns a set of aspect embeddings.
- Attention-based encoder extracts aspect-relevant words.
- Decoder reconstructs input sentence by a linear combination of aspect embeddings, to obtain reconstruction error or loss.
- Model output is a set of inferred aspects clusters.



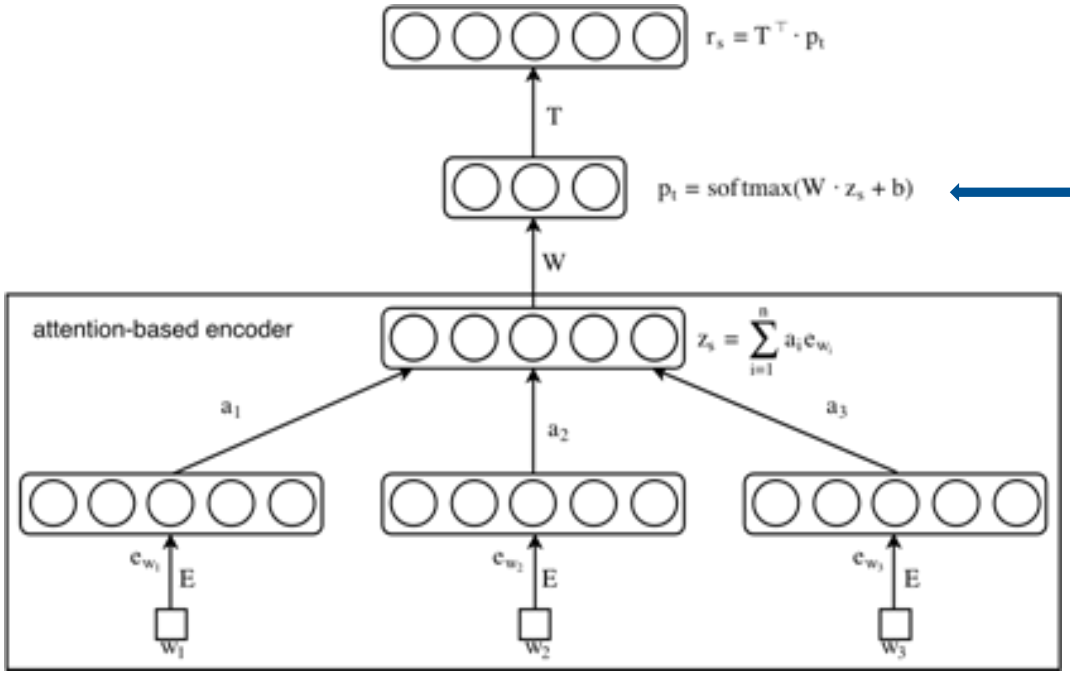
Non-aspect terms filtered using attention mechanism

(Attention weights are learnt in the context of the sentence)

Methodology

Attention-based Aspect Extraction (ABAE)

- Model learns a set of aspect embeddings.
- Attention-based encoder extracts aspect-relevant words.
- Decoder reconstructs input sentence by a linear combination of aspect embeddings, to obtain reconstruction error or loss.
- Model output is a set of inferred aspects clusters.

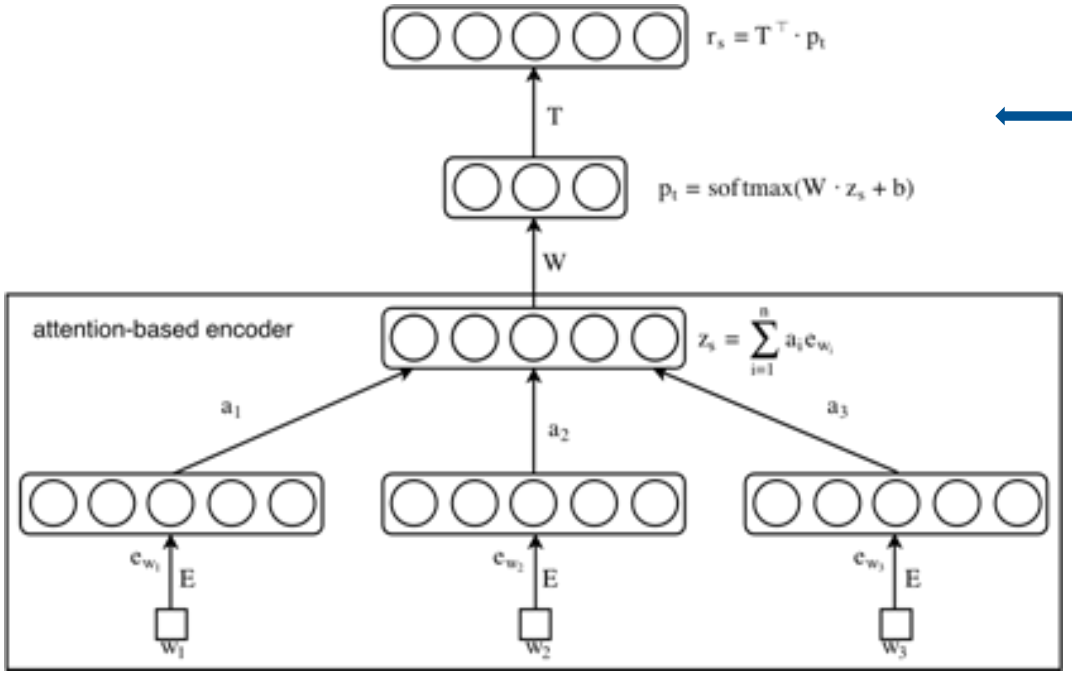


Distribution over inferred aspects for each sentence

Methodology

Attention-based Aspect Extraction (ABAE)

- Model learns a set of aspect embeddings.
- Attention-based encoder extracts aspect-relevant words.
- Decoder reconstructs input sentence by a linear combination of aspect embeddings, to obtain reconstruction error or loss.
- Model output is a set of inferred aspects clusters.

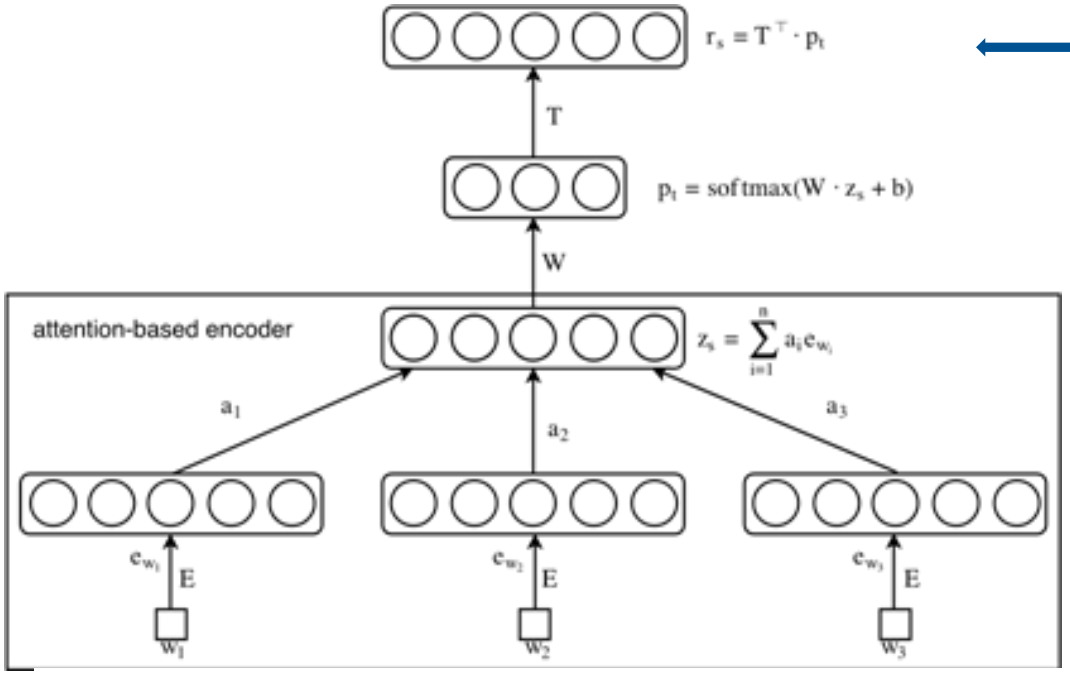


Aspect embedding matrix
learnt as a model parameter

Methodology

Attention-based Aspect Extraction (ABAE)

- Model learns a set of aspect embeddings.
- Attention-based encoder extracts aspect-relevant words.
- Decoder reconstructs input sentence by a linear combination of aspect embeddings, to obtain reconstruction error or loss.
- Model output is a set of inferred aspects clusters.



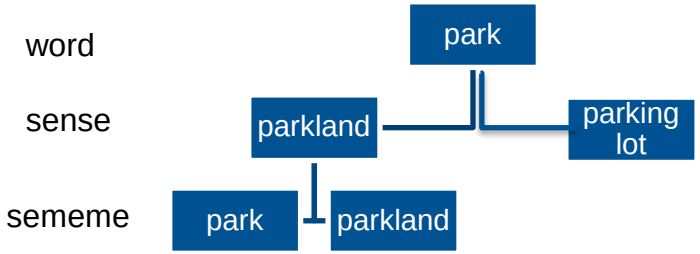
Input reconstruction error

Methodology

Aspect Extraction with Sememe Attentions (AE-SA) :

ABAE fails to extract less frequent aspects and implicit aspects.

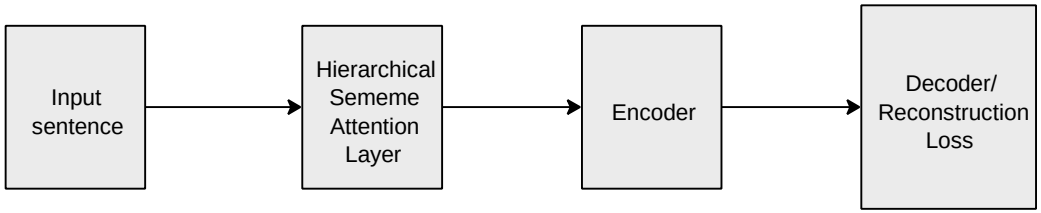
The Aspect Extraction with Sememe Attentions (AE-SA) model utilizes polysemy to learn word context.



Example of sense and sememe from WordNet

WordNet is a lexical database of English.

- A word sense is associated with sememes.
- A sememe is a minimum semantic unit in a human language.

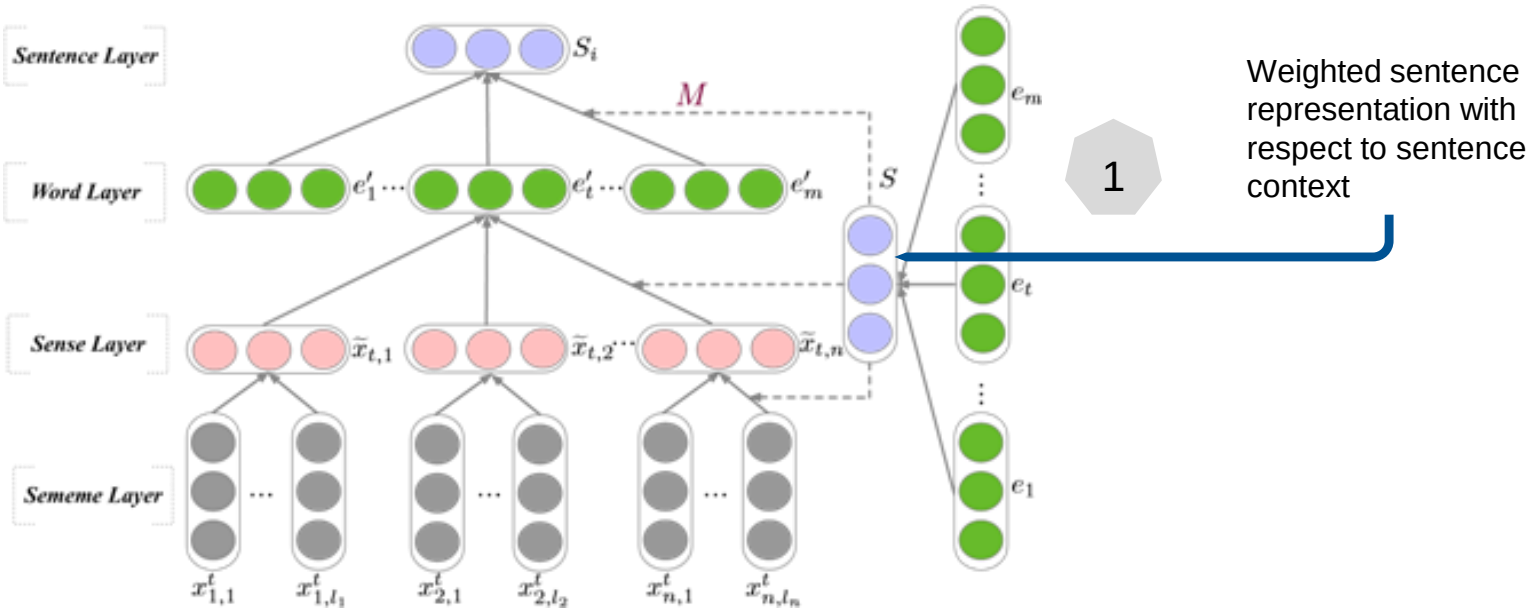


Aspect Extraction with Sememe Attentions (AE-SA)

- Hierarchical sememe attention chooses the most important senses and sememes for each word in the input sentence
- Model learns a set of aspect embeddings.
- Model output is a set of inferred aspects clusters.

Methodology

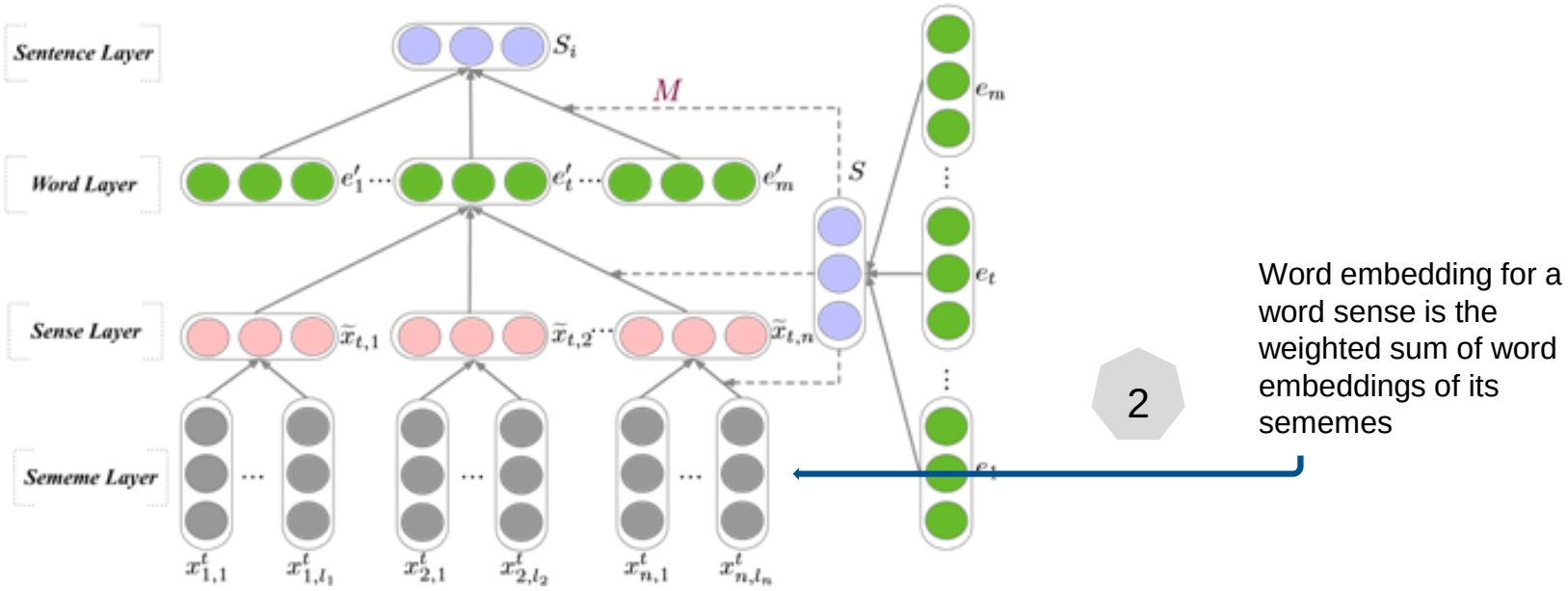
Aspect Extraction with Sememe Attentions (AE-SA)



Hierarchical sememe attention layer

Methodology

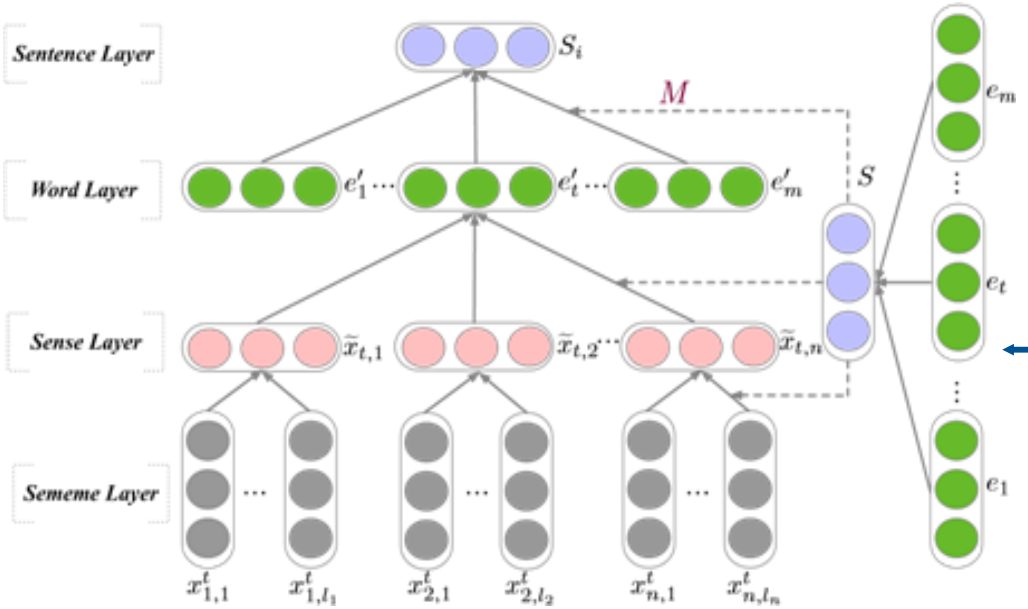
Aspect Extraction with Sememe Attentions (AE-SA)



Hierarchical sememe attention layer

Methodology

Aspect Extraction with Sememe Attentions (AE-SA)



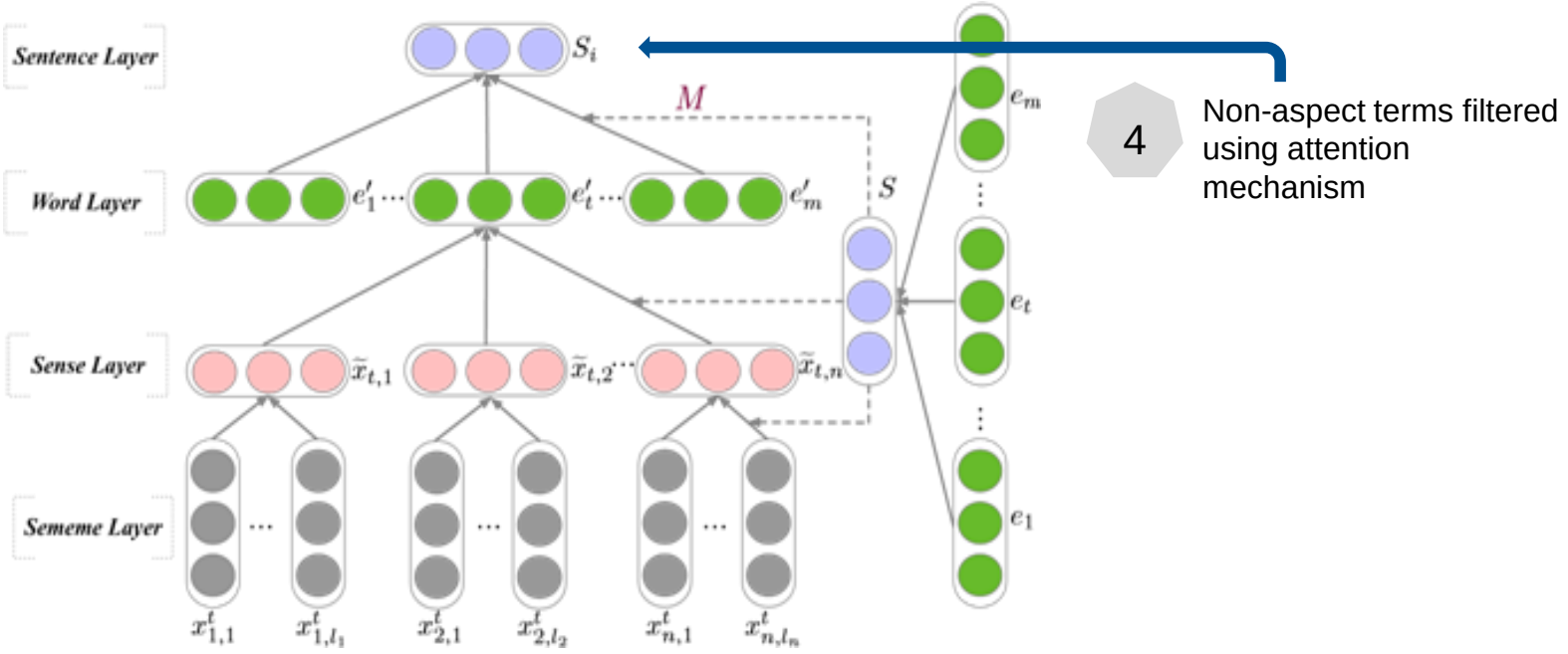
3

Word embedding for a word is the weighted sum of the embeddings of its word senses.

Hierarchical sememe attention layer

Methodology

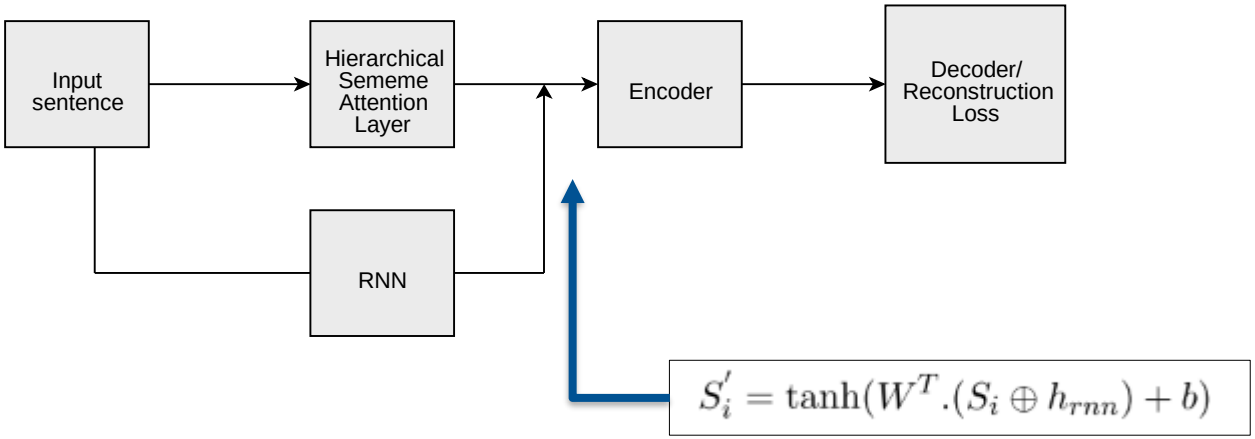
Aspect Extraction with Sememe Attentions (AE-SA)



Hierarchical sememe attention layer

Methodology

Aspect Extraction via Context-enhanced Sememe Attentions (AE-CSA)

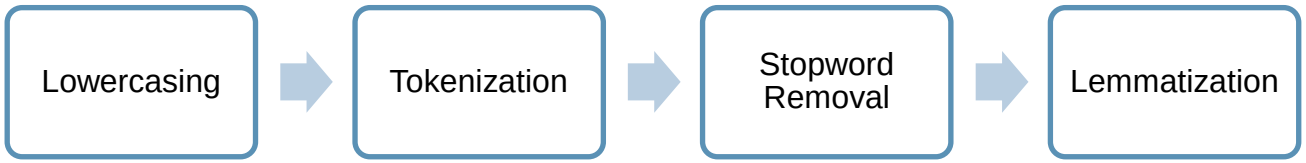


S_i': new sentence representation
S_i: sentence representation obtained from the hierarchical sememe attention layer
h_{rnn}: hidden representation of RNN
$\oplus$ denotes vector concatenation
$W \in R^{(d+d')\times d}$ and b are parameters learnt during training

- Model learns a set of aspect embeddings.
- Model output is a set of inferred aspects clusters.

Experiments

Data preprocessing:



Sample:

"LOCATION1 has been gentrified for over 40+ years, using London incomes to push property prices even higher than the national average, because of it's proximity to central London'

Preprocessed sample:

location1 gentrified 40 year using london income push property price even higher national average proximity central london

Experiments

Transfer Learning

- Experiments with ABAE, AE-SA and AE-CSA
- Models initialized with:
  - GloVe embeddings pretrained on Wikipedia and Gigaword dataset
  - Pretrained GloVe embeddings finetuned to Sentihood corpus.

Deep Learning

- Experiment performed with ABAE.
- Model initialized with GloVe embeddings trained on Sentihood corpus.

Number of inferred aspects is an experiment hyperparameter.
Experiments were performed as grid search over aspect size 10 to 100 with an interval of 10.

Experiments

Evaluation:

Quantitative:

Inferred aspects were evaluated with UMass coherence score and Word Embedding Topic Coherence score.

UMass coherence metric

$$C(z; S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} \log \frac{D_2(w_n^z, w_l^z) + 1}{D_1(w_l^z)}$$

$D_1(w)$ is the document frequency of word w,

$D_2(w1, w2)$ is the co-document frequency of the words w_1 and w_2

WETC metric

$$\mathrm{WETC}_{PW}(E) = \frac{1}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \langle E_{i,:}, E_{j,:} \rangle$$

$E \in^{N \times D}$, D is the embedding dimension, and $\langle ., . \rangle$ denotes matrix product.

Qualitative:

Coherent topics should be comprehensible by a human observer.

Cluster map:
<inferred aspects> : <predefined aspect labels>

Example:
0: 'general', 1: 'transit-location', 2: 'price', 3: 'touristy', 4: 'multicultural', 5: 'shopping'

Trained model was evaluated for the task of aspect-based sentence classification:

1. Inferred aspects mapped to predefined aspect labels as cluster maps.
2. Aspect label for an evaluation data-sample derived using latent vector parameter and cluster map.

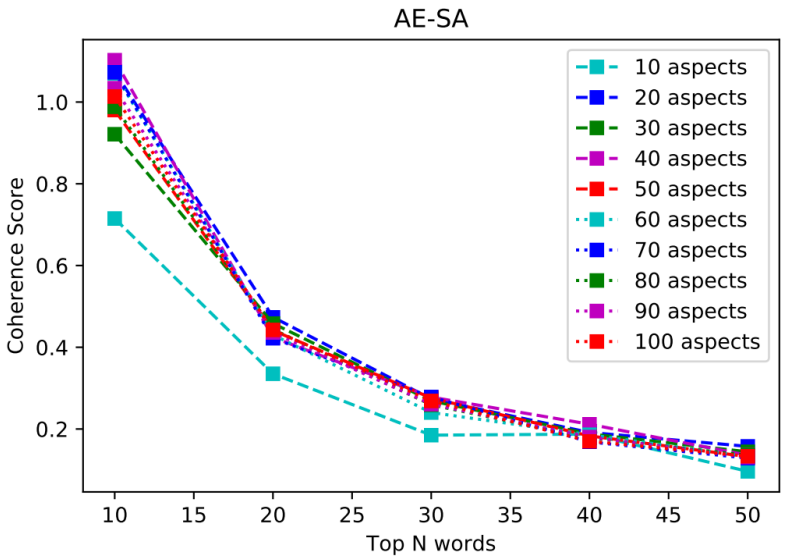
Experiments

Inferred Aspects	Representative Words	Predefined Aspect Label
Real estate worth	dollar, worth, equates, upwards, priced, rocketed, estimate	Price
City administration	prime, party, lebanese, movement, thai	Multicultural
Movement zones	parked, embankment, flooded, nearby, tunnel, parking, surrounded	Transit-Location
Royal family	lord, kingdom, king, abolished, viii, elizabeth, empire	Touristy
Brands	company, ikea, tesco, giant, kfc, ebay, bought, sale, buy, chain	Shopping
Entertainment	circus, cabaret, entertainer, comedy, concert, nightclub, musical	Nightlife

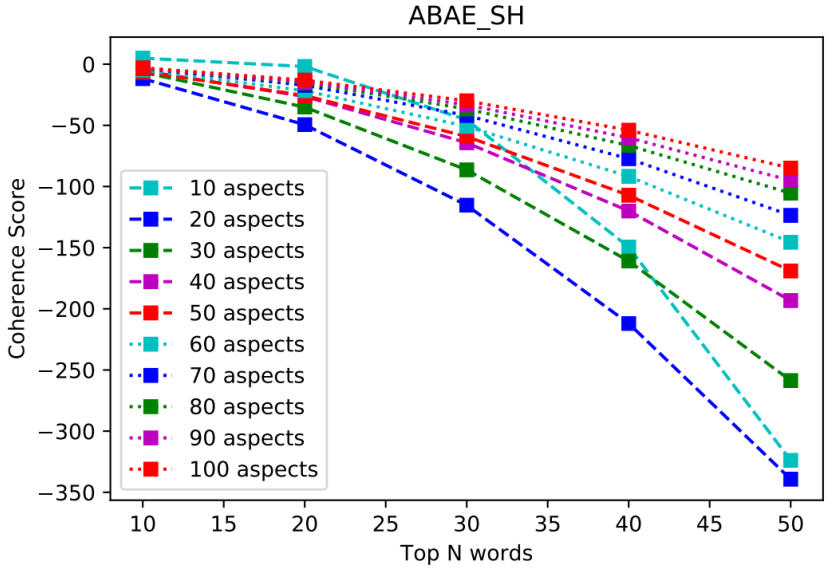
List of inferred aspects (left), subset of representative words used for cluster topic decision(middle), and the mapped gold standard label(right). The predefined aspect labels for Sentihood dataset are ’live’, ’safety’, ’price’, ’quiet’, ’dining’, ’nightlife’, ’transit-location’, ’touristy’, ’shopping’, ’green-culture’, ’multicultural’, ’general’

Results

Quantitative Evaluation



Coherence measured with WETC score



Coherence measured with UMass score

Results

Quantitative Evaluation

Aspect	ABAE	Kmeans	SA	SACSA	NMF	ABAE_SH	Sum of AUC
10	-148.874.201	-151.909.810	-145.356.715	-142.837.322	-104.959.758	-3.563.630	**-697.501.436**
20	-149.321.800	-154.448.616	-144.587.619	-146.663.804	-115.294.879	-5.523.148	-715.839.866
30	-150.333.167	-154.371.669	-146.056.441	-148.641.333	-120.255.853	-4.151.290	-723.809.753
40	-150.279.860	-155.410.926	-147.670.604	-149.235.099	-124.179.916	-3.099.355	-729.875.760
50	-150.610.692	-155.123.895	-148.138.198	-149.853.505	-127.195.119	-2.794.886	-733.716.295
60	-150.894.758	-155.509.730	-148.570.454	-149.694.998	-128.239.363	-2.394.719	-735.304.022
70	-151.183.497	-155.250.056	-149.043.904	-150.081.084	-129.863.126	-1.999.210	-737.420.877
80	-151.034.824	-155.081.512	-149.003.840	-150.061.577	-131.019.218	-1.728.078	-737.929.049
90	-151.236.573	-155.231.713	-149.284.444	-150.679.502	-132.296.331	-1.569.474	-740.298.037
100	-151.183.658	-155.316.483	-149.268.989	-150.320.469	-132.354.901	-1.409.682	-739.854.182
Sum of AUC	-1.504.953.030	-1.547.654.410	-1.476.981.208	-1.488.068.693	-1.245.658.464	**-28.233.472**	

Area under curve (AUC) table for cluster coherence of inferred aspects, calculated using the UMass Coherence score. A lower value denotes higher coherence. **ABAE-SH** achieves the highest aggregate AUC.

Results

Quantitative Evaluation

Aspect	ABAE	Kmeans	AE-SA	AE-CSA	NMF	ABAE_SH	Sum of AUC
10	9.535	9.264	11.133	11.816	15.700	16.022	73.470
20	13.414	8.753	15.550	13.230	11.642	14.673	**77.262**
30	13.467	9.488	14.483	12.832	12.242	11.639	74.151
40	13.553	9.252	15.332	13.073	10.342	9.770	71.322
50	12.394	9.583	14.565	11.503	10.087	7.968	66.100
60	12.493	9.596	14.580	11.797	9.203	7.589	65.258
70	12.822	9.794	14.698	12.566	9.132	7.742	66.754
80	12.573	9.456	14.434	12.590	8.651	6.978	64.682
90	12.483	10.325	14.599	11.089	8.820	6.598	63.914
100	11.490	9.752	14.546	10.821	9.136	6.619	62.364
Sum of AUC	124.224	95.263	**143.920**	121.317	104.955	95.598	

AUC table for cluster coherence calculated using WETC score. A higher value denotes higher coherence. **AE-SA** achieves the highest aggregate AUC.

Results

Additional implicit aspects:

- *outskirts, city administration*
- Mapped as *general* due to lack of suitable predefined aspect labels.

Coherent cluster which is mapped to 'general' label

"outskirt|0.58713293 suburb|0.5857716 bayswater|0.55934393 adjoining|0.5543636 situated|0.53446364 picturesque|0.5337991 neighbourhood|0.5108886 rosedale|0.50137115 overlooking|0.4937445 wooded|0.49081147"

Qualitative analysis of inferred aspects:

Table with number of coherent and relevant clusters,
number of coherent and not relevant clusters,
number of incoherent clusters obtained from experiments for aspect size 30.

A cluster is relevant if its topic is related to Sentihood dataset.

		Coherent Clusters		
Model	Aspect Count	Relevant	Not Relevant	Incoherent Clusters
AE-CSA	30	6	11	13
AE-SA	30	4	7	19
ABAE_SH	30	7	2	21
ABAE	30	6	7	17
NMF	30	2	0	28
K-means	30	3	2	25

Results

Dataset constraint:

- Incoherent clusters consist of nouns, alphanumericals, adverbs, adjectives.
- The words were not filtered as:
  - Adverbs and adjectives are used to express implicit aspects
  - Alphanumericals suggest street and highway names, related to label *transit-location*
  - Nouns are suggestive of labels *multicultural* and *touristy*.
  - Abbreviations words like *sqft* are suggestive of label *price*

Multicultural

"ortega|0.50206417 colombian |0.42718196 barrrre|0.3930954 da|0.38973352 barrio|0.38580525 spanish |0.37710303 brazilian |0.37064406 gabriel|0.36198694 esp|0.35027727 angel|0.335748 newell|0.32469124 pietra|0.3143616 la|0.31377035 villa|0.304713 portuguese|0.29728857 spain|"

Conclusion

- Domain knowledge is a key constituent in the task of aspect extraction.
- Knowledge about word meanings aids the model to learn better aspect embeddings and improves the coherence of clusters.
- Definition of coherent inferred aspect is subjective to the availability of an appropriate predefined aspect label.
- Transfer learning is not useful when the source corpus and the target corpus do not belong to related domain.

Future Work

- Experiments with all models using word embeddings trained on the Sentihood corpus.
- Using a set of seed words as a weak supervision for aspect term extraction.
- Use of parts of speech to improve the choice of word sense from WordNet.

References

1. Angelidis, Stefanos and Mirella Lapata (2018). „Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised“. In: arXiv preprint arXiv:1808.08858 (cit. on pp. 8, 50)
2. He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier (2017). „An Unsupervised Neural Attention Model for Aspect Extraction“. In: Proceedings of the 55th Annual Meeting ofthe Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, pp. 388–397 (cit. on pp. 4, 27–29, 38, 41–43, 45, 53, 57, 62).
3. Luo, Ling, Xiang Ao, Yan Song, et al. (2019). „Unsupervised Neural Aspect Extraction with Sememes“. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization, pp. 5123–5129 (cit. on pp. 4, 30, 33, 58, 59)
4. Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011). „Optimizing Semantic Coherence in Topic Models“. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP ’11. Edinburgh, United Kingdom: Association for Computational Linguistics, 262–272 (cit. on pp. 49, 60).
5. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). „GloVe: Global Vectors for Word Representation“. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (cit. on pp. 13, 42, 46)
6. Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b). „Efficient estimation of word representations in vector space“. In: arXiv preprint arXiv:1301.3781 (cit. on p. 12).
7. Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013a). „Distributed representations of words and phrases and their compositionality“. In: Advances in neural information processing systems, pp. 3111–3119 (cit. on p. 12).
8. Ding, Ran, Ramesh Nallapati, and Bing Xiang (2018). „Coherence-Aware Neural Topic Modeling“. In: Proceedings ofthe 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 830–836 (cit. on p. 49).
9. Dingwall, Nicholas and Christopher Potts (2018). „Mittens: an Extension of GloVe for Learning Domain-Specialized Representations“. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, pp. 212–217 (cit. on p. 44).
10. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). „GloVe: Global Vectors for Word Representation“. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (cit. on pp. 13, 42, 46)

Questions

Thank you!

Questions?

Appendix

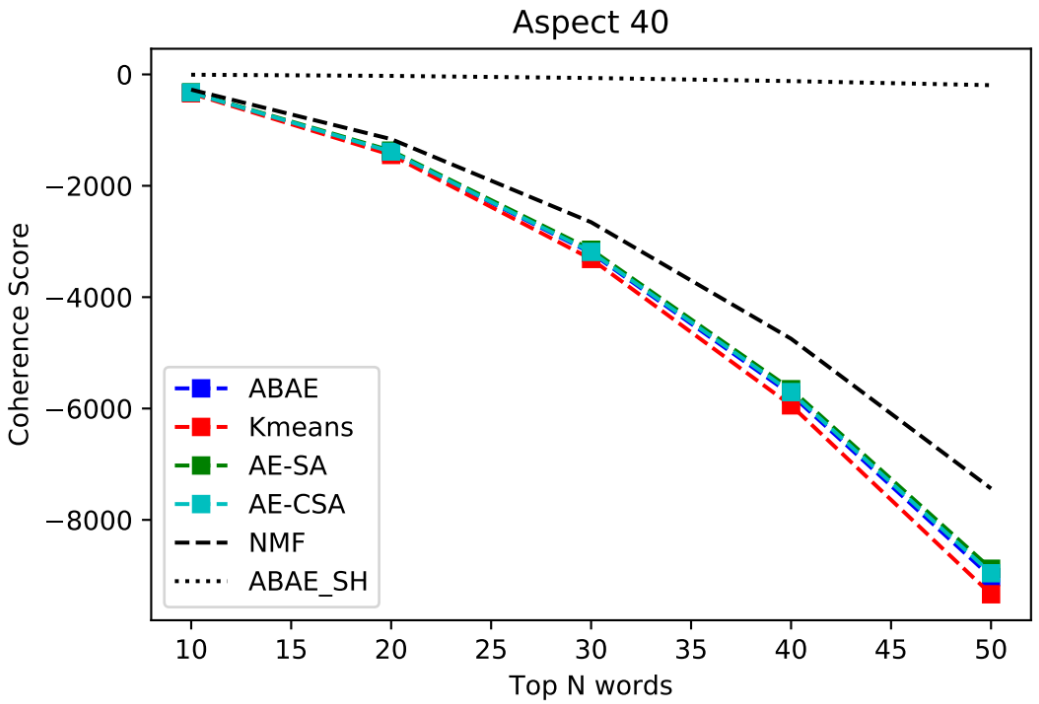
Results

Model	Accuracy	Micro-F1	Aspect Size
ABAE	0.349	0.349	37
ABAE_HP	0.411	0.411	37
ABAE_FT	0.266	0.266	37
ABAE_FT_HP	0.383	0.383	37
ABAE_SH	0.325	0.325	40
AE-SA	0.350	0.350	40
AE-CSA	0.341	0.341	50

Highest accuracy, highest micro F1 score and corresponding aspect size
for all models for the task of aspect extraction on Sentihood dataset.
ABAE_FT denotes ABAE model initialized with pre-trained GloVe embeddings finetuned to Sentihood dataset.
ABAE_HP and ABAE_FT_HP are the models ABAE and ABAE_FT with low precision classes removed.

Results

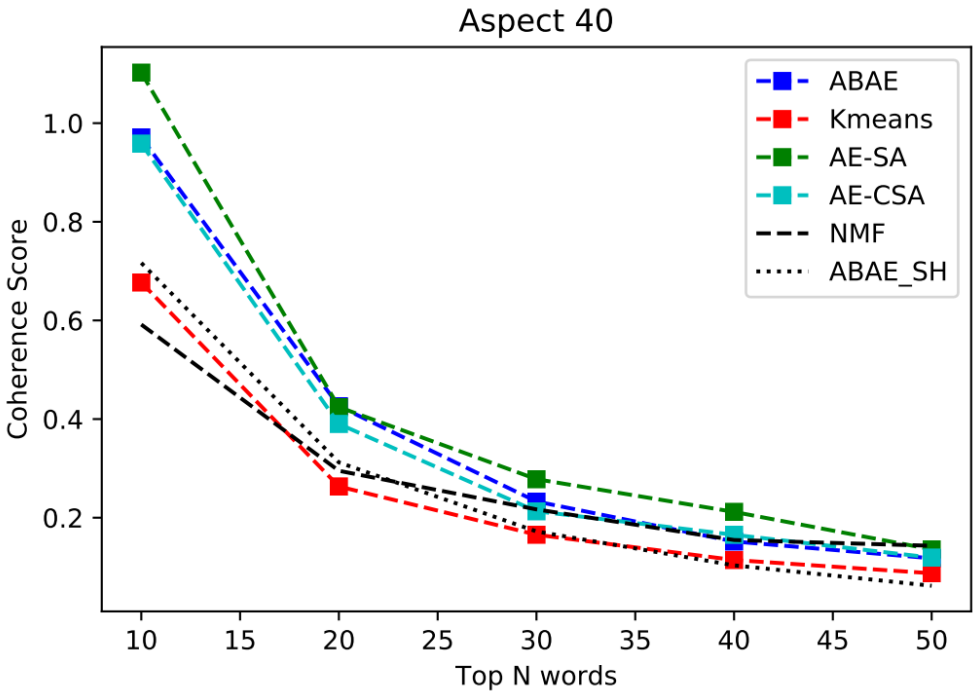
Quantitative Evaluation



UMass coherence for all models.
X axis represents number of top representative words in each cluster. ABAE_SH has highest coherence score.

Results

Quantitative Evaluation



WETC coherence score for all models.
X axis represents number of top representative words
in each cluster. AE-SA has highest coherence score.

Baseline models

- Kmeans
  - K-means is applied on the word embedding matrix of ABAE model to obtain cluster centroids.
  - The centroids are used as aspect embeddings to obtain inferred aspects.

- NMF
  - Corpus is converted to term-document matrix and factorized into a term-topic and a topic-document matrix.

Evaluation

Explicit aspect labels:
- *price* , *multicultural*
- Interpretable by human observer.

Implicit aspect labels:
- *quiet, live*.
- Difficult to interpret.

Price

" million |0.6206217 compared|0.5245044 per|0.52381444 total|0.49247202 income|0.47724462 excluding|0.46497005 sq|0.46454373 population|0.4597404 average|0.45596352 worth |0.436171 share|0.42922455 respectively|0.41132182 increase|0.4070933 increased|0.3890413 rate |0.38882133 roughly|0.3796498 housing|0.36251682 basis|0.36064082 lowest|0.3606255"

Multicultural

"ortega|0.50206417 colombian |0.42718196 barrrre|0.3930954 da|0.38973352 barrio|0.38580525 spanish |0.37710303 brazilian |0.37064406 gabriel|0.36198694 esp|0.35027727 angel|0.335748 newell|0.32469124 pietra|0.3143616 la|0.31377035 villa|0.304713 portuguese|0.29728857 spain|"