Capsule Based Aspect Extraction

Kirill Krasikov

May 2020

Abstract

In the world of data science, more and more data is required to train models. Getting labeled datasets is a very expensive and lengthy process. In this regard, in the problems of topic modeling, aspect extraction models from unstructured text are gaining popularity. These models allow, with minimal effort to annotate aspects to structure the text into a set of topics or intentions. In this paper, a comparison is made of a SotA model based on attention and the model in which the attention mechanism is replaced by a capsule network with dynamic routing. Codes for paper: `https://github.com/KirillKrasikov/TopicModelingWithCapsNet`.

1 Introduction

Aspect extraction is an important and challenging task in aspect-based sentiment analysis. For example, in the sentence "The beef was tender and melted in my mouth", the aspect term is "beef". Two sub-tasks are performed in aspects extraction: (1) extracting all aspects terms(e.g., "beef") from a review corpus, (2) clustering aspect terms with similar meaning into categories where each category represents a single aspect(e.g., cluster "beef", "pork", "pasta", and "tomato" into one aspect *food*) [He et al., 2017]. In 2017 year, Unsupervised Aspect Extraction [He et al., 2017] became the dominant unsupervised approach for aspect extraction. In that work word embeddings map words with the same context to nearby points in the embedding space [Mikolov et al., 2017]. Then attention mechanism [Bahdanau et al., 2015] filter the word embeddings within a sentence and filtered words are used to construct aspect embeddings. The training process for aspect embeddings is analogous with autoencoders, where dimension reduction is used to extract the common factors among embedded sentences and reconstruct each sentence through a linear combination of aspect embeddings. The attention mechanism deemphasizes words that are not part of any aspects, allowing the model to focus on aspect words. That model is called *Attention-based Aspect Extraction* (ABAE). In the same year "Google brain" suggested approach called *Dynamic Routing Between Capsules* [Sabour et al., 2017]. A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part. The length of the activity vector is used to represent the probability

that the entity exists and its orientation is used to represent the instantiation parameters. Active capsules at one level make predictions, via transformation n-dimension arrays, for the instantiation parameters of higher-level capsules. When multiple predictions mutually adjusted, a higher level capsule becomes active. To best performance, iterative routing-by-agreement mechanism is used. A lower-level capsule prefers to send its output to higher level capsules whose activity vectors have a big scalar product with the prediction coming from the lower-level capsule [Sabour et al., 2017]. In this work an ABAE model is implemented using pytorch, then attention mechanism is changed to capsule network and this model was evaluated on Citysearch corpus used by previous works [Ganu et al., 2009, Brody and Elhadad, 2010, Zhao et al., 2010, He et al., 2017].

2 Related Work

The problem of aspect extraction was well studied in the past decade. Initially, methods were mainly based on manually defined rules. [Hu and Liu, 2004] proposed to extract different product features through finding frequent nouns and noun phrases. They also extracted opinion terms by finding the synonyms and antonyms of opinion seed words through WordNet. Following this, a number of methods have been proposed based on frequent item mining and dependency information to extract product aspects [Zhuang et al., 2006, Somasundaran and Wiebe, 2009, Qiu et al., 2011]. These models heavily depend on predefined rules which work well only when the aspect terms are restricted to a small group of nouns. Supervised learning approaches generally model aspect extraction as a standard sequence labeling problem. [Jin and Ho, 2009, Li et al., 2010] proposed to use hidden Markov models(HMM) and conditional random fields(CRF) respectively with a set of manually-extracted features. More recently, different neural models [Yin et al., 2016, Wang et al., 2016] were proposed to automatically learn features for CRF-based aspect extraction. Rule-based models are usually not refined enough to categorize the extracted aspect terms. On the other hand, supervised learning requires large amounts of labeled data for training purposes.

Unsupervised approaches, especially topic models, have been proposed subsequently to avoid reliance on labeled data. Generally, the outputs of those models are word distributions or rankings for each aspect. Aspects are naturally obtained without separately performed extraction and categorization. Most existing works [Brody and Elhadad, 2010, Zhao et al., 2010, Mukherjee and Liu, 2012, Chen et al., 2014] are based on variants and extensions of LDA [Blei et al., 2003]. Recently, [Wang et al., 2015] relies on a substantial amount of prior knowledge such as part-of-speech(POS) tagging and sentiment lexicons. Biterm topic models (BTM) that generate co-occurring word pairs was proposed in [Yan et al., 2013].

Attention models [Mnih et al., 2014] have recently gained popularity in training neural networks and have been applied to various natural language processing tasks, including machine translation [Bahdanau et al., 2015], sentence

summarization [Rush et al., 2015], sentiment classification [**?**] and question answering [**?**]. Rather than using all available information, attention mechanism aims to focus on the most pertinent information for a task. [He et al., 2017] applies attention to an unsupervised neural model and demonstrate its effectiveness for aspect extraction.

[Zhao et al., 2015] explore capsule networks with dynamic routing process [Sabour et al., 2017] for supervised text classification. In this work capsule network with dynamic routing applied to unsupervised neural model and results are compared with others aspect extraction models.

3 Model Description

The ultimate goal of Capsule Based Aspect Extraction (CBAE) is to learn a set of aspect embeddings, where each aspect can be interpreted by looking at the nearest words (representative words) in the embedding space. Each word w in vocabulary with a feature vector $e_w \in^d$ [He et al., 2017]. Word embeddings are used to map words that often co-occur in a context to points that are close by in the embedding space [Mikolov et al., 2017]. The feature vectors associated with the words correspond to the rows of a word embedding matrix $E \in R^{V \times d}$, where V is the vocabulary size. Model tries to learn embeddings of aspects, where aspects share the same embedding space with words. This requires an aspect embedding matrix $T \in R^{K \times d}$, where K - the number of aspects (K is much smaller than V). The aspect embeddings are used to approximate the aspect words in the vocabulary, where aspect words are filtered through the capsule network with dynamic routing[1]. Each input sample to CBAE is a list of indexes for words in a review sentence. First, the model filter away non-aspect words by down-weighting them using a capsule network and then constructs a sentence embedding z_s from weighted word embeddings. Then, the model tries to reconstruct the sentence embedding as a linear combination of aspect embedding from T. This process of dimension reduction and reconstruction, where CBAE aims to transform sentence embeddings of the filtered sentences (z_s) into their reconstruction (r_s) with the least possible amount of distortion, preserves most of the information of the aspect words in the K embedded aspects. This process is described in detail in [He et al., 2017].

3.1 Sentence Embedding and Capsule Network with dynamic routing

Capsule network, depicted in Fig. 1 get sentence embedding vector to input. It consists of four layers: n-gram convolutional layer, primary capsule layer, secondary capsule layer and fully connected capsule layer.

N-gram convolutional layer is a standard convolutional layer which extracts n-gram features at different positions of a sentence through various convolu-

[1]This is the key difference from ABAE [He et al., 2017] , in which the mechanism of attention is used for filtering words.

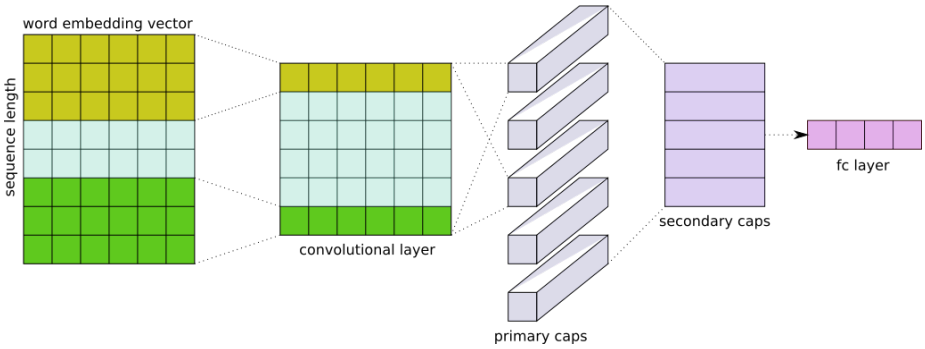

Figure 1: The Architecture of capsule network for aspect extraction (CBAE encoder part)

tional filters[Zhao et al., 2015] and activated by ELU function. Primary capsule layer is the first capsule layer in which the capsules replace the scalar-output feature detectors of CNNs with vector-output capsules to preserve the instantiated parameters such as the local order of words and semantic representation of words. Capsule network tries to address the representational limitation and exponential inefficiencies of convolutions with transformation matrices [Sabour et al., 2017]. It allows the networks to automatically learn child-parent relationships constituting viewpoint invariant knowledge that automatically generalizes to novel viewpoints. Dynamic routing is needed to construct a non-linear map in an iterative manner ensuring that the output of each capsule gets sent to an appropriate parent in the subsequent layer. For each potential parent, the capsule network can increase or decrease the connection strength by dynamic routing, which is more effective than the primitive routing strategies such as max-pooling in CNN that essentially detects whether a feature is present in any position of the text, but loses spatial information about the feature. Secondary capsule layer multiplies transformation matrices to learn child-parent relationships followed by routing agreement to produce parent capsules in the above. Then the capsules from secondary capsule layer gets flattened into fully connected final capsule, which is then activated by *tanh* function and normalized by *softmax*.

4 Dataset

The model have been evaluated on **Citysearch corpus**. This is restaurant review corpus widely used by previous works [Ganu et al., 2009, Brody and Elhadad, 2010, Zhao et al., 2010], which contains over 50,000 restaurant reviews from Citysearch New York. [Ganu et al., 2009] provided a subset of 3,400 sentences from the corpus with manually labeled aspects. These annotated sentences are used for evaluation of aspect identification. There are six manually defined aspect labels: *food*, *staff*, *ambience*, *price*, *anecdotes* and *miscellaneous*.

Corpus preprocessing is described in [He et al., 2017].

5 Experiments

5.1 Metrics

The model was evaluated by precision, recall and F1 scores. Classes of aspects were tagged manually.

5.2 Experiment Setup

For the CBAE implementation word embedding matrix E with word vectors trained by word2vec with negative sampling on each dataset. Settings of the word2vec parameters: embedding size = 200, window size = 10 and negative sample size = 5. The aspect embedding matrix initialized with the centroids of clusters results resulting from running k-means on word embeddings. Other parameters are initialized randomly. During the training process word embedding matrix E was fixed, other parameters were optimized using *Adam* with learning rate 1e-3 for 10 epochs and batch size of 50. The number of negative samples was set to 20. Following [Brody and Elhadad, 2010, Zhao et al., 2010, He et al., 2017] the number of aspects for the corpus was set to 14. In CBAE like in ABAE, representative words of an aspect can be found by looking at its nearest words in the embedding space using cosine as the similarity metric. N-gram convolutional layer setting the kernel size to 3, stride to 1 and channels to 200. Primary caps layer has 7 capsules with kernel size = 3, stride = 1 and out channels = 32, Secondary capsule layer has 7 capsules with 32 channels. Dynamic routing performs 3 iterations.

5.3 Baselines

To validate the performance of CBAE, the model was compared with next baselines:

1) **LocLDA** [Brody and Elhadad, 2010]: This method uses a standard implementation of LDA.

2) **k-means** [He et al., 2017]: Aspect matrix T with k-means centroids of the word embeddings.

3) **SAS** [Mukkerjee and Liu, 2012]: Hybrid topic model that jointly discovers both aspects and aspect-specific opinions.

4) **BTM** [Yan et al., 2013]: Bitermic topic model that is specially designed for short texts such as texts from social media and review sites. The major advantage of BTM over conventional LDA models is that it alleviates the problem of data sparsity in short documents by directly modeling the generation of unordered word-pair co-occurrences (biterms) over the corpus.

5) **ABAE** [He et al., 2017]: Attention based aspect extraction.

6 Results

Tab. 1 presents all 14 aspects inferred by CBAE.

Representative Words	Aspects
chicken potato onion tomato sauce	food
highly grilled seared roasted shrimp	food
pay money leave didn order	price
minute asked manager seated waiter	staff
wall room space ceiling lit	ambience
dish menu entree course flavor	food
birthday saturday reservation friend friday	anecdotes
year ny experience life nyc	miscellaneous
service staff waitstaff attentive friendly	staff
month time minute week hour	miscellaneous
drink wine sangria nice perfect	food
gras crust rib pork foie	food
chocolate butter fruit sauce sweet	food
chinese japanese american mexican indian	food

Table 1: List of top representative words for each aspect (left), and aspect labels (right). Aspect labels (right) were assigned manually.

Evaluation criterion of sentence-level-aspects identification is to judge how well the predictions match the true labels. It is measured by precision, recall and F_1 scores. The results are shown in Tab . 2

7 Conclusion

The analysis of the results shows that the attention-based model in general has more scores. Greater recall of capsule-based model is determined by the lower severity of aspects relative to the attention-based model (for example *nice* and *perfect* with *drink* and *wine* or *leave* with *pay* and *money*). Perhaps the orchestration of these models will make it possible to obtain higher marks in the problem of aspect extraction.

Aspect	Method	Precision	Recall	F_1
food	LocLDA	0.90	0.65	0.75
	SAS	0.88	0.77	0.82
	BTM	0.93	0.75	0.82
	k-means	0.93	0.65	0.76
	ABAE	**0.95**	0.74	0.83
	CBAE	0.89	**0.85**	**0.87**
staff	LocLDA	0.80	0.59	0.68
	SAS	0.77	0.56	0.65
	BTM	**0.83**	0.58	0.68
	k-means	0.79	0.69	0.66
	ABAE	0.80	0.73	**0.77**
	CBAE	0.68	**0.75**	0.71
ambience	LocLDA	0.60	0.68	0.64
	SAS	0.78	0.54	0.65
	BTM	0.81	0.6	0.69
	k-means	0.73	0.64	0.68
	ABAE	0.82	**0.70**	**0.74**
	CBAE	**0.83**	0.62	0.71

Table 2: Aspect identification results. The result of LocLDA is taken from [Zhao et al., 2010]; the result of SAS is taken from [Wang et al., 2015]; the results of BTM, k-means and ABAE are taken from [He et al., 2017]

.

References

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.

[Blei et al., 2003] Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. volume 3, pages 993–1022.

[Brody and Elhadad, 2010] Brody, S. and Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews.

[Chen et al., 2014] Chen, Z., Mukherjee, A., and Liu, B. (2014). Aspect extraction with automated prior knowledge learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

[Ganu et al., 2009] Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the 12th International Workshop on the Web and Databases*.

[He et al., 2017] He, R., Lee, W. S., Ng, H. T., and Dahlmeirer, D. (2017). An unsupervised neural attention model for aspect extraction.

[Hu and Liu, 2004] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Dicovery and Data Mining*.

[Jin and Ho, 2009] Jin, W. and Ho, H. H. (2009). A novel lexicallized hmm-based learning frameworks for web opinion mining. In *Proceedings of the 26th International Conference on Machine Learning*.

[Li et al., 2010] Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., Zang, S., and Yu, H. (2010). Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computer Linguistics*.

[Mikolov et al., 2017] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2017). Efficient estimation of word representations in vector space.

[Mnih et al., 2014] Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention.

[Mukkerjee and Liu, 2012] Mukkerjee, A. and Liu, B. (2012). Aspect extraction through semi-supervised modelling. In *Proceedings of the 50th Annual Meetings of the Association for Computational Linguistics*.

[Qiu et al., 2011] Qiu, G., Liu, B., Bu, J., and Chen, C. (2011). Opinion word expansion and target extraction thorugh double propagation. *Computation Linguistics*, 37:9–27.

[Rush et al., 2015] Rush, A., Chopa, S., and Weston, J. (2015). A neural attention model for sentence summarization. In *Proceeding of the 2015 Conference on Emperical Methods in Natural Language Processing*.

[Sabour et al., 2017] Sabour, S., Frost, N., and Hinton, G. E. (2017). Dynamic routing between capsules.

[Somasundaran and Wiebe, 2009] Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*.

[Wang et al., 2015] Wang, L., Liu, K., Cao, Z., Zhao, J., and de Malo, G. (2015). Sentiment-aspect extraction based on restricted boltzmann machines. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

[Wang et al., 2016] Wang, W., Pan, S. J., Dahlmeier, D., and Xiao, X. (2016). Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

[Yan et al., 2013] Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd International World Wide Web Conference*.

[Yin et al., 2016] Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., and Zhou, M. (2016). Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

[Zhao et al., 2015] Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., and Zhao, Z. (2015). Investigating capsule networks with dynamic routing for text classification. In *Proceeding of the 2015 Conference on Emperical Methods in Natural Language Processing*.

[Zhao et al., 2010] Zhao, W. X., Jiang, J., Yan, H., and Li, X. (2010). Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Emperical Methods in Natural Language Processing*.

[Zhuang et al., 2006] Zhuang, L., Jing, F., and Zhu, X.-Y. (2006). Movie review mining and summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*.