Mining What Developers Are Talking About Deep Learning

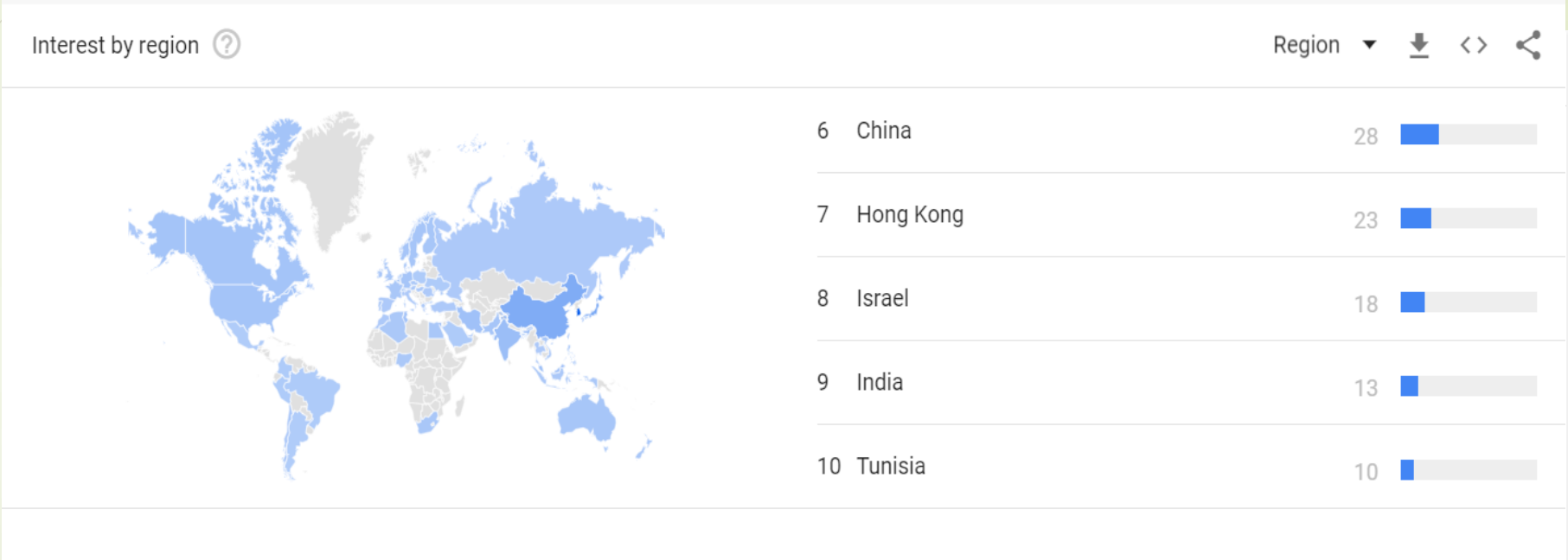
LYU1801

JIN Fenglei

Supervisor: Michael R. Lyu

Contents

Motivation



Search interest of deep learning

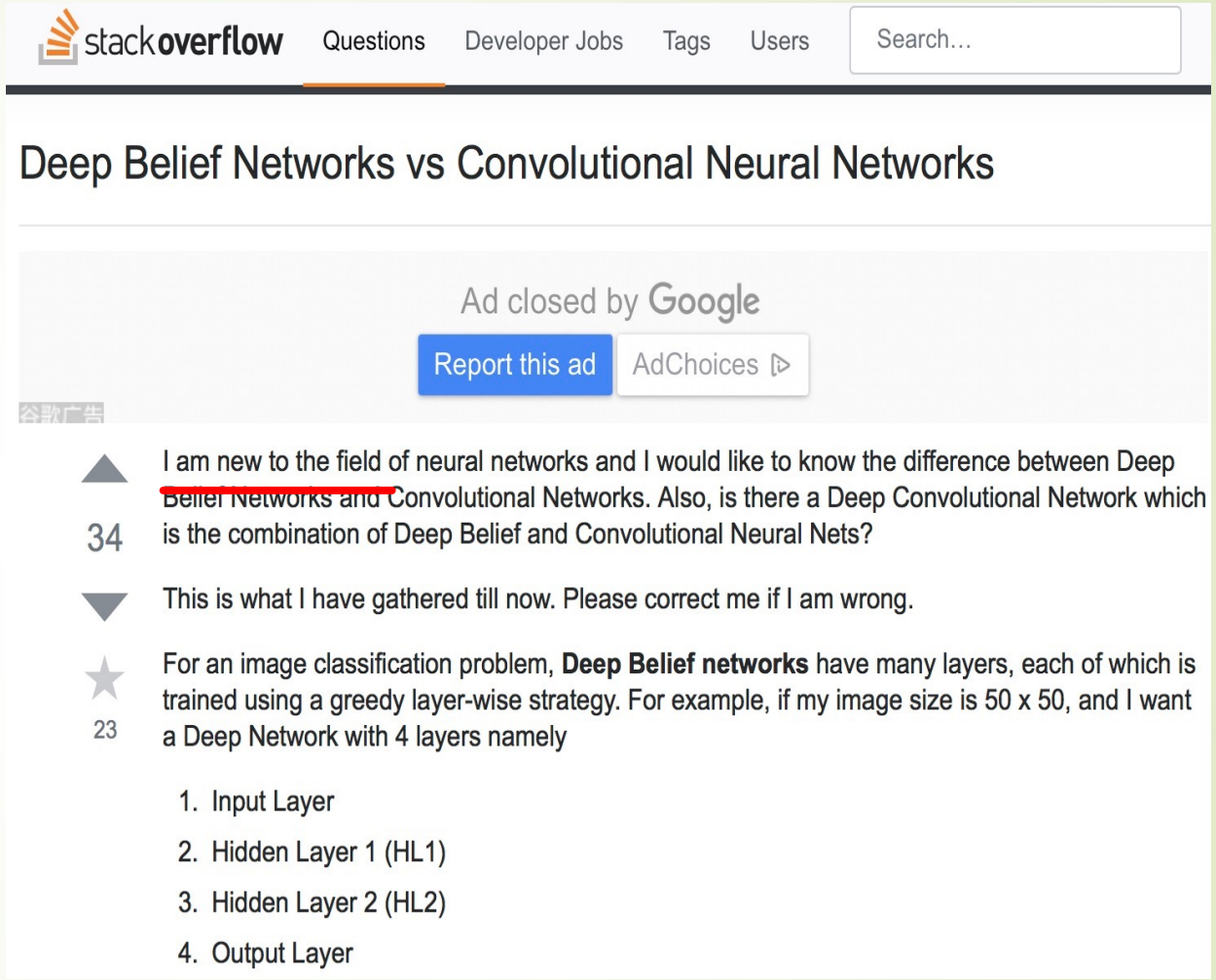
- Deep learning is popular!

Motivation

- Lots of engineers and researchers are jumping into this area.
  - More and more papers about deep learning
  - 36 FYP about deep learning this year!

Motivation

- Many **new** developers tend to enter this field and ask some basic questions.
- It is significant and necessary for the "newbies" to have a **brief understanding** about this field



Questions asked by "newbie"

Motivation

- Questions posted by developers directly reflect the **focus** of the deep leaning field.
  - In October 2017, lot of posts contain "Sophia", which is an AI robot and the first robot to receive citizenship at that time.
- For experienced developers, knowing the newest information gives them **inspiration**.

Are the dialogs at Sophia's (the robot) appearings scripted?

7

2

I talk about the robot from: Hanson Robotics, which was granted the right to citizenship from Saudi Arabia.

I have found the following articles:

Your new friend is a humanoid robot

source: theaustralian.com.au

> Like Amazon Echo, Google Assistant and Siri, **Sophia can ask and answer questions about discrete pieces of information**, such as what types of movies and songs she likes, the weather and whether robots should exterminate humans.
>
> But her general knowledge is behind these players and she doesn't do maths. **Her answers are mostly scripted** and, it seems, from my observation, her answer are derived from algorithmically crunching the language you use.
>
> Sometimes answers are close to the topic of the question, but off beam. Sometimes she just changes the subject and asks you a question instead.
>
> She has no artificial notion of self. **She can't say where she was yesterday, whether she remembers you from before**, and doesn't seem to amass data of past interactions with you that can form the basis of an ongoing association.
>
> Questions such as: *"What have you seen in Australia?"*, *"Where were you yesterday?"*, *"Who did you meet last week?"* and *"Do you like Australia?"* are beyond her.

Questions about "Sophia"

Contents

- Motivation
- Related work
- Methodology
- Experimentation
- Future work

Related Work

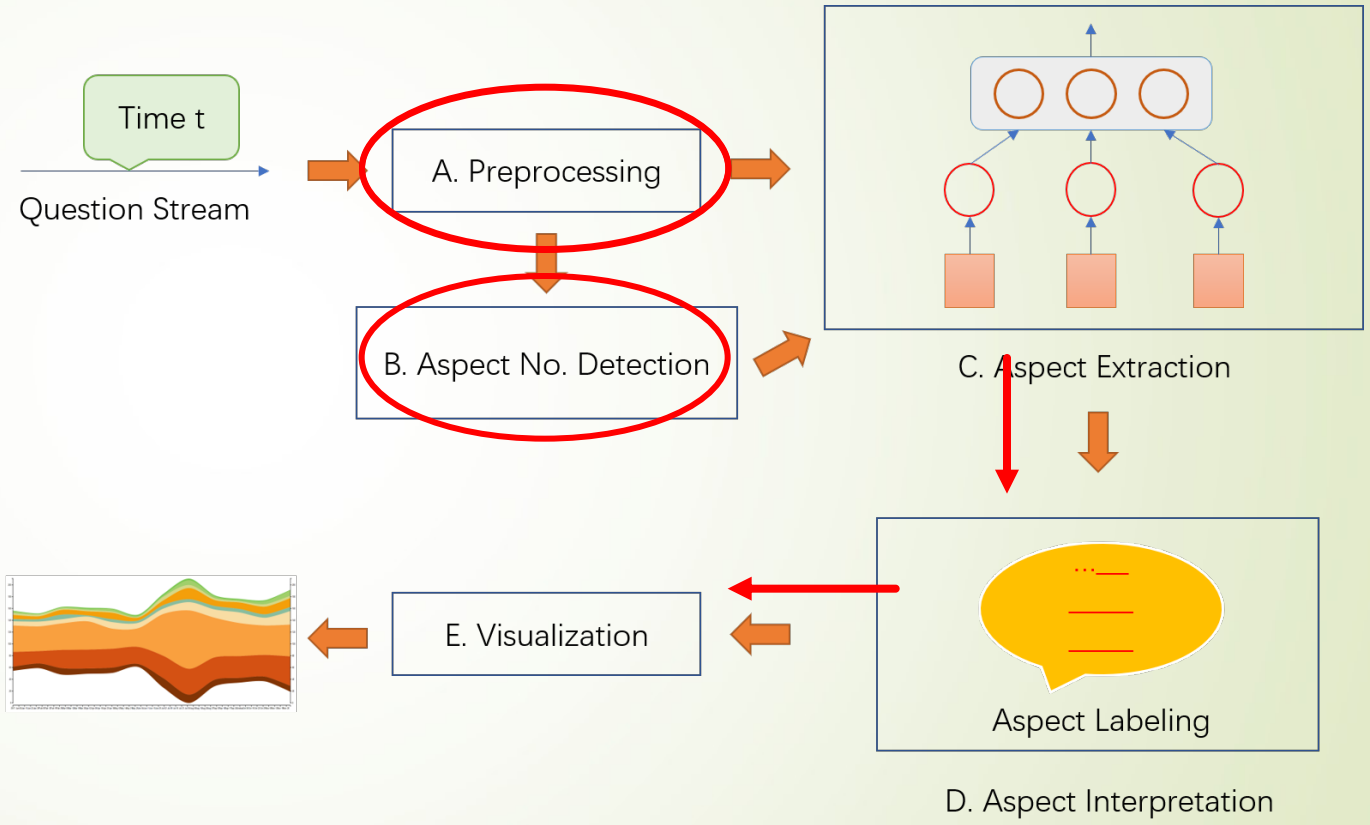
- Previous works for aspect extraction can be categorized into three approaches: rule-based, supervised, and unsupervised
  - Hu and Liu (2004) tried to extract different features by finding the frequency of nouns and noun phrases
  - Jin and Ho (2009) proposed hidden Markov models (HMM) and Li et al. (2010) proposed conditional random fields (CRF)
  - LDA (Blei et al., 2003) and its variants are the most popular unsupervised approaches
    - Attention-based Aspect Extraction (ABAE) model (He et al., 2017)

Contents

Methodology

overview

- Part A: preprocesses the raw questions
- Part B: get the most appropriate number for aspect
- Part C: extract aspects
- Part D: interpret the topic
- Part E: visualization

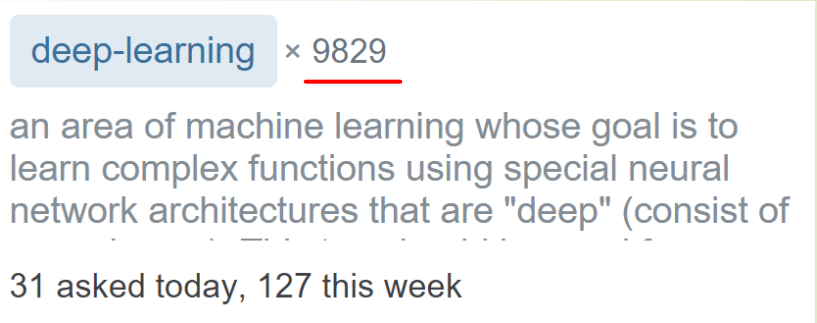


Framework of our model

Methodology

data crawling

- Over 5,500 questions provided by StackExchange
- Over 9,000 questions under the tag of deep-learning in StackOverflow
- Use a python package called **scrapy** to crawl the data in StackOverflow
- Enter the website of every question to crawl the detailed information



StackOverflow deep-learning tag

Methodology

A. *preprocessing*

- Difficulties:
  - massive noisy words
  - codes, terminologies and websites
  - HTML tags


```
{"title": "Reduce image dimensions in python",
 "question": "<div class=\"post-text\" itemprop=\"text\">\r\n\r\n<p>I have in input an
            image with dimensions (28, 28, 3).
            I trained a keras model with several images with dimensions (28, 28, 1). I
            want \n   to check a single test image with this model, but every time I get
            a dimension error. How can I reduce original dimensions (28, 28, 3) to (28,
            28, 1)?</p>\n\n<pre><code>test_image = image.load_img('test/number3.png' ,
            target_size = (28, 28))\ntest_image = image.img_to_array(
            test_image)\ntest_image = np.expand_dims(test_image, axis = 1)\nresult =
            classifier.predict(test_image)\n</code></pre>\n    </div>", "answer": "<div
            class=\"post-text\" itemprop=\"text\">\r\n<p>Depending on how you would like
            to reduce dimensionality you can just choose one of the colour channels like
            this</p>\n\n<pre><code>one_channel_image =
            test_image[:,:,0]\n</code></pre>\n\n<p>or you could find use the mean across
            the colour channels</p>\n\n<pre><code>one_channel_image = np.mean(
            test_image, axis=2)\n</code></pre>\n\n<p>In my experience of ML image
            problems just taking one channel works fine.</p>\n\n<p>If you need to
            increase dimensionality from (28, 28) to (28, 28, 1) you can use
            numpy.reshape</p>\n\n<pre><code>one_channel_image = test_image.reshape((28,
            28, 1))\n</code></pre>\n    </div>"},
```


Massive question

Methodology

A. *preprocessing*

- Word Formatting:
  - lowercase
  - lemmatization
- Word Filtering:
  - reduce the non-informative words
- Word Replacement:

Non-informative parts	Replacing words
Websites (eg: http://..., https://...)	url
All numbers	<num>
Image html tag	img
Code, pseudocode	code
Unknown words in dictionary	<unk>

Methodology

A. *preprocessing*

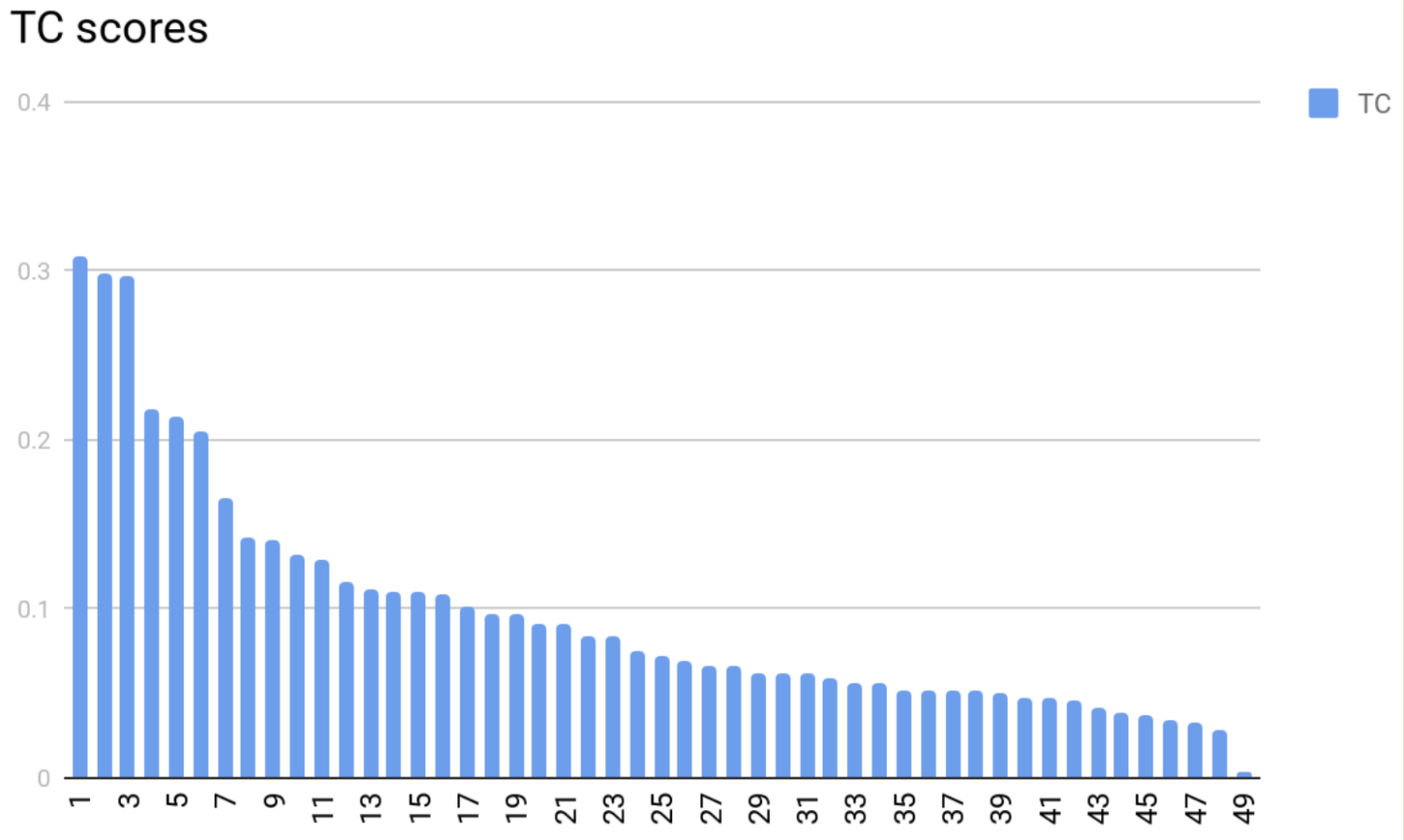
- HTML Tags Summarization:

Tags	Description	Tags	Description
 	new line		ordered list
<hr>	thematic change in the content	<blockquote>	a section that is quoted from another source
	stress emphasis	<pre>	a preformatted text
	important text	<code>	a code or pseudocode (handled before)
<h1>, <h2>, <h3>	define HTML headings	<img src=…>	image (handled before)
	unordered (bulleted) list	…	

Methodology

B. *Aspect Number Detection*

- Each aspect explains a **certain portion** of the total correlation
- Additional aspects should be added until additional aspects contribute **little** to the overall total correlation



Topic correlation distribution

Methodology

C. Model

- Goal: learn a certain number of aspects embeddings

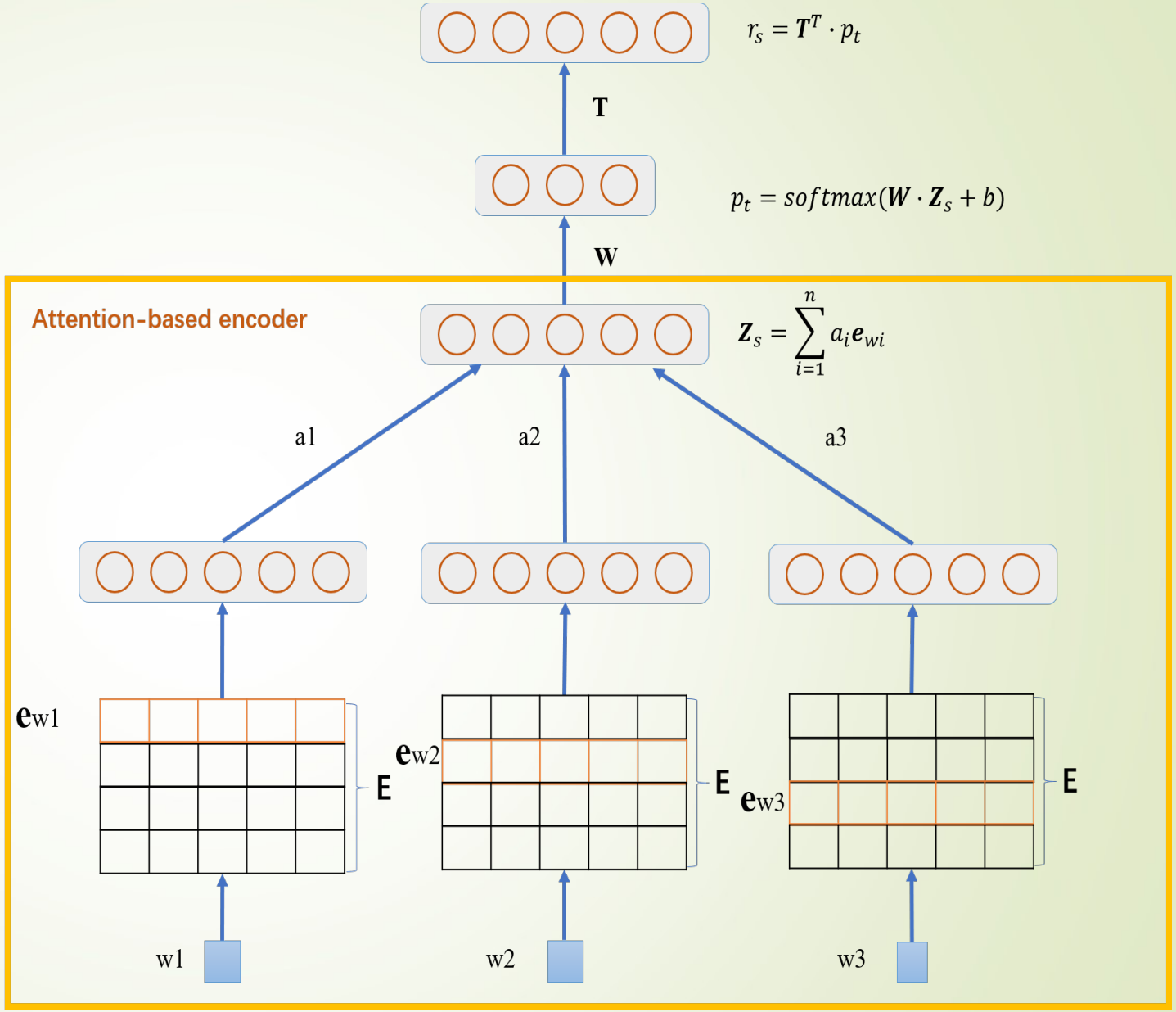


Illustration of ABAE

Methodology

C. Model

- Represent each word w with a **feature vector** (word embedding)
- Word embedding matrix ***E*** describes the feature vectors associated with the words by row locations

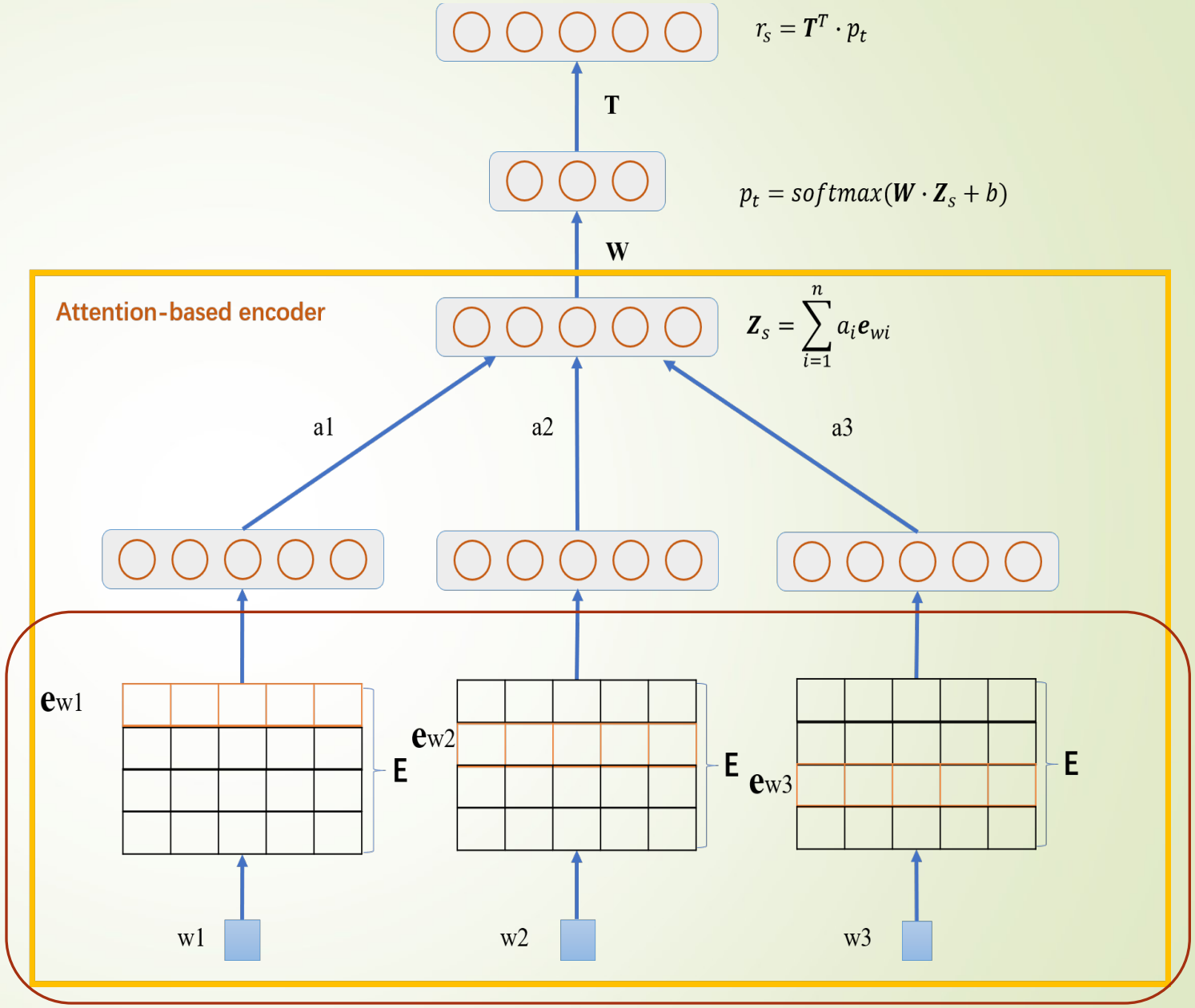


Illustration of ABAE

Methodology

C. Model

- Attention mechanism helps filter away non-aspect words

$$\boldsymbol{a}_i = \frac{\exp(d_i)}{\sum_{j=1}^{n} \exp(d_i)}$$

$$d_i = \boldsymbol{e}_{w_i}^T \cdot \boldsymbol{M} \cdot \boldsymbol{y_s}$$

$$\boldsymbol{y}_s = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{e}_{w_i}^T$$

- Vector representation $\mathbf{z}_s$ is constructed from

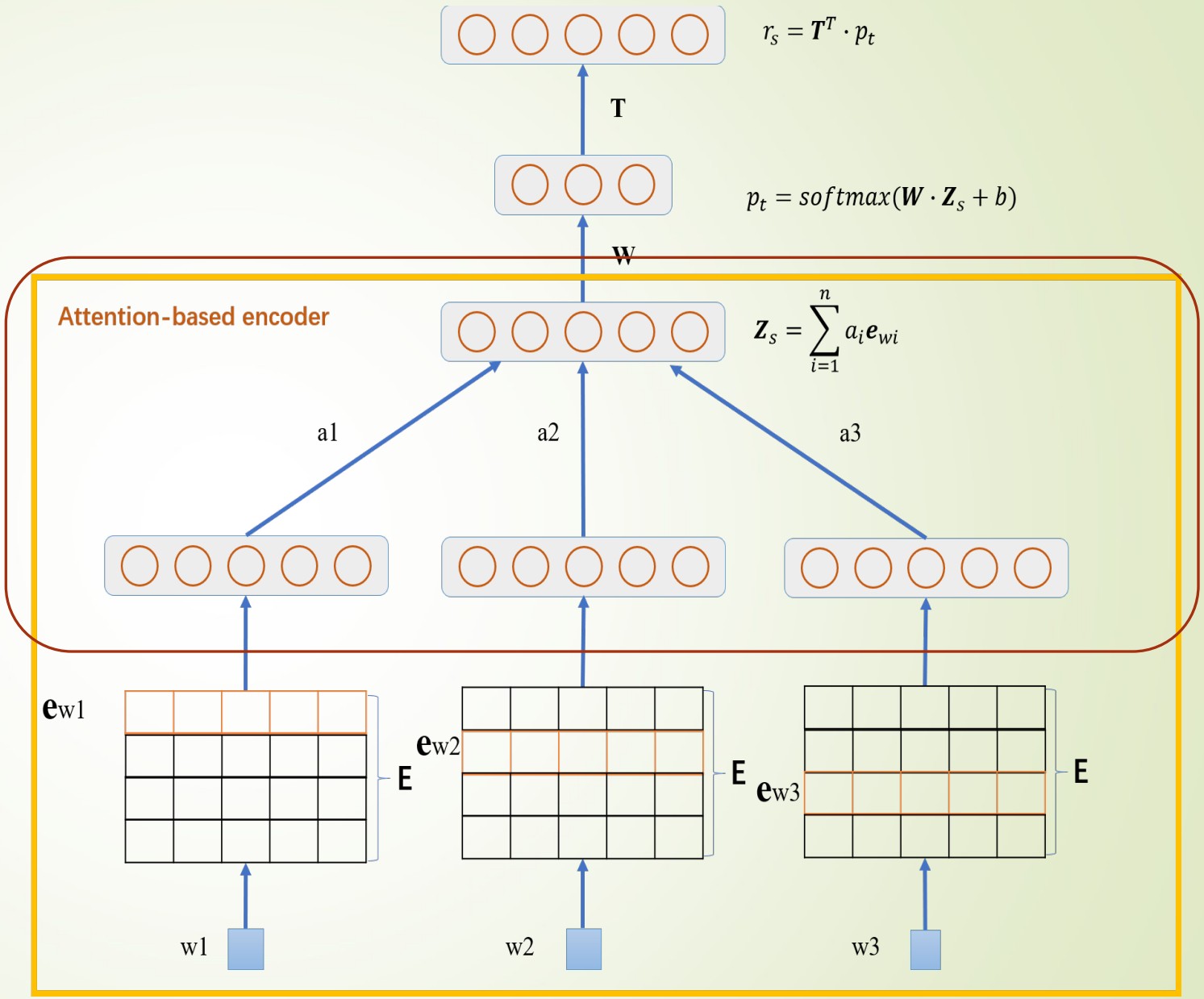


Illustration of ABAE

Methodology

C. Model

- Reconstruct the sentence embedding from aspect embedding matrix
- P_t: the probability that the input belongs to the related aspect

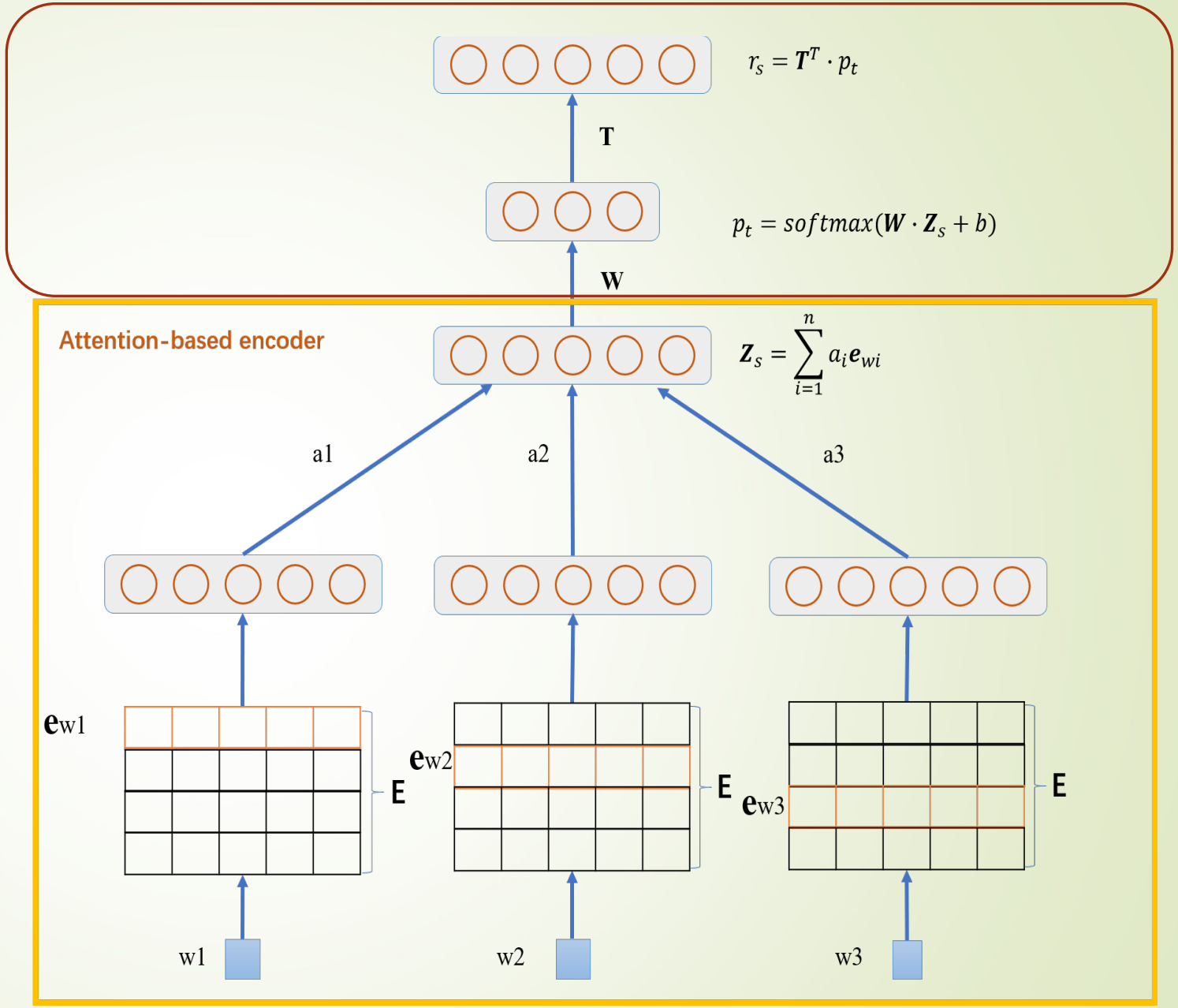


Illustration of ABAE

Methodology

C. Model

- Training Objective: minimize the re-construction error

$$J(\theta) = \sum_{s \in D} \sum_{i=1}^{m} \max(0, 1 - \boldsymbol{r}_s \boldsymbol{z}_s + \boldsymbol{r}_s \boldsymbol{n}_i)$$

$$U(\theta) = \left| \left| \boldsymbol{T}_n \cdot \boldsymbol{T}_n^T - \boldsymbol{I} \right| \right|$$

$$L(\theta) = J(\theta) + \lambda U(\theta)$$

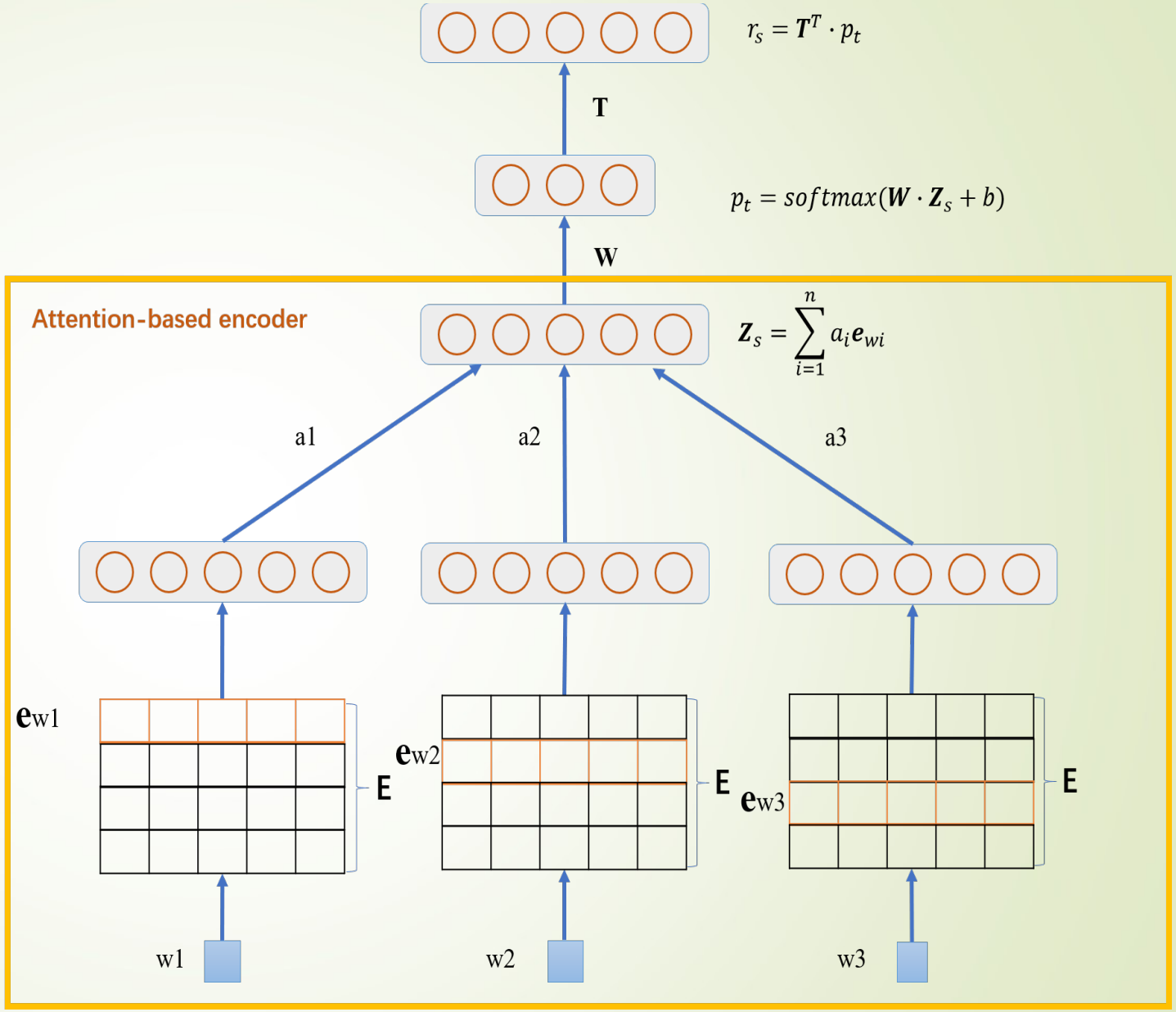


Illustration of ABAE

Contents

- Motivation
- Related work
- Methodology
- Experimentation
- Future work

Experimentation

Dataset

- StackExchange: 5,500

  StackOverflow: 7,000
- Divided test dataset in 2017 into 12 months

Month	Question No.	Month	Question No.
2017-01	147 questions	2017-07	179 questions
2017-02	113 questions	2017-08	229 questions
2017-03	144 questions	2017-09	179 questions
2017-04	153 questions	2017-10	187 questions
2017-05	136 questions	2017-11	175 questions
2017-06	114 questions	2017-12	189 questions
TOTAL	1945 questions		

Experimentation

Training with html tags

- **html structures** are learned rather than different topics
- Fake lower loss

```
Aspect 1:
[u'strong', u'h1', u'noreferrer', u'oxforddictionaries', u'ab', u'h2', ...]
...
Aspect 4:
[u'hr', u'answer', u'ask', u'emotion', u'think', u'question', ...]
Aspect 5:
[u'code', u'feature', u'camp', u'variable', u'vector', u'gt', ...]
Aspect 7:
[u'pre', u'en', u'wikipedia', u'rel', u'org', u'convolutional_neural_network', ...]
Aspect 8:
[u'stack', u'img', u'jpg', u'png', u'alt', u'imgur', ...]
Aspect 9:
[u'ol', u'li', u'general', u'human', u'intelligent', u'agi', ...]
Aspect 10:
[u'ul', u'exchange', u'post', u'overflow', u'stackexchange', u'datascience', ...]
...
Aspect 12:
[u'p', u'train', u'use', u'used', u'using', u'network', ...]
```

Aspect terms with tags and noisy word

Experimentation

Training

- 20,000 iteration
- Loss stop at 10
- Manually assign topic description

Experimentation

Training

Order No.	Top words	Label
Aspect 0	Goal, current, player, minimax, state, decision	Decision making algorithm
Aspect 1	Consume, restore, gpu	Storage
Aspect 2	Graffiti, identify	Image Identification
Aspect 3	Artificial, intelligence, resnets, neural	Deep learning model
Aspect 4	Enforcement, convnets, smoothness	Image Identification
…	…	…
Aspect 7	Neural, caffe2, stimulate	Deep learning platform
Aspect 12	Cocke(Cocke-Kasami-Younger algorithm), parsing	NLP

Experimentation

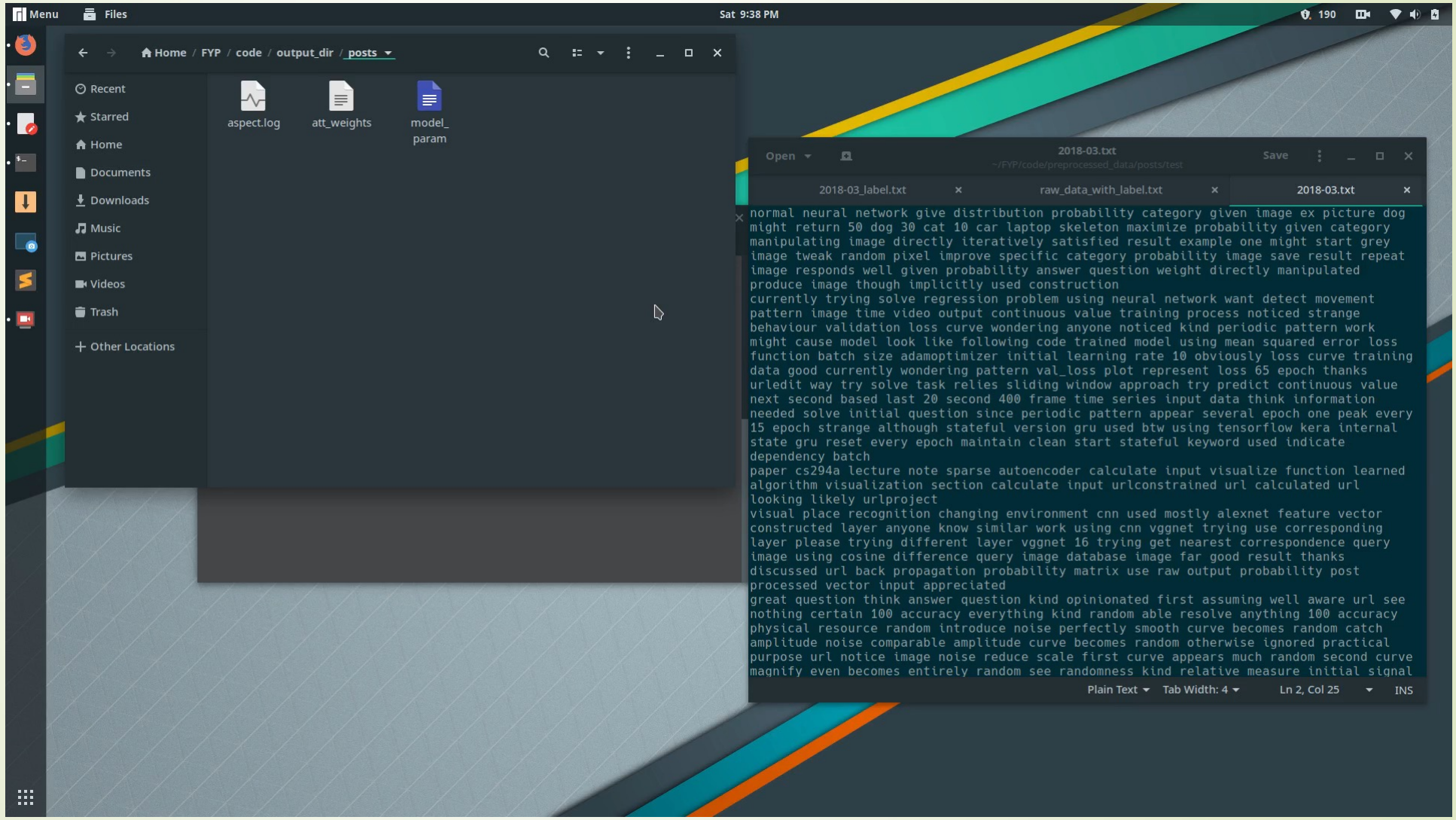
Training

Order No.	Top words	Label
Aspect 13	Data, training, set, test ,model, learning, recognition, algorithm	Dataset
…	…	…
Aspect 24	Melfrequency(MFCCs), recalibrate, electric	Voice recognition
…	…	…
Aspect 27	Learning, algorithm, procedural, reinforcement	Learning strategy
…	…	…
Aspect 44	Flu, south, sexual, elasticity, noob	Noise

Experimentation

Test

- Simple test



Experimentation

Test

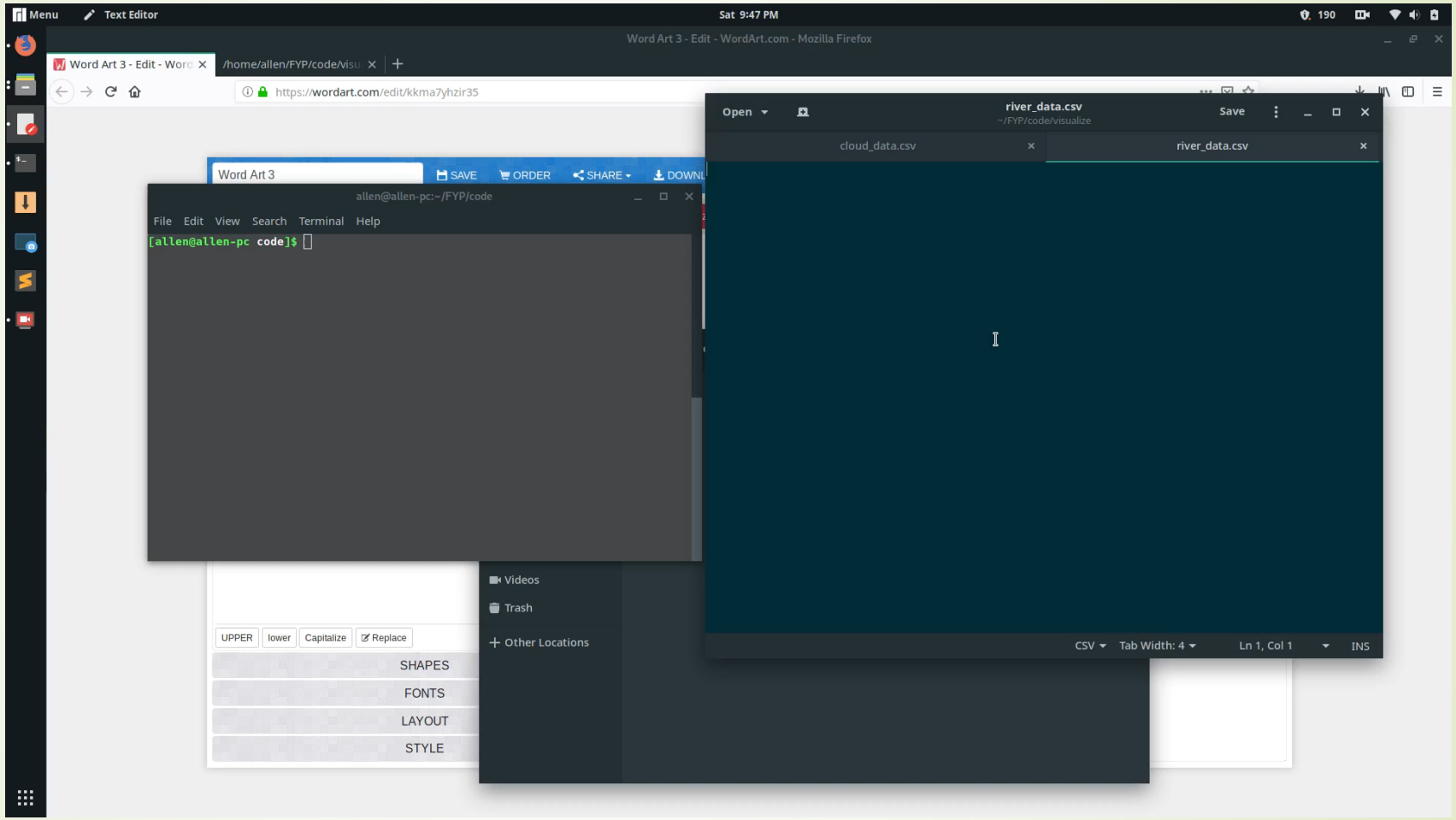
- Attention weight

	supervised		data		input		data		begin		function
0.016	0.301	0.23	0.016	0.016	0.016	0.016	0.016	0.119	0.016	0.119	0.119
performing		classification		vector		outcome		train		error	

The attention mechanism focus more on “supervised" and “classification" since it is predicted to be “learning strategy".

Experimentation

Visualization



- Word cloud: http://appsrv.cse.cuhk.edu.hk/~fljin7/fyp/cloud.html
- Topic river: http://appsrv.cse.cuhk.edu.hk/~fljin7/fyp/index.html

Summery

- Crawl over 7,000 questions about deep learning in StackOverflow
- Use Hierarchical Topic Model to detect appropriate aspect number in a corpus
- Simulate the Unsupervised Attention-based Aspect Extraction Model and learn the aspects embedded in deep learning related questions
- Visualize and analyze the extracted topics and their trends

Contents

- Motivation
- Related work
- Methodology
- Experimentation
- Future work

Future work

- Phrase extraction when preprocessing
- Manually label some test data to further quantitatively evaluate the prediction accuracy of the model
- Automatic aspect interpretation
- Use model taking time as one of the parameters to detect emerging issues

Q & A

Thank you!