

Product Function Need Recognition via Semi-supervised Attention Network

Hu Xu*, Sihong Xie[†], Lei Shu*, Philip S. Yu*[‡]

*Department of Computer Science, University of Illinois at Chicago, Chicago, IL, USA

[†]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA

[‡]Institute for Data Science, Tsinghua University, Beijing, China

hXu48@uic.edu, sxie@cse.lehigh.edu, lshu3@uic.edu, psyu@uic.edu

Abstract—Functionality is of utmost importance to customers when they purchase products. However, it is unclear to customers whether a product can really satisfy their needs on functions. Further, missing functions may be intentionally hidden by the manufacturers or the sellers. As a result, a customer needs to spend a fair amount of time before purchasing or just purchase the product on his/her own risk.

In this paper, we first identify a novel QA corpus that is dense on product functionality information¹. We then design a neural network called Semi-supervised Attention Network (SAN) to discover product functions from questions. This model leverages unlabeled data as contextual information to perform semi-supervised sequence labeling. We conduct experiments to show that the extracted function have both high coverage and accuracy, compared with a wide spectrum of baselines.

Keywords-product function need recognition; semi-supervised learning; deep learning; attention

I. INTRODUCTION

Functionality is a fundamental concern for customers when they decide to buy a new product. From *customers'* perspective, before they purchase a product (e.g., a laptop), it is natural for them to ask what the to-be-purchased one can do and cannot do. From *sellers'* perspective, selling fully-functioned products can increase sales, and yet selling products with missing functions can lead to catastrophic customer dissatisfaction. From *manufacturers'* perspective, missing functions reported by customers can help improve their products. In marketing, the term *product* is defined as “anything that can be offered to a market for attention, acquisition, use or consumption that might satisfy a want or need” [1]. It is crucial to ensure that the functions of a product can satisfy customers' needs. Therefore, conveying the information about functions successfully to customers is important for both manufacturers and sellers.

In e-commerce platforms, one issue to convey such information is that products cannot be physically presented to customers before purchasing. To overcome such limitation, many alternative approaches are deployed, i.e., using descriptions, pictures, and videos. However, detailed functionality information may not be readily available for the following reasons. 1) The cost of testing functions multiplied by a large number of products can be extremely

¹The annotated corpus can be found at <https://www.cs.uic.edu/~hXu/>.

Table I: A few QA pairs for a laptop: function expressions are underlined with function words (e.g., verbs, adjectives or prepositions) bolded.

Apple 13 ” MacBook Pro (2.5GHz Intel Core i5, 4GB RAM, 500GB HDD)	
Q:	Can I <u>use</u> this for video editing
A:	No, it does not support Google Play.
Q:	Can I <u>make</u> video calls to other non Apple computers ? ?
A:	yes you can if they have Skype , Tango , or oovoo
Q:	Will it be <u>useful</u> for music production ?
A:	I have not used it for music production ; however , I believe that it would be and have several friends who use it specifically for that purpose .
Q:	Can I <u>use</u> Microsoft Office on this MacBook Pro ?
A:	You can but maybe you wo n’t want to . The current Apple MacBook Pro is shipping with the Mavericks operating system , which includes Pages , Numbers , and Keynote at no cost .

high. For example, it is impossible to test so many PCs whether they can run specific high-performance PC games. 2) Some missing functions are deliberately hidden from descriptions by sellers to avoid hurting sales.

Fortunately, functionality information can be exchanged between customers and sellers via online platforms, such as forums and community QA. This allows us to adopt an NLP-based approach to automatically sense and harvest product functions on a large scale. We formulate a novel text mining task called Function Need Recognition (or FNR for short). A function need is defined as a sequence of words indicate a function expression (e.g., “make video calls”). In this paper, we only focus on product function needs and leave satisfiability issues (e.g., whether a product can “make video calls”) to future work².

This task is non-trivial and the following challenges have to be addressed. First, to ensure extraction quality, corpora that are dense and accurate in product functionality information are preferred. To the best of our knowledge, there is no existing study on such a corpus to meet these requirements. Second, the number of function needs can be unlimited. How to ensure unexpected function needs can be detected is important.

²A comprehensive study of product function satisfiability can be found at AAAI-2018 [2].

We address the challenges by first identify and annotate a high-quality corpus. In particular, Amazon.com allows potential consumers to communicate with existing product owners or sellers regarding product functions via Product Community Question Answering (PCQA for short). Four (4) QA pairs talking about a laptop sold on Amazon are shown in Table I. Observe that the name of target product (to-be-purchased) can be identified using the metadata of the target product. But 4 function needs (“use for video editing”, “make video calls”, “useful for music production”, and “use Microsoft Office”) should be identified from the questions.

Given the corpus, we then formulate the problem as a sequence labeling task on questions. We propose a deep sequence labeling model called Semi-supervised Attention Network (SAN) to solve this problem. The key property of SAN is to use attention mechanism to summarize unlabeled data as side information for short labeled questions. For example, let us assume only the 1st question is in the labeled data and all other 3 questions are in unlabeled data. Then words like “use” or “video” in other 3 questions can serve as side information to help identify that “use for video editing” is a function. Also, another advantage of using unlabeled data is that the embeddings of words do not appear in labeled data can still be tuned during training. To the best of our knowledge, this is the first attempt to use attention mechanism in a semi-supervised setting.

II. MODEL AND PRELIMINARY

A. Model Overview

We briefly introduce the proposed Semi-supervised Attention Network (SAN) in this section. The idea of the network is to couple RNN-based sequence labeling network with attention on unlabeled data. The proposed network is illustrated in Fig. 1. The left side can be viewed as a supervised sequence labeling model. It reads in a (labeled) question \mathbf{x}^q and outputs label sequence $\mathbf{y} = (y_1, \dots, y_t, \dots, y_{T_q})$, where $y_t = l \in L = \{F, O\}$. The right side is the semi-supervised part. A few unlabeled questions $U = \{\mathbf{x}^{u_1}, \dots, \mathbf{x}^{u_n}, \dots, \mathbf{x}^{u_{|U|}}\}$ are fed into a bank of BLSTMs (Bidirectional Long Short-Term Memory [3], [4], one for each unlabeled question) with attentions (called bank attention). The attended results are served as side information for the (labeled) question. The key point here is, given a labeled question, we need to learn the weights on how to attend (or read) unlabeled questions. Note that both supervised and semi-supervised parts share the same embedding layer. This also gives the opportunity to tune embeddings of words not appear in the labeled questions. Such a tuning is impossible in supervised settings. All unlabeled questions share the same weights for their BLSTM layers (not shown in the figure). After each word in the labeled question obtains the side formation, we feed the augmented labeled question into another BLSTM layer. Then we generate label sequence \mathbf{y} via a softmax layer. Overall, the labeled question can

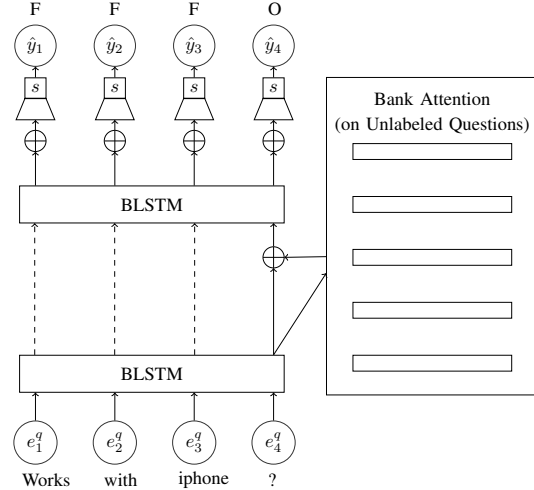


Figure 1: Semi-supervised Attention Network (SAN): the bottom 4 words are an input (labeled) question. They are labeled as F, F, F, O , indicating “Works with iphone” is a function expression. On the right is bank attention on unlabeled questions (sample questions are omitted).

leverage unlabeled questions to decide the output labels in an end-to-end manner.

B. Preliminary

Embedding Layer We pair each labeled question \mathbf{x}^q with a few unlabeled questions $U = \mathbf{x}^{u_1:|U|}$ (for both the training data and the test data). Unlabeled questions are similar questions from the same category as the labeled question returned by a search engine. Let the sequence $\mathbf{x}^q = (x_1^q, \dots, x_{T_q}^q)$ and $\mathbf{x}^{u_n} = (x_1^{u_n}, \dots, x_{T_{u_n}}^{u_n})$ denote the labeled question and the n -th unlabeled question, respectively. Here T_q and T_{u_n} denote their respective lengths. When a question contains multiple sentences, we concatenate them into a single sequence. We separate the sentences by a special token EOS . We set $T_q = T_{u_1:n} = 40$, which covers 99.5% of lengths of labeled questions. Questions longer (shorter) than 40 words are truncated (padded with zeros). We can view \mathbf{x}^q (\mathbf{x}^{u_n} , resp.) as a matrix of one-hot column vectors. \mathbf{x}^q is later transformed into embedded representation e^q (e^{u_n} , resp.). We pre-train the word embedding via skip-gram model [5]. Then we fine-tune the embeddings when optimizing the proposed model.

BLSTM Layer The embedded question sequences (e^q and $e^{u_1:|U|}$) are fed into the labeled BLSTM and the unlabeled BLSTMs, respectively. We use $h^{q,1}$ and $h^{u_1:|U|}$ to denote the outputs of these BLSTM layers for the labeled question and unlabeled questions, respectively. We show important notations in Table II, which is used in the next section.

Table II: Notations

Notation	Explanation
$h^{q,d}$	The d -th hidden representation of the labeled question q ($d = 1, 2, 3$)
h^{u_n}	Hidden representations of the n -th unlabeled question u_n
t	The t -th word in the labeled question
v	The v -th word in an unlabeled question
r	Indicator of transformed representation for the labeled question q
k, k'	Indicators of transformed representation for the unlabeled question u_n
$\alpha_{t,v}^{q,u_n}$	Level 1 attention weights for the t -th word in q on the v -th word in u_n .
α_t^{q,u_n}	Level 2 attention weights for the t -th word in q on u_n
h_t^{q,u_n}	Level 1 attended representation: the t -th word in q attends on unlabeled question u_n
s_t^q	Level 2 attended representation: the t -th word in q attends on all $U = u_{1:n}$

III. SEMI-SUPERVISED ATTENTION NETWORK

A. Bank Attention

The key point of SAN is to leverage attention mechanism for semi-supervised learning. We utilize attention mechanism to synthesize side information from unlabeled data for each word in a labeled question. The idea is that words in unlabeled data may have useful information for sequence labeling when they talk about similar products. We introduce a hierarchical attention mechanism. As traditional attention mechanism, we let each word in a labeled question to attend a word in an unlabeled question. This is level 1 attention. On the higher level, we pair a labeled question with multiple related unlabeled questions. Note that different questions may not equally contribute side information to the labeled question. So we allow one word in the labeled question to attend on the results of level 1 attention on multiple questions. We use the term *bank attention* to refer to one word in a labeled question hierarchically attending to unlabeled questions. The details are shown in Fig. 2.

We try to get the side information for the t -th word in the labeled question. We first transform the word representations of the labeled question $h^{q,1}$ and unlabeled question h^{u_n} via respective fully connected layers. Then the representations are activated by tanh:

$$\begin{aligned} h_t^{q,r} &= \tanh(W^r h_t^{q,1} + b^r) \\ h_v^{u_n,k} &= \tanh(W^k h_v^{u_n} + b^k), \end{aligned} \quad (1)$$

where W^r , b^r , W^k and b^k are trainable weights. The t -th word in the labeled question first obtain the attention weight for the v -th word in the n -th unlabeled question via a dot product. Then the weights are normalized by a softmax function:

$$\alpha_{t,v}^{q,u_n} = \frac{\exp((h_t^{q,r})^T h_v^{u_n,k})}{\sum_{v'=1}^{T_{u_n}} \exp((h_t^{q,r})^T h_{v'}^{u_n,k})}. \quad (2)$$

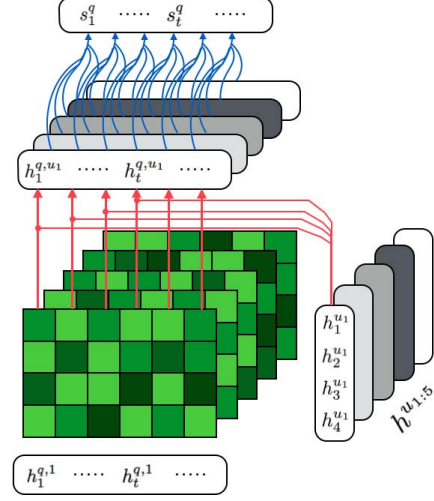


Figure 2: Bank Attention: the t -th word representation $h_t^{q,1}$ obtains its side information s_t^q from multiple unlabeled questions such as the 1st unlabeled question $h_{1:4}^{u_1}$. The red arrows indicate level 1 attention among different words in one unlabeled question (we omit the arrows for the other 4 questions). The blue arrows indicate level 2 attention among multiple representations of unlabeled questions.

This is the level 1 attention weights. Let h_t^{q,u_n} denote the side information of the t -th word in the labeled question for the n -th unlabeled question (representation after the first-level attention). It is the weighted sum over all words in the n -th unlabeled question.

$$h_t^{q,u_n} = \sum_{v=1}^{T_{u_n}} \alpha_{t,v}^{q,u_n} h_v^{u_n,k}. \quad (3)$$

Later, we have a level 2 attention over different unlabeled questions. Again we first transform the side information of the t -th word for each unlabeled question:

$$h_t^{q,u_n,k'} = \tanh(W^{k'} h_t^{q,u_n} + b^{k'}). \quad (4)$$

Then the level 2 attention weights are again obtained via dot products normalized by a softmax function:

$$\alpha_t^{q,u_n} = \frac{\exp((h_t^{q,r})^T h_t^{q,u_n,k'})}{\sum_{n'=1}^{|U|} \exp((h_t^{q,r})^T h_t^{q,u_{n'},k'})}. \quad (5)$$

And finally the side information vector for the t -th word in the labeled question (representation after level 2 attention) is:

$$s_t^q = \sum_{n=1}^{|U|} \alpha_t^{q,u_n} h_t^{q,u_n,k'}. \quad (6)$$

Lastly, we concatenate s_t^q with $h_t^{q,1}$ as the representation of the t -th word in the question: $h_t^{q,2} = h_t^{q,1} \oplus s_t^q$.

B. Sequence Labeling

After obtaining the representation of the labeled question with side information, we feed $h^{q,2}$ into another BLSTM layer. So we have two LSTM layers for the labeled question, which is similar to the stacked BLSTM [6] (S-BLSTM). We use S-BLSTM to obtain better sequence representation. Then we have $h^{q,3}$ for the labeled question sequence. We reduce the dimension of $h^{q,3}$ to the size of the label set via a fully connected layer:

$$c_t^q = Wh_t^{q,3} + b, \quad (7)$$

where $c_t^q \in \mathbb{R}^{|L|}$. We output the probability distribution over labels L for the t -th question word via a softmax function:

$$p^q(\hat{y}_t = l | \mathbf{x}^q, \mathbf{x}^{u_{1:|U|}}; \Theta) = \frac{\exp(c_{t,l}^q)}{\sum_{l' \in L} \exp(c_{t,l'}^q)}, \quad (8)$$

where Θ represents all trainable parameters, including parameters in LSTM cells and word embeddings. Finally, we optimize the cross entropy loss function over the training dataset:

$$J(\Theta) = - \sum_m \sum_t \sum_{l \in L} y_{t,l}^{(m)} \log p^q(\hat{y}_t^{(m)} = l | \mathbf{x}^q, \mathbf{x}^{u_{1:|U|}}; \Theta), \quad (9)$$

where M represents all the training examples. $y_{t,l}^{(m)} \in \{0, 1\}$ is the ground truth for the t -th question word and label l in the m -th training example. We leverage Adam optimizer [7] to optimize the whole network. We set the learning rate as 0.001 and keep other parameters the same as the original paper. We set the dropout rate to 0.2. The batch size is set to 256.

IV. EXPERIMENTAL RESULT

A. Corpus Annotation, Analysis, and Preprocessing

We crawled about 1 million QA pairs from the pages of products in the electronics department from Amazon as the training corpus for skip-gram model [5] to obtain word embedding matrix W_e .

We further annotated a subset of 4999 QA pairs from 18 products for model training and testing. The basic statistics of the corpus is shown in Table III. The corpus is labeled by 3 annotators independently. The general annotation guidelines are as follows:

- 1) only yes/no QAs should be labeled;
- 2) a function expression is labeled as a function target with an optional function verb;
- 3) a function target can be specific entities (e.g., “iPhone”), general entities like “video” or service providers like “AT&T”;
- 4) a function target should be labeled as token spans containing nouns, adjectives, or model numbers (e.g., “Samsung micro SD EVO”);

Table III: Statistics of 18 labeled products. QAs: number of QA pairs; % of QAs with Functions: percentage of QA pairs containing function needs.

Product	QA	% of QAs with Functions
DSLR	327	20.18
E-Reader	271	31.37
Speaker	153	30.72
Tablet	329	42.86
Cellphone 1	170	57.65
Cellphone 2	330	41.82
Laptop 1	297	18.86
Laptop 2	425	54.59
Netbook	199	44.72
TV	306	46.41
TV Console	183	54.1
Gaming Console	212	70.28
Apple Watch	331	28.1
VR Headset	444	76.13
Stylus	266	71.05
Micro SD Card	283	81.27
Mouse	259	66.02
Tablet Stand	214	88.79
Total	4999	51.07

- 5) expressions about specific aspects or accessories are not considered as function expressions. This is because aspects or accessories are not closely related to the functionality of the product as a whole;
- 6) nouns that are subjective are not regarded as function target (e.g., the word “need” in “Can it fit my need?”);
- 7) the optional function word can be a verb (e.g., “produce” in “produce music”) or its noun form (e.g., “production” in “music production”); we also include the adjunct word (e.g., “with” in “work with iPhone”) for extrinsic function expression;
- 8) some function expression does not have function word, e.g., “Does Skype ok on this?”;

All annotators initially agreed on their annotations (same function targets and function words) on 81% of all QA pairs. Disagreements are then resolved to reach final consensus annotations.

We observe that accessories (the last 5 products) have a higher percentage of the function need related questions than those of main products (the first 13 products). This is expected since one accessory may work with multiple devices and thus have more functions.

The annotated corpus is preprocessed using Stanford CoreNLP³. We have the following steps: sentence segmentation, tokenization, POS-tagging, lemmatization and dependency parsing. The last 3 steps provide features for the Conditional Random Fields (CRF) [8] baseline.

We also select the most similar 5 unlabeled questions under the same category as the labeled question returned by ElasticSearch⁴, as the question bank.

³<http://stanfordnlp.github.io/CoreNLP/>

⁴www.elastic.co

Table IV: Different methods for Function Need Recognition (FNR) in precision, recall and F1-score.

Method	\mathcal{P}	\mathcal{R}	\mathcal{F}_1
CRF	0.798	0.611	0.692
S-BLSTM	0.844	0.673	0.749
SAN (-) BLSTM2	0.83	0.7	0.759
SAN	0.839	0.721	0.776

We only perform sentence segmentation and tokenization on these unlabeled questions to save preprocessing time. Lastly, multiple sentences in both labeled and unlabeled questions are concatenated together. We set the maximum length of a question to be 40. This covers 99.5% labeled questions in full length.

After preprocessing, one example contains a labeled question, 5 unlabeled questions, and one labeled answer. We shuffle all examples and select 70% for training, 10% for validation and 20% for testing. The validation set is used to avoid overfitting on the training data.

B. Baselines

We compare the following baselines with SAN:

- 1) **CRF**: We use Mallet⁵ as the CRF implementation. We train a CRF model using exactly the same training data as the proposed method. We use the following manually created features:
 - a) the words within a 5-word window;
 - b) the POS tags within a 5-word window;
 - c) the number of characters;
 - d) binary indicators (camel case, digits, dashes, slashes and periods);
 - e) dependency relations for the current word obtained via dependency parsing.
- We use CRF as a baseline to show the performance of a non-deep learning method.
- 2) **S-BLSTM**: This baseline is a traditional S-BLSTM with 2 layers (by removing the bank attention from SAN). It is a supervised baseline. We use this baseline to show that using purely supervised data is not good enough. Unlabeled data can help to improve the performance.
- 3) **SAN (-) BLSTM2**: This baseline does not have the second layer of BLSTM for the labeled question. We use this baseline to show that S-BLSTM works better for our problem. We use 5 unlabeled questions in both this baseline and SAN.

Result Analysis From Table IV, we can see that the proposed SAN framework performs the best on F1-score. Although CRF is a non-deep learning model, its precision is not bad since we use dependency relations as features. However, the recall of CRF is very low since it can only

⁵<http://mallet.cs.umass.edu/>

train weights on words appear in the training data. All deep learning models have better recalls than CRF. S-BLSTM has the best precision as it is trained using only the training data. However, its recall is relatively low. It still suffers the problem that training data can not further tune embeddings of words not appeared in the training data. SAN (-) BLSTM2 shows that the additional BLSTM layer is effective in learning better representations. Lastly, SAN significantly improves the recall by further adjusting the weights for different unlabeled questions. It only loses 0.5% on precision compared that with S-BLSTM.

V. RELATED WORK

Both data mining and natural language processing communities study sentiment analysis on products [9]–[13]. However, Product Community Question and Answering (PCQA) only draws attention in recent years [14], [15]. PCQA is studied as a relevance ranking problem in [14], [15]. Given a question, they retrieve relevant reviews to augment existing answers. Instead, we observe that PCQA also contains valuable fine-grained information for extraction. Product function needs are an important type of such information. Functions may contain both intrinsic functions and extrinsic functions [2]. Extrinsic functions are closely related to complementary products (taking whether one product can work with another as a function) [16]–[18]. But we observe that from the perspective of functionality, how two products can work together is also important. For example, “install Windows 10” and “run Windows 10” are two different functions.

Although CNN [19], [20] and Long Short-Term Memory (LSTM) [3] are both used in NLP tasks, LSTM is more commonly used in sequence labeling [21], [22]. Attention mechanism is popular in image recognition [23], [24]. It is later used in natural language processing [25], [26]. However, attention mechanism is only used in supervised settings. We adapt attention for a semi-supervised setting [27]. Traditional semi-supervised learning uses unlabeled data as training examples [28] directly. Instead, we use unlabeled data as side information for labeled examples.

VI. CONCLUSION

In this paper, we propose the task of Function Need Recognition (FNR), which is to identify function needs queried by customers. We leverage a Semi-supervised Attention Network (SAN) to solve this problem by leveraging unlabeled data as attended side information. Experiments demonstrate that the SAN is better than a number of baselines.

ACKNOWLEDGMENT

This work is supported in part by NSF through grants IIS-1526499, and CNS-1626432, and NSFC 61672313.

REFERENCES

- [1] P. Kotler and G. Armstrong, *Principles of marketing*. pearson education, 2010.
- [2] H. Xu, S. Xie, L. Shu, and P. S. Yu, "Product function need recognition and satisfiability analysis," in *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [4] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [6] S. El Hahi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies." in *NIPS*, vol. 400. Citeseer, 1995, p. 409.
- [7] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [8] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, 2001, pp. 282–289.
- [9] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.
- [10] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 2013, pp. 165–172.
- [11] J. J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *KDD*, 2015.
- [12] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2015.
- [13] L. Shu, H. Xu, and B. Liu, "Lifelong learning crf for supervised aspect extraction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 148–154. [Online]. Available: <http://aclweb.org/anthology/P17-2023>
- [14] J. McAuley and A. Yang, "Addressing complex and subjective product-related queries with customer reviews," in *World Wide Web*, 2016.
- [15] M. Liu, Y. Fang, D. H. Park, X. Hu, and Z. Yu, "Retrieving non-redundant questions to summarize a product review," pp. 385–394, 2016.
- [16] H. Xu, S. Xie, L. Shu, and P. S. Yu, "Cer: Complementary entity recognition via knowledge expansion on large unlabeled product reviews," in *Proceedings of IEEE International Conference on Big Data*, 2016.
- [17] H. Xu, L. Shu, J. Zhang, and P. S. Yu, "Mining compatible/incompatible entities from question and answering via yes/no answer classification using distant label expansion," *arXiv preprint arXiv:1612.04499*, 2016.
- [18] H. Xu, L. Shu, and P. S. Yu, "Supervised complementary entity recognition with augmented key-value pairs of knowledge," *arXiv preprint arXiv:1705.10030*, 2017.
- [19] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.
- [20] L. Shu, H. Xu, and B. Liu, "Doc: Deep open classification of text documents," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 2901–2906. [Online]. Available: <https://www.aclweb.org/anthology/D17-1313>
- [21] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *arXiv preprint arXiv:1503.04069*, 2015.
- [22] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," *arXiv preprint arXiv:1603.01360*, 2016.
- [23] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order boltzmann machine," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1243–1251.
- [24] M. Denil, L. Bazzani, H. Larochelle, and N. de Freitas, "Learning where to attend with deep architectures for image tracking," *Neural computation*, vol. 24, no. 8, pp. 2151–2184, 2012.
- [25] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," in *Advances in Neural Information Processing Systems*, 2015, pp. 1693–1701.
- [26] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." in *ICML*, vol. 14, 2015, pp. 77–81.
- [27] X. Zhu, "Semi-supervised learning literature survey," 2005.
- [28] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 3546–3554.