Graph Enhanced Memory Networks for Sentiment Analysis

Zhao Xu[1], Romain Vial[1,2], and Kristian Kersting[3]

[1] NEC Laboratories Europe, Germany
zhao.xu@neclab.eu
[2] MINES ParisTech, PSL Research University, France
romain.vial@mines-paristech.fr
[3] Technical University of Dortmund, Germany
kristian.kersting@cs.tu-dortmund.de

Abstract. Memory networks model information and knowledge as memories which can be manipulated for prediction, inference and reasoning on the basis of attention mechanism in neural networks. In many cases, there exist complicated relations between memories, by which the memories are linked together into graphs. Typical examples include dependency tree of a sentence and knowledge graph in a dialogue system. In this paper, we present graph enhanced memory networks to integrate the relational information between memories into deep neural networks. Our approach can exploit two types of attentions, graph- and content-based ones, to effectively identify the important memories for the given question, and thus leads to a better inference and reasoning about the final response. We demonstrate the effectiveness of the proposed approach with an interesting application on aspect based sentiment classification. The empirical analysis on real data shows the advantages of incorporating relational dependencies into the memory networks.

1 Introduction

Memory network [42, 36, 12, 21] has recently attracted increasing attention due to its success in many applications, such as machine reading and understanding, visual and textual question answering [2, 44, 15, 45, 39, 13]. In general, a memory network embeds a set of facts and knowledge in vector spaces as memory cells (shorten as memories). Given a question (typically represented with natural language), the model searches the supporting memories, and infers the final answer from the retrieved memories. The major advantage of memory networks is that they introduce an external memory component and the associated computational modules in the neural network framework to explicitly store, update, access, and manipulate the knowledge and facts for prediction, inference and reasoning. The reader and writer functions are fully differentiable such that the entire architectures can be learned end-to-end with backpropagation.

The recent works on memory networks only model the contents of the facts and knowledge. However the relations between them are not taken into account.

In many cases, the facts and knowledge are not independent of each other, but are linked into a graph. The information exists not only in the content of the facts, but also in the relations between them. The importance of relational information has been demonstrated in the literature, see e.g., probabilistic models [7, 10, 27] and neural network based models [3, 6, 19, 43].

In this paper, we propose graph enhanced memory networks (GeMN) to integrate the relational information between memories into (deep) memory networks to enhance attentions. Memory networks use attention mechanism [1, 24] to access and manipulate memories. In the GeMN approach, we introduce an extra attention, *graph attention*, learned from relations between memories to better position the most important memories w.r.t. a given question. We assume that the attention weights of memories follow a Gaussian random field, i.e., a Gaussian distribution with graph Laplacian as kernels [46, 47, 33]. The memories with a short distance on the graph show strong correlation and thus likely have similar relevance with the question. Our approach can exploit two types of attentions, graph- and content-based ones, to effectively identify the important memories for the given question, and thus lead to a better inference and reasoning about the final response. There are few works investigating relational information in memory networks. As far as we know, the related work on structured attention networks [17] is the closest. It models probabilistic dependencies with conditional random field that mainly focuses on sequence structures, rather than relations in general. Our approach introduces relational information into memory networks and explicitly models relations in an elegant way. The graph attention, together with the content-based one, improves inference and reasoning of memory networks.

We apply the proposed method to address aspect level sentiment classification problem. With the exponential growth of user-generated content on online social network services, extracting useful insights such as preferences and opinions of users is of growing interest. Sentiment analysis [23, 29, 5, 34] focuses on detecting opinions and emotions of users on products, services and social events from large collections of texts. Typically the sentiment analysis is to estimate the positive or negative polarity of a given sentence. A more important and complicated task is to extract the sentiment polarity towards aspects [32, 23, 29, 31]. For example, in a customer review on a laptop, *price is ok, but resolution is low!*, there are positive emotion on the aspect *price*, and negative emotion on the aspect *resolution*. Simply classifying the sentence as positive or negative may not properly elicit user's opinions, thus a fine-grained analysis on aspect level sentiment is necessary. We consider a supervised case where the aspects are given. There are different approaches explored in the literature, such as SVM [20, 41, 4], conditional random field [14, 40], and neural networks [8, 38, 39, 28]. Our approach exploits graph- and content-attentions to position the related words (i.e. memories) in a sentence w.r.t. a given aspect, and estimates the aspect level polarity based on the discovered relevant words. The empirical analysis on the real data about customer reviews on laptops and restaurants [31] demonstrates the superiority of the proposed approach.

We start the rest of this paper with a brief review on memory networks, and then introduce the graph enhanced memory networks with the application to aspect based sentiment classification. Before conclusion, the empirical analysis of the proposed approach is presented.

2 Memory Networks

Memory Networks [42, 36, 12, 21] are a class of learning methods with a memory component that can be read and written to for prediction, inference and reasoning. The memory networks typically consist of memories and four computation modules, including I (input), G (generalization), O (output), and R (response). They are defined as follows:

- Memories are an array of objects or facts;
- Input module computes the feature representation of the input;
- Generalization module updates the old memory with the new input;
- Output module produces an output given the input representation and the current memories;
- Response module generates the final response (such as a textual answer to a question) conditioned on the output.

Output module manipulates memories for the final response. In the recent literature, the output module is typically based on the attention mechanism [1, 24]. In particular, the output can be computed as (see e.g. [36]):

$$o = \sum\nolimits_i p_i c_i, \;\; p_i = softmax(f(u, m_i)), \tag{1}$$

where $softmax(x_i) = \exp(x_i)/\sum_j exp(x_j)$. m_i denotes the i-th memory in the input space, and c_i is the vector of the memory i in the output space. u is the vector representation of a question q that can be a question sentence in question-answering, or a target word in machine translation, etc. The output is a weighted sum of output memories. The computation of the weight p_i is based on a function of the question vector u and the input memory m_i. There are some variants of the weight function. The typical ones include:

$$f(u, m_i) = \begin{cases} \exp(u^T m_i) & \text{dot} \\ \exp(u^T A m_i) & \text{general} \\ \exp(v^T \tanh(A[u; m_i]) & \text{concatenation,} \end{cases} \tag{2}$$

where the matrix A and the vector v are parameters to be learned with back-propagation [24].

3 Graph Enhanced Memory Networks for Sentiment Analysis

Attentions play an important role in memory networks. They provide random access to memories conditioned on the features of the question q. The attention mechanism can be viewed as an additional hidden layer to compute a categorical distribution $(p_1, \dots, p_N)$ for soft selection over the number N of memories. However the existing methods in the literature only consider the content of the memories without the relations between the memories (e.g. the dependency tree of a sentence and knowledge graph). It is obvious that integrating the relational information into the learning process will lead to a better attention distribution. To reach the goal, we propose a graph based method, inspired by [46, 47, 33].

We introduce a random variable z_i to each memory, which value specifies to what extent the memory i contributes to the output o. z_i's are not independent of each other, but are interconnected into a weighted graph $\mathcal{G} = (Z, E, W)$. Each memory is denoted as a vertex $z \in Z$ of the graph $\mathcal{G}$, and is linked with undirected and weighted edges $e \in E$.

The weighted graph $\mathcal{G}$ is represented as an adjacency matrix W of size $N \times N$, where N denotes the number of memories. Each entry $W_{i,j}$ represents the weight of an edge $e_{i,j}$ between the memories i and j. Intuitively, the larger the weight, the stronger the correlation between the two memories, and thus the more likely the memories are assigned similar attentions for the output. We formulate the weight as a function of the distance $d_{i,j}$ between i and j on the graph $\mathcal{G}$. The function can be of any form, but non-negative and monotonically decreasing. It can be defined as, e.g.,:

$$\text{Squared exponential:} \quad \exp(-d^2/2\ell^2) \tag{3}$$

$$\text{Rational quadratic:} \quad (1 + d^2/2\alpha\ell^2)^{-\alpha} \tag{4}$$

$$\gamma\text{-exponential:} \quad \exp(-(d/\ell)^\gamma),\ 0 < \gamma \leq 2 \tag{5}$$

With the adjacency matrix, we now model the distribution of z_i's for a soft selection over memories. To be more flexible, it is not necessary to assume the sum-to-one constraint Eq. (1). Thus multiple memories can make significant contributions (analogous to the multi-label classification case). Inspired by [48, 47, 46], the distribution is modeled as Gaussian random field. In particular, the state of z_i is only conditioned on the connected random variables, and follows a Gaussian distribution. The energy, i.e. sum of clique potentials of a Markov random field, can be defined as :

$$E(\mathbf{z}) = \frac{1}{4} \sum_{i,j} W_{i,j} (z_i - z_j)^2. \tag{6}$$

Therefore, the distribution of z_i's is

$$\begin{aligned} p(\mathbf{z}) &\propto \exp\left(-E(\mathbf{z})\right), \\ &= \exp\left(-\frac{1}{2}\mathbf{z}^T \Delta \mathbf{z}\right). \end{aligned} \tag{7}$$

Δ denotes combinatorial graph Laplacian: $\Delta = D - W$, where D is a diagonal degree matrix with $D_{i,i} = \sum_j W_{i,j}$.

We now descript the graph enhanced memory networks with more details. For a better understanding, we illustrate the network structure with an application

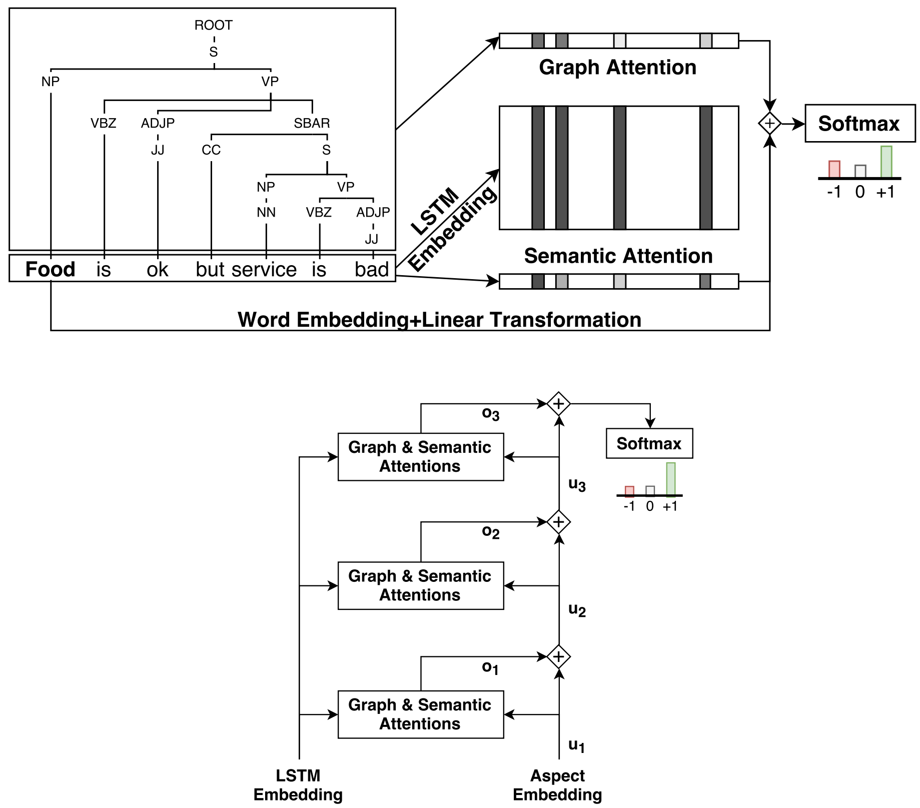

Fig. 1. Graph enhanced memory networks for aspect-based sentiment classification: a single layer version (top) and a multiple layers version (bottom).

of aspect based sentiment classification. The key idea should be applicable in other cases where the memories are linked into graphs.

Assume that there is a sentence consisting of a sequence of words $\{w_1, \ldots, w_N\}$ and multiple aspects $\{a_1, \ldots, a_M\}$. For instance, let consider a guest comment on a restaurant, *the food is ok, but service is bad*, with two aspect words *food* and *service*. The task is to detect aspect level sentiment (i.e., positive emotion on *food* and negative emotion on *service*) by exploiting the sentence itself and the corresponding dependency tree. Here we assume each aspect only involves single word in the sentence (e.g., *food* and *service*). In the case of multi-word aspects, the computation will be similar.

Let start with the single level version of our model, shown as top panel of Fig. 1. The words $\{w_1, \ldots, w_N\}$ of the sentence are formulated as memories using word embedding [26, 30, 22], which are facts to be used to identify sentiment polarities. The memories are mapped into the input space with *long short-term memory* (LSTM) [16, 9, 11, 37]. In particular, the output vector of the LSTM cell, one for each word, is denoted as the input memory m_i. The output memory embedding here is the same as the input one (i.e., $c_i \equiv m_i$). Each aspect word is formulated as a question q, and is mapped as a vector u with word embedding.

Given the aspect word u, the output vector consists of two components: graph-based output o_g and content-based output o_c

$$o_c = \sum\nolimits_i p_i m_i, \; p_i = \tanh(m_i^T A u); \tag{8}$$

$$o_g = \sum\nolimits_i z_i m_i, \; z = \Delta_{m,m}^{-1} W_{m,a} z_a \tag{9}$$

The content based output is computed as usual [36]. The matrix A makes linear mapping from aspect space to input memory space. Here the activation function for computing p_i can be flexible, though the tanh function is theoretically more reasonable (refer to categorical distributions for multi-label classification problem). The graph based output is also computed as a weighted sum of memories. But the attentions are computed with memory graph (dependency tree of the sentence). As z_i's follow a Gaussian distribution Eq. (7), we can then estimate the graph attentions of memories given that of the aspect word. To characterize the properties of the graph attentions explicitly in terms of matrix operations, the distribution is expanded as:

$$\begin{bmatrix} z_a \\ \mathbf{z}_m \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} D_{a,a} - W_{a,a} & -W_{a,m} \\ -W_{m,a} & D_{m,m} - W_{m,m} \end{bmatrix}\right) \tag{10}$$

where z_a denotes the attention of the aspect word, which is always $z_a \equiv 1$ as the word is directly related to the aspect. The vector $\mathbf{z}_m$ denotes the unknown graph attentions of memories. The Laplacian Δ is split into four corresponding blocks for the aspect word and the memory words. Then the maximum likelihood estimation of $\mathbf{z}_m$ conditioned on the attention of the aspect word z_a is:

$$\mathbf{z}_m = \left(D_{m,m} - W_{m,m}\right)^{-1} W_{m,a} z_a. \tag{11}$$

There are different ways to combine the graph- and content-outputs, such as *addition* and *concatenation*.

$$o = \begin{cases} o_c + o_g & \text{addition} \\ B[o_c; o_g] & \text{concatenation} \end{cases} \tag{12}$$

$$s = softmax(C(o + Au)) \tag{13}$$

The parameter matrix B makes a linear transformation from the concatenation space to the memory space. Conditioned on the total outputs, the final response s (i.e., the aspect level sentiment polarity) is estimated by a softmax with a final weight matrix C.

We also extend our model to a multiple level version, shown as the bottom panel of Fig. 1. The structure of the deep network is stacked as follows:

$$u^t = Au^{t-1} + o^{t-1}, \;\; p_i^t = \tanh(m_i^T A u^t), \tag{14}$$

$$o_c^t = \sum\nolimits_i p_i^t m_i, \qquad o^t = o_g + o_c^t, \tag{15}$$

Table 1. Statistics of the datasets

Dataset	Positive	Negative	Neutral
Laptop Train	987	866	460
Laptop Test	341	128	169
Restaurant Train	2164	805	633
Restaurant Test	728	196	196

where the stacking strategy of memory embeddings $\{m_1, \ldots, m_N\}$ is RNN-like, i.e., keeping the memories the same across layers [36]. At the top of the network, the final response is again computed with softmax: $s = softmax(C(o^t + Au^t))$.

4 Experiments

To evaluate the performance of the graph enhanced memory network, we apply the approach to address the aspect-based sentiment classification problem. The experimental analysis is performed on real data with comparison against the state-of-the-art methods.

4.1 Datasets

The data is from the Task 4.2 of SemEval2014 [31], which includes two domain-specific English datasets for laptop and restaurant customer reviews. Each dataset has been manually labeled with annotations at the sentence level. The statistics of the datasets are summarized in Table 1. We follow the settings as in [39] that removes the sentences with the label "conflict" due to the small size of the category. The goal is to predict the aspect level polarity (three polarities: positive, negative and neutral) of a sentence given the labeled aspect terms. Note that one sentence can include multiple aspects. For example, given the sentence "*Great food but the service was dreadful!*" and the aspect terms {"*food*" and "*service*"}, successful predictions would be {"*food*": positive and "*service*": negative}.

4.2 Baselines

The proposed method is compared with multiple recent baselines to evaluate its performance on both datasets. The baselines include:

- *Majority*: assigns to each sentence in the test set the majority sentiment label in the training set.
- *SVM* [20]: is ranked at the 1st (Laptops) and 2nd (Restaurants) places in the SemEval2014 contest. The features used in the method are sophisticated hand-crafted, including n-gram, lexicon and parse features.

Table 2. Classification accuracy of different methods

Baselines	Laptops	Restaurants	GeMN	Laptops	Restaurants
Majority	53.45	65.00	Semantic Attention	70.69	78.84
Feature+SVM	72.10	80.89	Graph Attention	73.51	80.36
LSTM	66.45	74.28	Graph + Semantic (1 hop)	73.82	80.00
TDLSTM	68.13	75.63	Graph + Semantic (2 hops)	**74.29**	80.71
TDLSTM+ATT	66.24	74.31	Graph + Semantic (3 hops)	73.20	80.18
MemNet(1)	67.66	76.10	Graph + Semantic (4 hops)	72.88	80.54
MemNet(3)	71.74	79.06	Graph + Semantic (5 hops)	72.72	80.80
MemNet(5)	71.89	80.14	Graph + Semantic (6 hops)	72.41	**81.43**
MemNet(7)	72.37	80.32	Graph + Semantic (7 hops)	72.72	80.09
MemNet(9)	72.21	80.95	Graph + Semantic (8 hops)	72.26	80.62

- Three LSTM based models [38]: the *LSTM* method directly uses the output vector of the LSTM cell for the last word of a sentence as input of a softmax to estimate the sentiment polarity. The *TDLSTM* method extends the LSTM to consider the content similarity with the aspect words. The *TDLSTM+ATT* method further extends the TDLSTM with the attention mechanism.
- *MemNet* [39]: uses several layers of attentions over the word embeddings. In the experiments, MemNet(k) denotes that the model uses k layers of attentions.

4.3 Quantitative Analysis

We first perform quantitative analysis of the proposed method: predicting aspect level sentiment polarity (positive, negative and neutral) for each test sentence. The performance is measured with classification accuracy.

The graph of a sentence, used in the proposed approach, is extracted with Stanford's CoreNLP Toolkit [25]. Here we use the constituency tree of a sentence. The adjacency matrix is computed using squared exponential kernel with $\ell = 0.1$. The distance $d_{i,j}$ between two words is defined as the number of edges of the shortest path connecting them. The distance is normalized by the diameter of the sentence tree. The questions (i.e. the aspects) and the words are mapped as 300-dimensional Glove vectors [30], and the weights of the embedding matrix are freezed during training. The LSTM is then used to map each word in a sentence into a 128-dimensional memory space. We use an aggressive dropout of 0.7 before the final softmax layer to prevent the model from overfitting [35]. Dropout of 0.5 and 0.3 are respectively used at the input nodes and the recurrent connections of the LSTM cells. The optimization is done with Adam method [18]. The learning rate is set to 0.05. The model learns during 10 epochs with a batch size of 16 sentences.

To get detailed performance of the proposed approach, we consider different ways to compute the total output vector for the final softmax layer:

- Variant 1: only models semantic attentions, $o \equiv o_c$. In this case we do not use any information extracted from the graph of the sentence.

- Variant 2: only models graph attentions, $o \equiv o_g$. Here the content based information is ignored.
- Variant 3: combines both attentions, $o \equiv o_c + o_g$.
- Variant 4: models stacked and combined attentions (k hops).

The experimental results are summarized in Table 2. For a fair comparison, we directly use the results of the baselines reported in [39]. Our approach using both graph and semantic attentions refined over multiple layers considerably outperforms the baselines on the Laptop dataset, and provide (slightly) better results on the Restaurant dataset. It is interesting to note that our approach with only graph attentions performs rather well. It means that the dependency tree structures of the sentences contain already meaningful information about which words are relevant in the given context. The semantic information, which may not be fully contained in the parse tree, can further improve the predictions. Therefore, combining both graph and semantic attentions leads to a notable gain in accuracy. Stacking several layers to get a deep network performs well, e.g., a 0.47% increase in accuracy on the Laptops dataset. In summary, the empirical results demonstrate that, as the reltional information reveals additional correlations among attention candidates, the proposed graph attention helps the memory model to focus on the important memories w.r.t. the given aspects and thus offers better predictions.

4.4 Qualitative Analysis

To better understand the performance of the proposed approach, we further analyse the computed attentions and reveal interesting insights. Figure 2 and Figure 3 show two example sentences with one and two aspects, respectively. One can see the structure of the sentences and the computed attention weights with respect to the aspect word. On one hand, graph attention, which only models the sentence structure, appears to effectively identify the important words related to the context (aspect), and assigns them high weights. On the other hand, semantic attention only considers the meanings of the words without syntactic clues. As a consequence, it highlights all the sentiment keywords, even if it is not related to the context. For example, "fresh" is actually not for the aspect "food" in Fig. 2, and "ideal" not for the aspect "weight" in Fig. 3. The combined attention (graph + semantic) takes advantage of both syntactic and semantic information to identify in-context words, and thus better discovers the important words w.r.t. the given aspects. One can see that the words "delicious" (Fig. 2 for the aspect "food"), "acceptable" (Fig. 3 for the aspect "weight") and "ideal" (Fig. 3 for the aspect "size") capture the major part of the attention with a score of 0.51, 0.33 and 0.44 respectively.

We also perform analysis on the effectiveness of a deep structure (i.e., multiple layers attentions). As shown in Fig. 4, one can find that one layer of combined attention is not always enough to handle complex sentences. With more layers, the attention weights appear to be refined at each pass and gradually focus on the important word. For example, the word "delicious" has its score increasing from 0.41 to 0.59, and simultaneously the noise caused by other words is reduced.

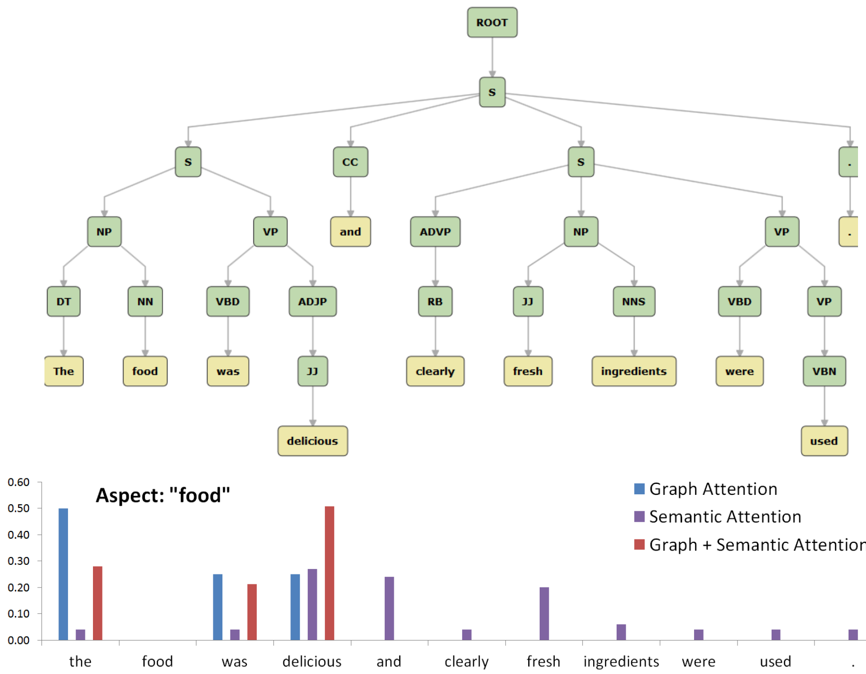

Fig. 2. Example sentence with single aspect: sentence tree structure (top) and computed attention weights (bottom).

5 Conclusion

In this paper we present a graph enhanced memory network to integrate relational information between memories to improve prediction, inference and reasoning. We propose graph attentions learned with graph kernels to identify the most related memories w.r.t. a given question, together with content-based attentions, for inference of the final response. The proposed approach is applied to aspect-based sentiment classification and demonstrates superior performance on real data. Our work provides interesting avenues for future work, such as graph enhanced memory networks for question-answering and knowledge graph reasoning.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations (2015)
2. Bordes, A., Usunier, N., Chopra, S., Weston, J.: Large-scale simple question answering with memory networks. ArXiv preprint:1506.02075 (2015)

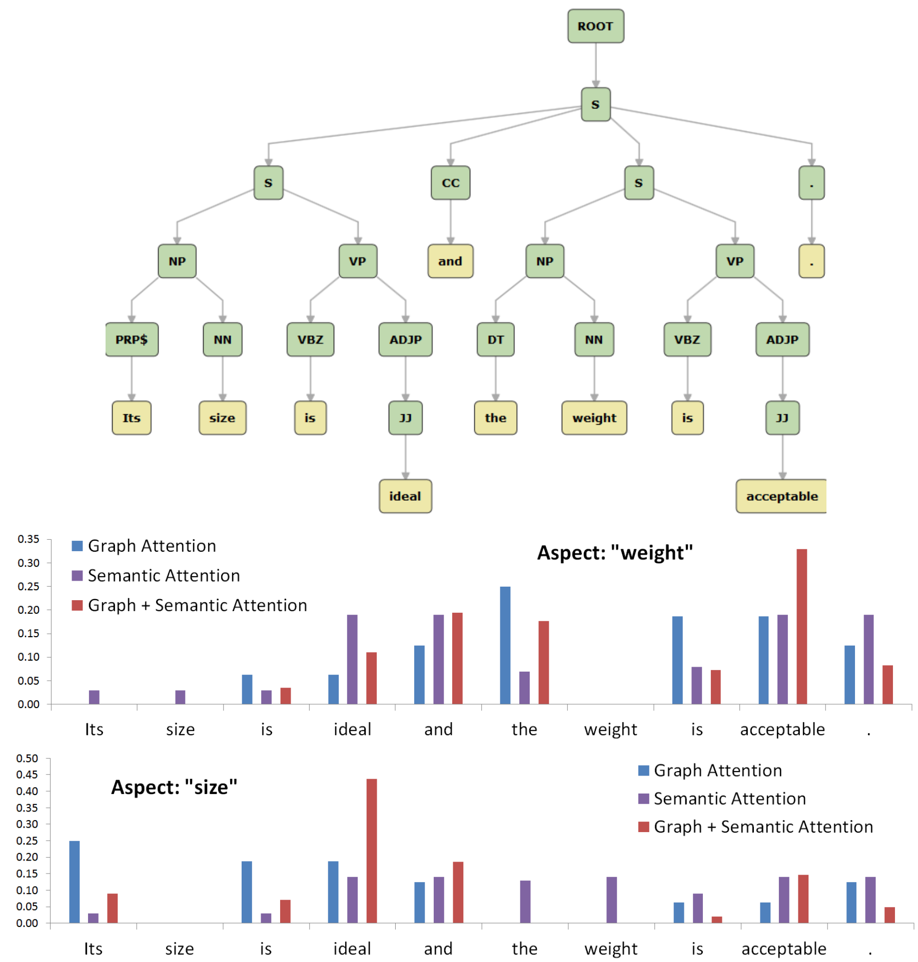

Fig. 3. Example sentence with two aspects: sentence tree structure (top) and computed attention weights (bottom).

3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26. pp. 2787–2795 (2013)
4. Brychcin, T., Konkol, M., Steinberger, J.: Uwb: Machine learning approach to aspect-based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 817–822 (2014)
5. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in Neural Information Processing Systems 28 (2015)
6. Dai, H., Dai, B., Song, L.: Discriminative embeddings of latent variable models for structured data. In: ICML (2016)

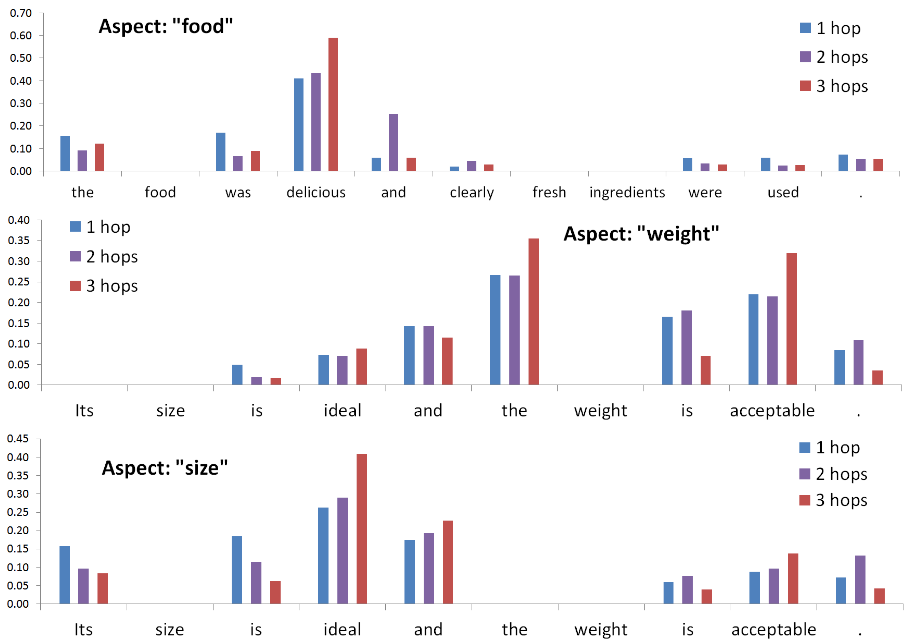

Fig. 4. Examples of the predicted attention weights at each hop.

7. De Raedt, L., Kersting, K., Natarajan, S., Poole, D.: Statistical Relational Artificial Intelligence: Logic, Probability, and Computation. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers (2016)
8. Dong, L., Wei, F., Tan, C., Tang, D., et al.: Adaptive recursive neural network for target-dependent twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (2014)
9. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm. Neural Computation 12(10), 2451–2471 (2000)
10. Getoor, L., Taskar, B. (eds.): Introduction to Statistical Relational Learning. MIT Press (2007)
11. Graves, A.: Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, Springer (2012)
12. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. ArXiv preprint:1410.5401 (2014)
13. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., et al.: Hybrid computing using a neural network with dynamic external memory. Nature 538, 471476 (2016)
14. Hamdan, H., Bellot, P., Bechet, F.: Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation. pp. 719–724 (2015)
15. Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children's books with explicit memory representations. ArXiv preprint:1511.02301 (2015)

16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
17. Kim, Y., Denton, C., Hoang, L., Rush, A.M.: Structured attention networks. In: Proceedings of the International Conference on Learning Representations (2017)
18. Kingma, D., Ba, J.: Adam: A method for stochastic optimisation. In: Proceedings of the International Conference on Learning Representations (2015)
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: Proceedings of the International Conference on Learning Representations (2017)
20. Kiritchenko, S., Zhu, X., Cherry, C., Mohammad, S.: Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 437–442 (2014)
21. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: Proceedings of the International Conference on Machine Learning (2016)
22. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning. pp. 1188–1196 (2014)
23. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
24. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attentionbased neural machine translation. In: Proceedings of the Conference on Empirical Methods in NLP (2015)
25. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., et al.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014)
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26. pp. 3111–3119 (2013)
27. Miller, K.T., Griffiths, T.L., Jordan, M.I.: Nonparametric latent feature models for link prediction. In: Advances in Neural Information Processing Systems 22. p. 12761284 (2009)
28. Nguyen, T.H., Shirai, K.: Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. p. 25092514 (2015)
29. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
30. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1532–1543 (2014)
31. Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., et al.: Semeval-2014 task 4: Aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 27–35 (2014)
32. Schouten, K., Frasincar, F.: Survey on aspect-level sentiment analysis. IEEE Trans. Knowledge and Data Engineering 28(3), 813–830 (2016)
33. Smola, A.J., Kondor, R.: Kernels and regularization on graphs. In: Proceedings of the conference on Learning Theory (2003)
34. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP. pp. 1631–1642 (2013)

35. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15(1), 1929–1958 (2014)
36. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. In: NIPS. pp. 2431–2439 (2015)
37. Sutskever, I.: Training Recurrent Neural Networks. Ph.D. thesis, University of Toronto (2013)
38. Tang, D., Qin, B., Feng, X., Liu, T.: Target-dependent sentiment classification with long short term memory. ArXiv preprint:1512.01100 (2015)
39. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. In: EMNLP. pp. 214–224 (2016)
40. Toh, Z., Wang, W.: Dlirec: Aspect term extraction and term polarity classification system. In: Proceedings of the 8th International Workshop on Semantic Evaluation. pp. 235–240 (2014)
41. Wagner, J., Arora, P., Cortes, S., et al.: Dcu: Aspect-based polarity classification for semeval task 4. In: Proceedings of the 8th International Workshop on Semantic Evaluation. p. 223229 (2014)
42. Weston, J., Chopra, S., Bordes, A.: Memory networks. In: ICLR (2015)
43. Xie, R., Liu, Z., Sun, M.: Representation learning of knowledge graphs with hierarchical types. In: IJCAI (2016)
44. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: ICML (2016)
45. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Proceedings of the European Conference on Computer Vision (2016)
46. Xu, Z., Kersting, K., Tresp, V.: Multi-relational learning with gaussian processes. In: IJCAI. pp. 1309–1314 (2009)
47. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions. In: ICML (2003)
48. Zhu, X., Lafferty, J., Ghahramani, Z.: Semi-supervised learning: From gaussian fields to gaussian processes. Tech. Rep. CMU-CS-03-175, Carnegie Mellon University (2003)