Joint Multi-Label Attention Networks for Social Text Annotation

Hang Dong[1,2], **Wei Wang**[2], **Kaizhu Huang**[3], and **Frans Coenen**[1]

[1]Dept. of Computer Science, University of Liverpool
[2]Dept. of Computer Science and Software Engineering, Xi'an Jiaotong-Liverpool University
[3]Dept. of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University
{HangDong,Coenen}@liverpool.ac.uk
{Wei.Wang03,Kaizhu.Huang}@xjtlu.edu.cn

Abstract

We propose a novel attention network for document annotation with user-generated tags. The network is designed according to the human reading and annotation behaviour. Usually, users try to digest the title and obtain a rough idea about the topic first, and then read the content of the document. Present research shows that the title metadata could largely affect the social annotation. To better utilise this information, we design a framework that separates the title from the content of a document and apply a title-guided attention mechanism over each sentence in the content. We also propose two semantic-based loss regularisers that enforce the output of the network to conform to label semantics, i.e. similarity and subsumption. We analyse each part of the proposed system with two real-world open datasets on publication and question annotation. The integrated approach, Joint Multi-label Attention Network (JMAN), significantly outperformed the Bidirectional Gated Recurrent Unit (Bi-GRU) by around 13%-26% and the Hierarchical Attention Network (HAN) by around 4%-12% on both datasets, with around 10%-30% reduction of training time.

1 Introduction

Social annotation, or tagging, is a popular functionality allowing users to assign "keywords" to online resources for better semantic search and recommendation (Vander Wal, 2007; Singer et al., 2014; Gedikli and Jannach, 2014). Common socially annotated textual resources include questions, papers, (micro-)blogs, product reviews, etc. In practice, however, only a limited number of resources is annotated with tags. Annotating a large number of documents requires much cognitive effort and can be time-consuming. This has driven research on document annotation based on existing tag sets (Belém et al., 2017; Nie et al., 2014).

Recent studies formalise the automated social annotation task as a multi-label classification problem (Gibaja and Ventura, 2015) and apply deep learning approaches (Li et al., 2016; Huang et al., 2016; Hassan et al., 2018). A strong baseline is the use of Bi-directional RNN (Schuster and Paliwal, 1997) with GRU (Cho et al., 2014) or LSTM (Hochreiter and Schmidhuber, 1997). Another more recent improvement is achieved through Hierarchical Attention Network (HAN) (Yang et al., 2016) which discriminates important words and sentences from others, as adapted in (Hassan et al., 2018) for annotation. These models, however, suffer from two issues: (i) simply scanning over the words and sentences, the models do not fully mimic the way users read and annotate documents, and (ii) semantic relations, similarity and subsumption, among the labels are not considered.

Our model focuses on simulating users' reading and annotation behaviour with attention mechanisms. The title of a document is highly abstract while informative about the topics and has a direct impact on users' annotation choice (Lipczak and Milios, 2010), showing high descriptive capacity and effectiveness for annotation (Figueiredo et al., 2013); the content provides complementary information for annotation. Usually, users firstly read the title, and based on their understanding of the title, proceed to the content of the document. To simulate this behaviour, we propose an attention network with separated inputs (title and content) and parallelled attention layers at both the word-level and the sentence-level. One major distinction to previous approaches is to represent the content with a title-guided attention mechanism; this enables the network to discriminate among sentences based on its understanding of the title.

In addition, in the social context, users tend to annotate documents collectively with tags of various semantic forms and granularities (Peters, 2009; Heymann and Garcia-Molina, 2006). One challenging issue is how to exploit the relations among labels (user-generated tags) (Zhang and Zhou, 2014; Gibaja and Ventura, 2015) to improve the learning performance. Among neural network based methods, a recent attempt is to initialise weights for dedicated neurons in the last layer to memorise the label relations (Kurata et al., 2016; Baker and Korhonen, 2017), however, the limitation is the large number of neurons to be assigned, making it inefficient (or inapplicable) for systems with large number of labels. To incorporate the label semantics inferred from the data or from external knowledge bases into the network, we design two loss regularisers, for similarity and subsumption relations, respectively. The regularisers enforce the output layer of the network to satisfy the semantic constraints of the labels.

2 Proposed Method

We propose a parallelled two-layered attention network that simulates users' reading and annotation behaviour for document annotation. The proposed Joint Multi-label Attention Network (JMAN) approach is depicted in Figure 1. The model inputs the title and content separately into two Bidirectional-RNNs with word-level attention and sentence-level attention mechanisms to capture the important words and sentences. Each target is a multi-hot (as opposed to an one-hot) representation of the labels in the label set $y_d \in \{0,1\}^{|T|}$, where T is a list all labels, "1" indicates that a label appears in the label set of the document d, "0" otherwise. In Figure 1, attention mechanisms are indicated with dotted edges. One key distinction from the HAN model (Yang et al., 2016) is the title-guided sentence-level attention that models the reading order for annotation (the dotted edges linking c_t and c_{ta}). The output layer $s_d = \sigma(W_c c_d + b_c)$, activated with the sigmoid function σ, is further constraint by two loss regularisers, emphasising two types of label relations, similarity and subsumption, respectively.

For the RNN encoder, we apply the Gated Recurrent Unit (GRU) which can capture long term dependencies and is usually more time-efficient than LSTM (Hochreiter and Schmidhuber, 1997) in training. The Bidirectional-GRU (Bi-GRU) encoder (Cho et al., 2014) concatenates the hidden states generated from two GRUs, one reading the words (or sentences) forward and the other reading them backwards. This helps form a more complete understanding of the current word (or sentence).

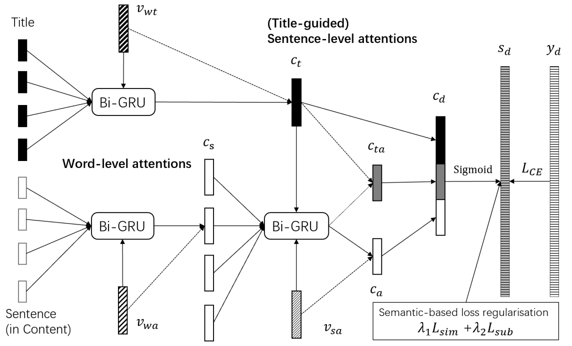

Figure 1: The Proposed Joint Multi-label Attention Network (JMAN) for Social Text Annotation

2.1 Hierarchical Attention

Hierarchical Attention captures the structure of a document by a word-level attention on each word's hidden state to create a sentence representation, then a sentence-level attention to form a content representation (Yang et al., 2016). The attention coefficients are computed based on the dot product between a non-linearly transformed weight vector of the hidden state and an "informative" vector, which encodes "what is the most informative word (or sentence)" in the sequence. This "informative" vector is commonly treated as a sequence of weights (Yang et al., 2016; Kumar et al., 2018; Hassan et al., 2018), trained along with other weights in the network. We applied parallelled word-level attention on the title and each sentence in the content. The attention coefficient and the final representation of a sequence is calculated as (taking words in title as an example):

$$c_t = \sum_i \alpha_i h_i = \sum_i \frac{\exp(v_{wt} \bullet v_i)}{\sum_j \exp(v_{wt} \bullet v_j)} h_i \tag{1}$$

where $v_i = \tanh(W_t h_i + b_t)$ is the output of a fully-connected layer of the hidden state h_i for each word in the title, v_{wt} is the "informative" vector for titles, and c_t is the resulting title representation. We can compute each sentence representation c_s and the content representation c_a in a similar manner (see Figure 1).

2.2 Title-guided Sentence-level Attention

The attention mechanisms above do not capture the interaction between the title and content of the document. Title represents highly abstract while important information about the topics of a document. Selection of the important sentences in the content should conform to the document's general topic, e.g. title. We can thus model the title-guided sentence-level attention as:

$$c_{ta} = \sum_r \alpha_r h_r = \sum_r \frac{\exp(c_t \bullet v_r)}{\sum_k \exp(c_t \bullet v_k)} h_r \qquad (2)$$

where $v_r = \tanh(W_s h_r + b_s)$ is a fully connected layer with the hidden state of the rth sentence h_r as input and c_t is the title representation obtained from Equation 1.

Guiding sentence reading through title representation facilitates content understanding, but may lead to an overemphasis on the title in the annotation. In fact, the content itself, carrying more terms, conveys detailed information not covered by the title and may help suggest further tags for annotation (Figueiredo et al., 2013). We thus concatenate the title guided content representation c_{ta} and the content representation c_a from the original sentence-label attention, to form a more comprehensive representation of the content. The final content representation is then concatenated with the title representation $c_d = [c_t, c_{ta}, c_a]$. In the experiment, we will show the effectiveness of this design against several variations of the model.

2.3 Semantic-based Loss Regularisers

Users tend to annotate documents collectively with semantically related tags. Two major semantic relations in user-generated tags are similarity and subsumption (Stock, 2010; Peters, 2009). To deal with this label correlation issue, we propose two loss regularisers jointly learned with the binary cross entropy loss function. The intuition is that the output values of the neural network s_d, having the dimensions as the label space $|T|$, should satisfy semantic relations among labels. Such relations can be inferred from the label sets or observed in external knowledge bases. The whole joint loss is defined as $L = L_{CE} + \lambda_1 L_{sim} + \lambda_2 L_{sub}$. L_{CE} is the binary cross entropy loss adopted for multi-label text classification (Nam et al., 2014). L_{sim} and L_{sub} are defined as:

$$\begin{aligned} L_{sim} &= \frac{1}{2}\sum_d \sum_{(j,k)|T_j,T_k \in y_d} Sim_{jk}|s_{dj} - s_{dk}|^2 \\ L_{sub} &= \frac{1}{2}\sum_d \sum_{(j,k)|T_j,T_k \in y_d} Sub_{jk} R(s_{dj})(1 - R(s_{dk})) \end{aligned} \qquad (3)$$

where y_d is the label set (annotated tags) of the document d. T is a list of all labels, where j and k are the indices of the list T, corresponding to the indices of nodes s_{dj} and s_{dk} in the output layer s_d. $R()$ is the rounding function for binary prediction, $R(s_{dj}) = 0$ if $S_{dj} < 0.5$, otherwise $R(s_{dj}) = 1$.

The similarity matrix $Sim \in (0,1)^{|T|*|T|}$ indicates pairwise similarity between labels, the larger the value of Sim_{jk}, the more similar the labels T_j and T_k are. Each element Sub_{jk} in the subsumption matrix $Sub \in \{0,1\}^{|T|*|T|}$ indicates whether the label T_j is a child label of T_k. Both the Sim and Sub matrix can be inferred from the training data or from external knowledge bases before training. In implementation, Sim (if thresholded) and Sub can be treated as sparse matrix to reduce computational complexity. We also used an adapted version of the loss regularisers in mini-batch training (the same set of label pairs that co-occurred within all documents in the same batch) to further to reduce computational complexity.

The rationale is that the less the difference of the two outputs of the similar labels is, the lower the L_{sim}. On the contrary, for output values not reflecting the label similarity, i.e. large $|s_{dj} - s_{dk}|^2$ when Sim_{jk} is close to 1, the error will be penalised with higher L_{sim}.

Given a document and a subsumption pair of labels, if the child label is used for annotation, its parent label has a relatively higher chance being used as well. In L_{sub}, if a subsumption relation $< T_j \rightarrow T_k >$ presents in the label set y_d, the case that the parent label T_k is predicted as false, i.e. $R(s_{dk}) = 0$, when its child label T_j is predicted as true, i.e. $R(s_{dj}) = 1$, will be penalised. Such a case will result in a positive penalty, while the penalty will be 0 in all other cases.

Thus, L_{sim} constrains similar labels to have similar outputs, while L_{sub} reinforces each co-occurring subsumption pair to satisfy the dependency of the parent label on the child label.

3 Experiments

3.1 Datasets

We evaluate our proposed approach for automated social annotation on two representative open datasets in social tagging, Bibsonomy[1] (academic publication annotation) and Zhihu[2] (general domain social question annotation). For Bibsonomy, we used the cleaned dataset from (Dong et al.,

[1] https://www.kde.cs.uni-kassel.de/bibsonomy/dumps
[2] https://biendata.com/competition/zhihu/

Bibsonomy	Precision	Recall	F_1 Score	Time/Fold
Bi-GRU	.522±.020*	.217±.016*	.306±.019*	1480±92s
HAN	.572±.008*	.246±.012*	.344±.013*	1164±52s
JMAN-s-tg	.591±.010	.269±.006*	.370±.007*	1075±87s
JMAN-s-att	.586±.009	.269±.005*	.369±.006*	968±81s
JMAN-s	.586±.004	.282±.005	.380±.005	**894±55s**
JMAN	**.592±.009**	**.284±.006**	**.384±.007**	1044±73s

* Paired t-tests at 95 percent significance level against the JMAN model.

Table 1: Comparison Results on the Bibsonomy dataset

Zhihu	Precision	Recall	F_1 Score	Time/Fold
Bi-GRU	.238±.011*	.154±.009*	.187±.010*	1455±69s
HAN	.257±.012	.167±.010*	.203±.011*	1387±78s
JMAN-s-tg	.257±.005	.175±.003*	.208±.006**	1220±81s
JMAN-s-att	.254±.007**	.174±.005*	.207±.005*	1275±99s
JMAN-s	.257±.008	.177±.005	.210±.007	1147±44s
JMAN	**.260±.006**	**.179±.003**	**.212±.004**	**1135±52s**

* Paired t-tests at 95 percent significance level against the JMAN model.
** Paired t-tests at 90 percent significance level against the JMAN model.

Table 2: Comparison Results on the Zhihu dataset

2017) and further selected the tags related to Computer Sciences according to the ACM Computing Classification System[3] and selected the document that have both title and abstract (content); for Zhihu, we randomly sampled around 100,000 questions from the original data dump.

The cleaned Bibsonomy dataset has 12,101 documents, 17,619 vocabularies and 5,196 labels; the average number of labels per document is 11.59. The sample Zhihu dataset has 108,168 documents (questions), 62,519 vocabularies and 1,999 labels; the average number of labels per document is 2.45.

3.2 Implementation Details

To calculate Sim, we used cosine similarity, normalised to between 0 and 1, of self-trained skip-gram embedding (Mikolov et al., 2013) on all label sets in each dataset. To obtain Sub, about subsumption relations, for Bibsonomy, we resorted to an external knowledge source Microsoft Concept Graph[4] for label mapping and semantic grounding; for Zhihu, we used the provided crowd-sourced label subsumption relations. We tuned the λ_1 and λ_2 in L based on 10-fold cross-validation[5].

We implemented the proposed Joint Multi-label Attention Network (JMAN) model in Figure 1 on Tensorflow (Abadi et al., 2016) along with the baselines[6] based on brightmart's implementation[7] of TextRNN and HAN under the MIT license. Two strong baselines were chosen **Bi-GRU** (Schuster and Paliwal, 1997; Cho et al., 2014) and **HAN** (Yang et al., 2016; Hassan et al., 2018). Several variations of **JMAN** were also considered: (i) **JMAN-s**, the proposed model without semantic-based loss regularisers; (ii) **JMAN-s-tg**, the proposed model without semantic-based regularisers and title guided sentence-level attention, $c_d = [c_t, c_a]$; (iii) **JMAN-s-att**, the proposed model without semantic-based regularisers and the original sentence-level attention, $c_d = [c_t, c_{ta}]$.

We optimised the joint loss L using the Adam optimiser (Kingma and Ba, 2014) and set the number of hidden units as 100, learning rate as 0.01 and dropout rate as 0.5 (Srivastava et al., 2014) for all models. The batch sizes for Bibsonomy and Zhihu were set as 128 and 1,024, respectively. The sequence lengths of the title (also the length of each sentence) and the content were padded to 30 and 300 for Bibsonomy and 25 and 100 for Zhihu. Non-static input embedding for the title and the sentences were initialised as 100-dimension self-trained skip-gram embedding (Mikolov et al.,

[3] https://www.acm.org/publications/class-2012
[4] https://concept.research.microsoft.com/Home
[5] λ_1, λ_2 were tuned to 1e-4, 1e-1 for Bibsonomy and 1e-3, 1e-1 for Zhihu, respectively.
[6] Our code and datasets are available at https://github.com/acadTags/Automated-Social-Annotation.
[7] https://github.com/brightmart/text_classification

2013). We decayed the learning rate by half when the loss on validation set increased and set an early stopping point when learning rate is below 2e-5. All experiments were run on a GPU server, NVIDIA GeForce GTX 1080 Ti.

3.3 Results

We report the mean and the standard deviation of the testing results on models trained with 10-fold cross-validation. The cleaned user-generated tags, i.e. labels, for each dataset were taken as the ground truth and the widely used example-based metrics, Precision, Recall and F_1 score (Godbole and Sarawagi, 2004; Tsoumakas et al., 2010; Zhang and Zhou, 2014), were adopted. The average training time per fold was also recorded.

The results with respect to the two datasets are presented in the Table 1 and 2 respectively. Our proposed JMAN model significantly outperforms Bi-GRU and HAN. In terms of F_1, with the Bibsonomy dataset, the proposed JMAN model provides a 7.8% absolute increase (by 25.5%) over Bi-GRU and 4.0% (by 11.6%) over HAN; on the Zhihu dataset, our model is 2.5% absolutely (by 13.4%) better than Bi-GRU and 0.9% (by 4.4%) than HAN. This is mostly attributed to the boost of recall through modeling the title metadata and the title-guided attention mechanism. The JMAN model also converges ("understands") much faster than HAN with around 10.3% (for Bibsonomy) and 18.2% (for Zhihu) less training time per fold and converges even faster than Bi-GRU (by 29.5% and 22.0% for the Bibsonomy and Zhihu dataset in terms of training time, respectively). Recall and F_1 score drop significantly, with training time increased, when the title-guided or the original sentence-level attention is removed. Adding semantic-based loss regularisers further boosts the precision, recall and F_1 of the model.

We also noticed that, compared to the results on the Bibsonomy dataset, the improvement on the Zhihu dataset with the proposed model is less significant. This may be related to the characterstics of the dataset: Zhihu has shorter texts (padded to 1/3 of the Bibsonomy dataset), more vocabularies (over 3 folds), less number of labels (about 40%) and less average number of labels per document (about 1/5) than the Bibsonomy dataset. This would warrant further study on the datasets and on validating the model with datasets from other social media platforms.

4 Conclusion

We proposed a parallelled two-layer attention network for text annotation based on user-generated tags. It models the behaviour how human users read and understand document with the title-guided attention mechanism and leverages label semantics through two loss regularisers to constrain the network outputs. Experimental results show the effectiveness of this method with superior performance and training speed. This system can be applied to various types of social media platforms to support document organisation.

Future studies will explore the possibility of applying the title-guided attention mechanism to other large datasets on major social media platforms. It is also interesting to see whether the semantic-based loss regularisers can be adapted to improve the performance of the recent pre-trained transferable deep learning models, such as the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018).

Acknowledgments

We thank all the anonymous reviewers for their constructive feedback. The implementation is based on brightmart's TextRNN and Hierarchical Attention Network under the MIT license[8]. This research is funded by the Research Development Fund at Xi'an Jiaotong-Liverpool University, contract number RDF-14-01-10 and partially supported by the following: The National Natural Science Foundation of China under no. 61876155; The Natural Science Foundation of the Jiangsu Higher Education Institutions of China under no. 17KJD520010; Suzhou Science and Technology Program under no. SYG201712, SZS201613; Natural Science Foundation of Jiangsu Province BK20181189; Key Program Special Fund in XJTLU under no. KSF-A-01, KSF-P-02, and KSF-E-26.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016.

[8] https://github.com/brightmart/text_classification

Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA. USENIX Association.

Simon Baker and Anna Korhonen. 2017. Initializing neural networks for hierarchical multi-label text classification. *BioNLP 2017*, pages 307–315.

Fabiano M Belém, Jussara M Almeida, and Marcos A Gonçalves. 2017. A survey on tag recommendation methods. *Journal of the Association for Information Science and Technology*, 68(4):830–844.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hang Dong, Wei Wang, and Coenen Frans. 2017. Deriving dynamic knowledge from academic social tagging data: a novel research direction. In *iConference 2017 Proceedings*, pages 661–666. iSchools.

Flavio Figueiredo, Henrique Pinto, Fabiano Belm, Jussara Almeida, Marcos Gonalves, David Fernandes, and Edleno Moura. 2013. Assessing the quality of textual features in social media. *Information Processing & Management*, 49(1):222 – 247.

Fatih Gedikli and Dietmar Jannach. 2014. Recommender systems, semantic-based. In *Encyclopedia of Social Network Analysis and Mining*, pages 1501–1510, New York, NY. Springer New York.

Eva Gibaja and Sebastián Ventura. 2015. A tutorial on multilabel learning. *ACM Computing Survey*, 47(3):52:1–52:38.

Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hebatallah A. Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, and Alessandro Micarelli. 2018. Semantic-based tag recommendation in scientific bookmarking systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, pages 465–469, New York, NY, USA. ACM.

Paul Heymann and Hector Garcia-Molina. 2006. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford InfoLab.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Haoran Huang, Qi Zhang, Yeyun Gong, and Xuanjing Huang. 2016. Hashtag recommendation using end-to-end memory networks with hierarchical attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 943–952.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Abhishek Kumar, Daisuke Kawahara, and Sadao Kurohashi. 2018. Knowledge-enriched two-layered attention network for sentiment analysis. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 253–258.

Gakuto Kurata, Bing Xiang, and Bowen Zhou. 2016. Improved neural network-based multi-label classification with better initialization leveraging label cooccurrence. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–526.

Yang Li, Ting Liu, Jing Jiang, and Liang Zhang. 2016. Hashtag recommendation with topical attention-based lstm. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3019–3029.

Marek Lipczak and Evangelos Milios. 2010. The impact of resource title on tags in collaborative tagging systems. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 179–188, New York, NY, USA. ACM.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification — revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452, Berlin, Heidelberg. Springer Berlin Heidelberg.

Liqiang Nie, Yi-Liang Zhao, Xiangyu Wang, Jialie Shen, and Tat-Seng Chua. 2014. Learning to recommend descriptive tags for questions in social forums. *ACM Transactions on Information Systems (TOIS)*, 32(1):5:1–5:23.

Isabella Peters. 2009. Knowledge representation in Web 2.0: Folksonomies. In *Folksonomies. Indexing and Retrieval in Web 2.0*, Knowledge and Information, pages 153–282. De Gruyter.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Philipp Singer, Thomas Niebler, Andreas Hotho, and Markus Strohmaier. 2014. Folksonomies. In *Encyclopedia of Social Network Analysis and Mining*, pages 542–547, New York, NY. Springer New York.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Wolfgang G. Stock. 2010. Concepts and semantic relations in information science. *Journal of the American Society for Information Science and Technology*, 61(10):1951–1969.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining multi-label data. In *Data Mining and Knowl. Discovery Handbook*, pages 667–685, Boston, MA. Springer US.

Thomas Vander Wal. 2007. Folksonomy. `http://vanderwal.net/folksonomy.html`. [Online; accessed 7-March-2019].

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837.