

Improving Distributed Representations of Tweets - Present and Future

Ganesh Jawahar

Information Retrieval and Extraction Laboratory

IIIT Hyderabad

Telangana, India

ganesh.j@research.iiit.ac.in

Abstract

Unsupervised representation learning for tweets is an important research field which helps in solving several business applications such as sentiment analysis, hashtag prediction, paraphrase detection and microblog ranking. A good tweet representation learning model must handle the idiosyncratic nature of tweets which poses several challenges such as short length, informal words, unusual grammar and misspellings. However, there is a lack of prior work which surveys the representation learning models with a focus on tweets. In this work, we organize the models based on its objective function which aids the understanding of the literature. We also provide interesting future directions, which we believe are fruitful in advancing this field by building high-quality tweet representation learning models.

1 Introduction

Twitter is a widely used microblogging platform, where users post and interact with messages, “tweets”. Understanding the semantic representation of tweets can benefit a plethora of applications such as sentiment analysis (Ren et al., 2016; Giachanou and Crestani, 2016), hashtag prediction (Dhingra et al., 2016), paraphrase detection (Vosoughi et al., 2016) and microblog ranking (Huang et al., 2013; Shen et al., 2014). However, tweets are difficult to model as they pose several challenges such as short length, informal words, unusual grammar and misspellings. Recently, researchers are focusing on leveraging unsupervised representation learning methods based on neural networks to solve this problem. Once these representations are learned, we can use off-

the-shelf predictors taking the representation as input to solve the downstream task (Bengio, 2013a; Bengio et al., 2013b). These methods enjoy several advantages: (1) they are cheaper to train, as they work with unlabelled data, (2) they reduce the dependence on domain level experts, and (3) they are highly effective across multiple applications, in practice.

Despite this, there is a lack of prior work which surveys the tweet-specific unsupervised representation learning models. In this work, we attempt to fill this gap by investigating the models in an organized fashion. Specifically, we group the models based on the objective function it optimizes. We believe this work can aid the understanding of the existing literature. We conclude the paper by presenting interesting future research directions, which we believe are fruitful in advancing this field by building high-quality tweet representation learning models.

2 Unsupervised Tweet Representation Models

There are various models spanning across different model architectures and objective functions in the literature to compute tweet representation in an unsupervised fashion. These models work in a semi-supervised way - the representations generated by the model is fed to an off-the-shelf predictor like Support Vector Machines (SVM) to solve a particular downstream task. These models span across a wide variety of neural network based architectures including average of word vectors, convolutional-based, recurrent-based and so on. We believe that the performance of these models is highly dependent on the objective function it optimizes – predicting adjacent word (within-tweet relationships), adjacent tweet (inter-tweet relationships), the tweet itself (autoencoder), mod-

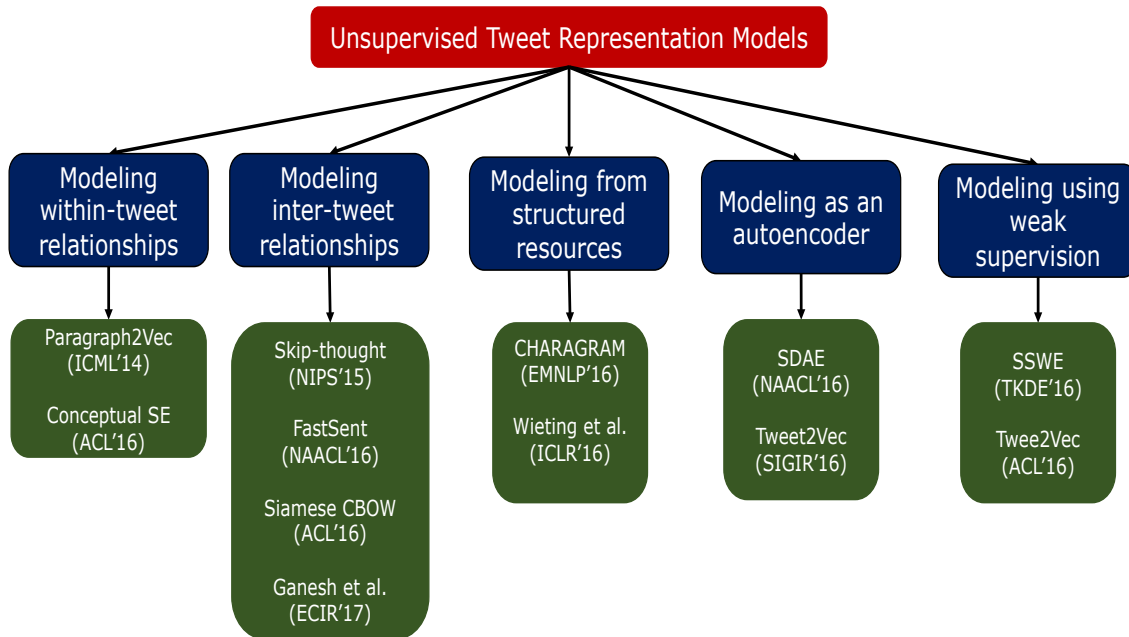


Figure 1: Unsupervised Tweet Representation Models Hierarchy based on Optimized Objective Function

eling from structured resources like paraphrase databases and weak supervision. In this section, we provide the first of its kind survey of the recent tweet-specific unsupervised models in an organized fashion to understand the literature. Specifically, we categorize each model based on the optimized objective function as shown in Figure 1. Next, we study each category one by one.

2.1 Modeling within-tweet relationships

Motivation: Every tweet is assumed to have a latent topic vector, which influences the distribution of the words in the tweet. For example, though the appearance of the phrase *catch the ball* is frequent in the corpus, if we know that the topic of a tweet is about “technology”, we can expect words such as *bug* or *exception* after the word *catch* (ignoring *the*) instead of the word *ball* since *catch the bug/exception* is more plausible under the topic “technology”. On the other hand, if the topic of the tweet is about “sports”, then we can expect *ball* after *catch*. These intuitions indicate that the prediction of neighboring words for a given word strongly relies on the tweet also.

Models: (Le and Mikolov, 2014)’s work is the first to exploit this idea to compute distributed document representations that are good at predicting words in the document. They propose two models: PV-DM and PV-DBOW, that are extensions of Continuous Bag Of Words (CBOW) and Skip-

gram model variants of the popular Word2Vec model (Mikolov et al., 2013) respectively – PV-DM inserts an additional document token (which can be thought of as another word) which is shared across all contexts generated from the same document; PV-DBOW attempts to predict the sampled words from the document given the document representation. Although originally employed for paragraphs and documents, these models work better than the traditional models: BOW (Harris, 1954) and LDA (Blei et al., 2003) for tweet classification and microblog retrieval tasks (Wang et al., 2016). The authors in (Wang et al., 2016) make the PV-DM and PV-DBOW models *concept-aware* (a rich semantic signal from a tweet) by augmenting two features: attention over contextual words and conceptual tweet embedding, which jointly exploit concept-level senses of tweets to compute better representations. Both the discussed works have the following characteristics: (1) they use a shallow architecture, which enables fast training, (2) computing representations for test tweets requires computing gradients, which is time-consuming for real-time Twitter applications, and (3) most importantly, they fail to exploit textual information from related tweets that can bear salient semantic signals.

2.2 Modeling inter-tweet relationships

Motivation: To capture rich tweet semantics, researchers are attempting to exploit a type of *sentence-level Distributional Hypothesis* (Harris, 1954; Polajnar et al., 2015). The idea is to infer the tweet representation from the content of adjacent tweets in a related stream like users' Twitter timeline, topical, retweet and conversational stream. This approach significantly alleviates the context insufficiency problem caused due to the ambiguous and short nature of tweets (Ren et al., 2016; Ganesh et al., 2017).

Models: Skip-thought vectors (Kiros et al., 2015) (STV) is a widely popular sentence encoder, which is trained to predict adjacent sentences in the book corpus (Zhu et al., 2015). Although the testing is cheap as it involves a cheap forward propagation of the test sentence, STV is very slow to train thanks to its complicated model architecture. To combat this computational inefficiency, FastSent (Hill et al., 2016) propose a simple additive (log-linear) sentence model, which predicts adjacent sentences (represented as BOW) taking the BOW representation of some sentence in context. This model can exploit the same signal, but at a much lower computational expense. Parallel to this work, Siamase CBOW (Kenter et al., 2016) develop a model which directly compares the BOW representation of two sentence to bring the embeddings of a sentence closer to its adjacent sentence, away from a randomly occurring sentence in the corpus. For FastSent and Siamese CBOW, the test sentence representation is a simple average of word vectors obtained after training. Both of these models are general purpose sentence representation models trained on book corpus, yet give a competitive performance over previous models on the tweet semantic similarity computation task. (Ganesh et al., 2017)'s model attempt to exploit these signals directly from Twitter. With the help of attention technique and learned user representation, this log-linear model is able to capture salient semantic information from chronologically adjacent tweets of a target tweet in users' Twitter timeline.

2.3 Modeling from structured resources

Motivation: In recent times, building representation models based on supervision from richly structured resources such as Paraphrase Database (PPDB) (Ganitkevitch et al., 2013) (containing

noisy phrase pairs) has yielded high quality sentence representations. These methods work by maximizing the similarity of the sentences in the learned semantic space.

Models: CHARAGRAM (Wieting et al., 2016a) embeds textual sequences by learning a character-based compositional model that involves addition of the vectors of its character n-grams followed by an elementwise nonlinearity. This simpler architecture trained on PPDB is able to beat models with complex architectures like CNN, LSTM on SemEval 2015 Twitter textual similarity task by a large margin. This result emphasizes the importance of character-level models that address differences due to spelling variation and word choice. The authors in their subsequent work (Wieting et al., 2016b) conduct a comprehensive analysis of models spanning the range of complexity from word averaging to LSTMs for its ability to do transfer and supervised learning after optimizing a margin based loss on PPDB. For transfer learning, they find models based on word averaging perform well on both the in-domain and out-of-domain textual similarity tasks, beating LSTM model by a large margin. On the other hand, the word averaging models perform well for both sentence similarity and textual entailment tasks, outperforming the LSTM. However, for sentiment classification task, they find LSTM (trained on PPDB) to beat the averaging models to establish a new state of the art. The above results suggest that structured resources play a vital role in computing general-purpose embeddings useful in downstream applications.

2.4 Modeling as an autoencoder

Motivation: The *autoencoder* based approach learns latent (or compressed) representation by reconstructing its own input. Since textual data like tweets contain discrete input signals, *sequence-to-sequence models* (Sutskever et al., 2014) like STV can be used to build the solution. The encoder model which encodes the input tweet can typically be a CNN (Kim, 2014), recurrent models like RNN, GRU, LSTM (Karpathy et al., 2015) or memory networks (Sukhbaatar et al., 2015). The decoder model which generates the output tweet can typically be a recurrent model that predicts a output token at every time step.

Models: Sequential Denoising Autoencoders (SDAE) (Hill et al., 2016) is a LSTM-based sequence-to-sequence model, which is trained to

recover the original data from the corrupted version. SDAE produces robust representations by learning to represent the data in terms of features that explain its important factors of variation. Tweet2Vec (Vosoughi et al., 2016) is a recent model which uses a character-level CNN-LSTM encoder-decoder architecture trained to construct the input tweet directly. This model outperforms competitive models that work on word-level like PV-DM, PV-DBOW on semantic similarity computation and sentiment classification tasks, thereby showing that the character-level nature of Tweet2Vec is best-suited to deal with the noise and idiosyncrasies of tweets. Tweet2Vec controls the generalization error by using a data augmentation technique, wherein tweets are replicated and some of the words in the replicated tweets are replaced with their synonyms. Both SDAE and Tweet2Vec has the advantage that they don't need a coherent inter-sentence narrative (like STV), which is hard to obtain in Twitter.

2.5 Modeling using weak supervision

Motivation: In a weakly supervised setup, we create labels for a tweet automatically and predict them to learn potentially sophisticated models than those obtained by unsupervised learning alone. Examples of labels include sentiment of the overall tweet, words like hashtag present in the tweet and so on. This technique can create a *huge* labeled dataset especially for building data-hungry, sophisticated deep learning models.

Models: (Tang et al., 2016) learns sentiment-specific word embedding (SSWE), which encodes the polarity information in the word representations so that words with contrasting polarities and similar syntactic context (like *good* and *bad*) are pushed away from each other in the semantic space that it learns. SSWE utilizes the massive distant-supervised tweets collected by positive and negative emoticons to build a powerful tweet representation, which are shown to be useful in tasks such as sentiment classification and word similarity computation in sentiment lexicon. (Dhingra et al., 2016) observes that *hashtags* in tweets can be considered as topics and hence tweets with similar hashtags must come closer to each other. Their model predicts the hashtags by using a Bi-GRU layer to embed the tweets from its characters. Due to subword modeling, such character-level models can approximate the representations

for rare words and new words (words not seen during training) in the test tweets really well. This model outperforms the word-level baselines for hashtag prediction task, thereby concluding that exploring character-level models for tweets is a worthy research direction to pursue. Both these works fail to study the model's generality (Weston et al., 2014), i.e., the ability of the model to transfer the learned representations to diverse tasks.

3 Future Directions

In this section we present the future research directions which we believe can be worth pursuing to generate high quality tweet embeddings.

- (Ren et al., 2016) propose a supervised neural network utilizing contextualized features from conversation, author and topic based context about a target tweet to perform well in classification of tweet. Apart from (Ganesh et al., 2017)'s work which utilizes author context, there is no other work which builds unsupervised tweet representation model on Twitter-specific contexts such as conversation and topical streams. We believe such a solution directly exploits semantic signals (or nuances) from Twitter, unlike STV or Siamese CBOW which are trained on books corpus.
- (dos Santos and Gatti, 2014) propose a supervised, hybrid model exploiting both the character and word level information for Twitter sentiment analysis task. Since the settings when the character level model beats the word level model is not well understood yet, we believe it would be interesting to explore such a hybrid compositional model to build unsupervised tweet representations.
- Twitter provides a platform for the users to interact with other users. To the best of our knowledge, there is no related work that computes unsupervised tweet representation by exploiting the user profile attributes like profile picture, user biography and set of followers, and social interactions like retweet context (set of surrounding tweets in a users retweet stream) and favorite context (set of surrounding tweets in a users favorite tweet stream).

- DSSM (Huang et al., 2013; Shen et al., 2014) propose a family of deep models that are trained to maximize the relevance of clicked documents given a query. Such a ranking loss function helps the model cater to a wide variety of applications¹ such as web search ranking, ad selection/relevance, question answering, knowledge inference and machine translation. We observe such a loss function has not been explored for building unsupervised tweet representations. We believe employing a ranking loss directly on tweets using a large scale microblog dataset² can result in representations which can be useful to Twitter applications beyond those studied in the tweet representation learning literature.
- Linguists assume that language is best understood as a hierarchical tree of phrases, rather than a flat sequence of words or characters. It's difficult to get the syntactic trees for tweets as most of them are not grammatically correct. The average of word vectors model has the most simplest compositional architecture with no additional parameters, yet displays a strong performance outperforming complex architectures such as CNN, LSTM and so on for several downstream applications (Wieting et al., 2016a,b). We believe a theoretical understanding of why word averaging models perform well can help in embracing these models by linguists.
- Models in (Wieting et al., 2016a,b) learn from noisy phrase pairs of PPDB. Note that the source of the underlying texts is completely different from Twitter. It can be interesting to see the effectiveness of such models when directly trained on structural resources from Twitter like Twitter Paraphrase Corpus (Xu et al., 2014). The main challenge with this approach is the small size of the annotated Twitter resources, which can encourage models like (Arora et al., 2017) that work well even when the training data is scarce or nonexistent.
- Tweets mostly have an accompanying image which sometimes has visual correspondence with its textual content (Chen et al.,

¹<https://www.microsoft.com/en-us/research/project/dssm/>

²<http://trec.nist.gov/>

2013; Wang et al., 2014) ('visual' tweet). To the best of our knowledge, there is no work which explores the following question: *can we build multimodal representations for tweets accompanying correlated visual content and compare with traditional benchmarks?* We can leverage insights from multimodal skip-gram model (Lazaridou et al., 2015) which builds multimodally-enhanced word vectors that perform well in the traditional semantic benchmarks. However, it's hard to detect visual tweets and learning from a non-visual tweet can degrade its tweet representation. It would be interesting to see if a dispersion metric (Kiela et al., 2014) for tweets can be explored to overcome this problem of building a nondegradable, improved tweet representation.

- Interpreting the tweet representations to unearth the encoded features responsible for its performance on a downstream task is an important, but a less studied research area. (Ganesh et al., 2016)'s work is the first to open the blackbox of vector embeddings for tweets. They propose elementary property prediction tasks which predicts the accuracy to which a given tweet representation encodes the elementary property (like slang words, hashtags, mentions, etc). The main drawback of the work is that they fail to correlate their study with downstream applications. We believe performing such a correlation study can clearly highlight the set of elementary features behind the performance of a particular representation model over other for a given downstream task.

4 Conclusion

In this work we study the problem of learning unsupervised tweet representations. We believe our survey of the existing works based on the objective function can give vital perspectives to researchers and aid their understanding of the field. We also believe the future research directions studied in this work can help in breaking the barriers in building high quality, general purpose tweet representation models.

References

- Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *CoRR*.
- Yoshua Bengio. 2013a. Deep Learning of Representations: Looking Forward. In *Proc. of the 1st Intl. Conf. on Statistical Language and Speech Processing*. pages 1–37.
- Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. 2013b. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1798–1828.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.
- Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. 2013. Understanding and classifying image tweets. In *ACM Multimedia Conference, MM '13*. pages 781–784.
- Bhuvan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W. Cohen. 2016. Tweet2Vec: Character-Based Distributed Representations for Social Media. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*.
- Cícero Nogueira dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *25th Intl. Conf. on Computational Linguistics*. pages 69–78.
- J Ganesh, Manish Gupta, and Vasudeva Varma. 2016. Interpreting the syntactic and social elements of the tweet representations via elementary property prediction tasks. *CoRR* abs/1611.04887.
- J Ganesh, Manish Gupta, and Vasudeva Varma. 2017. Improving tweet representations using temporal and user context. In *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR*. pages 575–581.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: the paraphrase database. In *Proc. of conf. of the North American Chapter of the Association of Computational Linguistics*. pages 758–764.
- Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys (CSUR)* 49(2):28.
- Zellig S Harris. 1954. Distributional Structure. *Word* 10(2-3):146–162.
- Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1367–1377.
- Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *Proc. of the 22nd ACM Intl. Conf. on Information and Knowledge Management*. pages 2333–2338.
- Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and understanding recurrent networks. *CoRR* abs/1506.02078.
- Tom Kenter, Alexey Borisov, and Maarten de Rijke. 2016. Siamese CBOW: Optimizing Word Embeddings for Sentence Representations. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*.
- Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*. pages 835–841.
- Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing*. pages 1746–1751.
- Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Proc. of the 2015 Intl. Conf. on Advances in Neural Information Processing Systems*. pages 3294–3302.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proc. of conf. of the North American Chapter of the Association of Computational Linguistics*. pages 153–163.
- Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proc. of the 31st Intl. Conf. on Machine Learning*. pages 1188–1196.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proc. of the 26th Intl. Conf. on Neural Information Processing Systems*. pages 3111–3119.
- Tamara Polajnar, Laura Rimell, and Stephen Clark. 2015. An exploration of discourse-based sentence spaces for compositional distributional semantics. In *Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*. page 1.

- Yafeng Ren, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Proc. of the 13th AAAI Conference on Artificial Intelligence*. pages 215–221.
- Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval. In *Proc. of the 23rd ACM Intl. Conf. on Conference on Information and Knowledge Management*. pages 101–110.
- Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*. pages 2440–2448.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*. pages 3104–3112.
- Duyu Tang, Furu Wei, Bing Qin, Nan Yang, Ting Liu, and Ming Zhou. 2016. Sentiment Embeddings with Applications to Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering* 28(2):496–509.
- Soroush Vosoughi, Prashanth Vijayaraghavan, and Deb Roy. 2016. Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder. In *Proc. of the 39th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*. pages 1041–1044.
- Yashen Wang, Heyan Huang, Chong Feng, Qiang Zhou, Jiahui Gu, and Xiong Gao. 2016. CSE: conceptual sentence embeddings based on attention model. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics*.
- Zhiyu Wang, Peng Cui, Lexing Xie, Wenwu Zhu, Yong Rui, and Shiqiang Yang. 2014. Bilateral correspondence model for words-and-pictures association in multimedia-rich microblogs. *TOMCCAP* 10(4):34:1–34:21.
- Jason Weston, Sumit Chopra, and Keith Adams. 2014. #tagspace: Semantic embeddings from hashtags. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing*. pages 1822–1827.
- John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016a. Charagram: Embedding words and sentences via character n-grams. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP*. pages 1504–1515.
- John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016b. Towards universal paraphrastic sentence embeddings. *CoRR* abs/1511.08198.
- Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from twitter. *TACL* 2:435–448.
- Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *IEEE Intl. Conf. on Computer Vision, ICCV*. pages 19–27.