Commonsense for Making Sense of Data

Sreyasi Nag Chowdhury
Supervised by Prof. Dr. Gerhard Weikum
Max Planck Institute for Informatics
Saarbrücken, Germany
sreyasi@mpi-inf.mpg.de

ABSTRACT

In my doctoral research, I address the problem of automatically acquiring commonsense knowledge from text corpora and also from data-sets containing visuals (images, videos) along with textual descriptions. I also aim to exploit the acquired commonsense knowledge for domain-specific and domain-independent applications such as fine-grained search, retrieval and prediction, data integration and analytics using qualitative reasoning.

1. INTRODUCTION

Motivation: Commonsense knowledge (CSK) can be defined as a set of facts that human beings inherently use for analysis and decision making in their daily activities; for example - *heavy objects always fall to the ground.* "Intelligent Machines" therefore also need to be equipped with similar latent general knowledge in order to be of meaningful assistance to humans. With the introduction of ubiquitous devices like smart phones and wearable devices, intelligent applications and personal assistance systems have become the need of the hour. For example, a smart phone should be able to turn on the silent mode if the user is in a concert hall (which it can potentially detect from the location settings and calender events). However, since it does not 'know' that *people get disturbed if a phone rings at a concert hall'*, it is unable to do so. CSK of the form *(phone, make, noise), (concert, requires, silence), (noise, oppositeOf, silence)* create an inference chain to conclude that the phone needs to be switched off/put on silent mode. Taking an example from a more data-centric perspective, recommender systems may benefit from CSK to make the recommendations more personal. For example a recommender system equipped with CSK such as *(Indians, love, spicy food), (Indians, dislike, alcohol), (wine, contains, alcohol)* would recommend its customer to buy spices rather than wine as a gift for his/her Indian friend. Such instances introduce the need for the integration of CSK into everyday computer applications.

Proceedings of the VLDB 2016 PhD Workshop, September 9, 2016. New Delhi, India.

Opportunity and Challenges: Challenging problems which would benefit from Commonsense Knowledge are natural language understanding and machine translation tasks, object/scene recognition or interpretation, fine-grained search, retrieval and prediction applications, intelligent assistant systems and household robotics, to name a few. Improving the performance of each of these applications by CSK would eventually pave the way for achieving the greater goal of producing "intelligent machines".

Unfortunately, the acquisition and canonicalization of CSK is the foremost bottleneck. Firstly, this is because the traditional source of CSK acquisition is text. However, text is prone to omission of useful trivia simply because these are too obvious to state in the written form and humans automatically make assumptions while reading. For example, a text piece about a person failing an examination may not often mention that the person is sad as a result. This makes it impossible to gather CSK like *(failure, causeFor, sadness).* Interestingly, images express what text may not. So, sophisticated computer vision mechanisms like object, scene, and emotion detection on images and videos may allow for capture of more human-like CSK. Secondly, the definition of "commonsense" lies in a somewhat grey area since these inherent knowledge may depend on socio-cultural background of a person. Since standard or general-purpose CSK knowledgebases may not be able to bridge socio-cultural gaps, it could be interesting to curate domain-specific or culture-specific CSK knowledgebases.

Approach: To overcome the shortcomings of CSK acquisition from text, I want to look into integrating visual cues from images/videos for curating new CSK knowledgebases or enhancing existing ones. I would also like to investigate the effect and importance of CSK in domain-specific applications. To emulate a human-like understanding of the digital world my effort would be to bridge the gaps between text and visuals through commonsense knowledge.

2. RELATED WORK

Existing Commonsense Knowledge Bases: Early efforts to consolidate a database of CSK was mostly manual, either by experts [8] or through crowd-sourcing [20]. Since such manual creation of a knowledge base is expensive, the paradigm gradually shifted to automatic acquisition of CSK from text corpora [11, 25] or the web [23].

NLP and Computer Vision: Existing research on automatic image annotations [28], description generation [27, 16, 14], scene understanding [5], and image extraction through

Detected visual objects: traffic light, bus, person

(a) Correct object detection – may aid in visual search

Detected visual objects: swimming cap, bookshelf, wine bar, firebox

(b) Faulty object detection – will worsen visual search

Figure 1: Examples of visual object detection which may or may not improve search and retrieval

"passionate street music"

"environment friendly sport"

"side effects of mountaineering"

Figure 2: Abstract queries and expected retrieved results

natural language queries [13] point towards the ongoing collaboration of the NLP and CV communities. Although much have been achieved, human level of understanding of both text and visuals is still far-fetched.

Learning CSK from text and vision: To leverage the vast resources of hidden knowledge in visuals, CSK has been acquired from real images [2] as well as from non-photo-realistic abstractions [26]. Visual tasks like verification of relational phrases [18] and also non-visual tasks like fill-in-the-blanks by intelligent agents [10] have used CSK. Because of the challenge in detecting visual content with perfection, learning CSK from visual cues is a difficult problem.

3. ONGOING WORK

Commonsense Knowledge for Visual Search

The boost in the use of social media and the internet have led to a huge collection of images with accompanied text on the web. In spite of the vast expanse of visual content, search and retrieval still depend solely on textual cues. The imperfection in state-of-the-art computer vision mechanisms is one of the reasons for the conservative use of visual cues (Figure 1). Traditional search engines also do not use additional knowledge about the query. Our hypothesis in order to improve search results is that background CSK on query terms can be used along with textual and visual cues. To this end we deploy three different modalities - text, visual cues, and CSK pertaining to the query - as a recipe for efficient search and retrieval .

Inter-related work from the databases, information retrieval, multimedia and computer vision communities have addressed the problem of image retrieval by visual contents [12, 3]. Popular search engines like Google, Bing, Baidu crucially rely on tags, caption, URL string, and adjacent text of the images for this task. Although lately with deep learning fine-grained object detection has been possible [19, 7, 15, 6], these come with uncertainty and cannot be always efficiently used for search and retrieval.

More sophisticated search considering human factors like emotions evoked on the viewer call for the necessity to bridge the gap between query vocabulary and image features (textual and visual). Let us take the following fictive queries as examples: *passionate street music, environment friendly sport, side effects of mountaineering.* These queries contain abstract words like *environment friendly, sport, passionate* which pertain more to emotions than to visual objects like *bicycle, tent, guitar* etc., making it difficult to retrieve relevant images. Figure 2 shows example images which would be considered as relevant results for the corresponding queries. To address this problem we propose an approach that harness CSK. Recent attempts at automatic CSK acquisition have produced huge collections of CSK with regards to properties of commonplace objects [22], relationships [24] and comparisons [23] between entities, activities and their participants [21]. We believe that use of these or newly acquired CSK would help bridge the semantic gap between queries and retrieved results. For example CSK such as *(street musicians, havePassion, music), (bicycling, isA, sport), (bicycling, is, eco-friendly), (mountaineering, causeFor, garbage), (garbage, causeFor, pollution), (pollution, destroys, environment)* would successfully associate the queries and the results in Figure 2.

We develop a system architecture – *Know2Look* – to incorporate CSK into image retrieval. It consists of a query processor which expands queries with commonsense, and an answer-ranking component based on statistical language models [30]. Our model unifies three kinds of features: *textual features* from the page context of an image, *visual features* obtained from recognizing fine-grained object classes in an image, and *CSK features* in the form of additional properties of the concepts referred to by query words.

Language Model for Ranking We devise a query-likelihood language model (LM) for ranking images x with regard to a given query q (Equation 1). We assume that a query is simply a set of keywords $q_i (i = 1..L)$. The following equation for a unigram LM can be simply extended to a bigram LM by using word pairs instead of single ones:

$$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x] \qquad (1)$$

where β_{CS} is a hyper-parameter weighing the commonsense features of the expanded query, $P_{CS}[q|x]$ represents

query expansion by CSK, and $P_{smoothed}[q|x]$ takes care of smoothing the results with respect to a background corpus.

Datasets and Experiments Since automatic acquisition of CSK from the web can be costly, we conjecture that noisy subject-predicate-object (SPO) triples extracted through Open Information Extraction [1] may be used as CSK. We use OpenIE tool ReVerb [4] on a corpus of Wikipedia articles to collect ~22,000 assertions. To evaluate *Know2Look* ~50,000 images with descriptions are collected from the following datasets: Flickr30k [29], Pascal Sentences [17], SBU Captioned Photo Dataset [16], and MSCOCO [9]. We compare *Know2Look* to Google search with a set of queries devised from co-occurring Flickr tags. Our initial experiments have produced promising results.

This work is in pursuit of improving search and retrieval by the use of CSK along with textual and visual information in images. Emboldened by the initial results we would like to study the effectiveness and use of various existing commonsense knowledgebases along with our present collection of noisy OpenIE triples through ablation studies. We can also naturally extend this work to search on documents with visual contents (like blog posts).

4. PROSPECTIVE USE CASES

- **Searching Multimedia Content:** In Section 3 we have proposed a framework for improvement of search and retrieval of images by the incorporation of commonsense knowledge for better 'understanding' of query terms. Use of CSK would help tackle abstract queries like the ones shown in Figure 2. A natural extension of this work would be on video search by content.
- **Data Integration and Cleansing:** In the age of Big Data a major challenge is to standardize and consolidate data from various sources in order to make it useful. Commonsense knowledge can act as the glue between two disjoint data sources. We motivate this idea further by the following example. Consider the disjoint data sources in Tables 1[1] and 2[2]. Specialized factual knowledge would be required to make sense of Table 1. However, with CSK such as *(air pollution, causeFor, difficult breathing), (nose masks, usedFor, breathing clean air), (air pollution, causeFor, respiratory ailments)*, it is conceivable to combine the two tables and infer that Delhi has a high air pollution level.
- **Data Analytics with Commonsense Qualitative Reasoning:** As e-market emerge as the most convenient marketplace, comprehensive and personalized analysis of a product's worthiness affects the buyer's choice. Qualitative analysis of market trends with commonsense may help in providing simplistic, yet strong cues for buyers and sellers. Let us consider a scenario where a mother plans a birthday party for her 10-year-old and wants recommendations for food, drinks and activities. Recommender systems that learn merely from purchase data do not work well here, as they do not know which items parents buy for their children and which ones for themselves. CSK can help to overcome this bottleneck. For example, the following CSK triples could be beneficial:

Table 1: PM 2.5 level in cities around the world

City	PM 2.5 (mg/m^3)
Delhi	153
Beijing	56
Singapore	17

Table 2: Traveler experience in cities around the world

City	Tourist's opinion
Delhi	Could hardly breathe in traffic-prone areas; touristic places far away from traffic were relatively low on air pollution
Beijing	Had to use masks on a regular basis; couldn't step outdoors at peak traffic times
Singapore	Comfortably clean air; no troubles for people with respiratory conditions

(children, like, sweet drinks), (apple juice, hasTaste, sweet), (liquor, hasTaste, sour),
(children, like, sweet food), (marshmallows, hasTaste, sweet), (pretzels, hasTaste, salty).
Combining this background knowledge with a database of item-item correlations may be able to generate good recommendations.

5. OUTLOOK

In addition to topics described in the previous sections, I would also like to consider the following interesting problems the course of this PhD:

- Activity commonsense knowledge: *Eventuality Prediction* from video frames – the task of predicting the final series of events/activities to follow given a particular frame or a few consecutive frames of a video.
- Spatial commonsense knowledge: Study the importance of spatial relations (*in, on, left of, behind* etc.) in natural language and whether they can be harnessed in a meaningful way for applications such as real-time video understanding and analysis.
- Domain-specific commonsense knowledgebases: Venturing into useful domains and their intersections may give rise to meaningful applications. For example, the intersection of commonsense knowledge on the domains *traffic* and *environment* may be used for an application dedicated to the concept of *smart-city*. Domain-specific CSK knowledgebases may also be used for prediction tasks.

The bigger vision of this thesis is to pave the way for making commonsense knowledge a natural ingredient in applications, be it naive search and retrieval or more sophisticated intelligent assistance systems. Let us hope that in the near future "commonsense" in machines will be more common[3].

6. REFERENCES

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

[1]data from WHO Report 2014
[2]fictional data
[3]Voltaire, 1694-1778: *"Commonsense is not so common."*

[2] X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1409–1416, 2013.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (CSUR)*, 40(2):5, 2008.

[4] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics, 2011.

[5] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[6] J. Hoffman, S. Guadarrama, E. S. Tzeng, R. Hu, J. Donahue, R. Girshick, T. Darrell, and K. Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014.

[7] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[8] D. B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[10] X. Lin and D. Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2984–2993, 2015.

[11] H. Liu and P. Singh. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[12] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, 2007.

[13] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690, 2014.

[14] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012.

[15] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 20, 2015.

[16] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

[17] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics, 2010.

[18] F. Sadeghi, S. K. Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1456–1464. IEEE, 2015.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu. Open mind common sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems 2002: Coopis, doa, and odbase*, pages 1223–1237. Springer, 2002.

[21] N. Tandon, G. de Melo, A. De, and G. Weikum. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proc. CIKM*, 2015.

[22] N. Tandon, G. de Melo, F. Suchanek, and G. Weikum. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 523–532. ACM, 2014.

[23] N. Tandon, G. de Melo, and G. Weikum. Acquiring comparative commonsense knowledge from the web. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 166–172. AAAI Press, 2014.

[24] N. Tandon, C. D. Hariman, J. Urbani, A. Rohrbach, M. Rohrbach, and G. Weikum. Commonsense in parts: mining part-whole relations from the web and image tags. In *Thirtieth AAAI Conference on Artificial Intelligence*. AAAI, 2015.

[25] N. Tandon, G. Weikum, G. d. Melo, and A. De. Lights, camera, action: Knowledge extraction from movie scripts. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 127–128. International World Wide Web Conferences Steering Committee, 2015.

[26] R. Vedantam, X. Lin, T. Batra, C. Lawrence Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2542–2550, 2015.

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[28] J. K. Wang, F. Yan, A. Aker, and R. Gaizauskas. A poodle or a dog? evaluating automatic image annotation using human descriptions at different levels of granularity. *V&L Net 2014*, page 38, 2014.

[29] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[30] C. Zhai. Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies*, 1(1):1–141, 2008.