Usage Based Tag Enhancement of Images

Balaji Vasan Srinivasan
Big data Experience Lab,
Adobe Research, Bangalore
balsrini@adobe.com

Noman Ahmed Sheikh
Indian Institute of Technology,
Delhi
nomanahmedsheikh11@gmail.com

Roshan Kumar
Indian Institute of Technology,
Kanpur
roshankr1995@gmail.com

Saurabh Verma
Indian Institute of Technology,
Roorkee
saurv4u@gmail.com

Niloy Ganguly
Indian Institute of Technology,
Kharagpur
ganguly.niloy@gmail.com

ABSTRACT

Appropriate tagging of images is at the heart of efficient recommendation and retrieval and is used for indexing the image content. However, existing technologies in image tagging either focus on what the content contains or takes the tags of the entire accompanying text as the image tags. Neither of these are sufficient to get a complete understanding of the images. In this paper, we propose a system that analyzes the usage of an image via its accompanying content and utilizes the tags thus obtained to enhance the tags from a visual algorithm to obtain a deeper understanding of the image. Evaluation based of annotators and using existing metrics from baselines show superior performance.

CCS Concepts

•Information systems → Web searching and information discovery;

Keywords

image tagging, usage content, YAGO

1. INTRODUCTION

A popular English idiom says "An image is worth a thousand words". Content writers always look out for good visual supplements to enrich their content and make it more appealing to their target audience. In the era of data explosion, it is therefore necessary to annotate content (images, video, etc.) with appropriate tags for efficient organization that can be leveraged for such retrieval and recommendation needs. However, the size of visual data on the Web today clearly calls for automatic approach to tag these visual data.

Existing tagging systems work towards capturing the denotational aspects of the image, viz. what the image denotes/contains. These details are either captured via the visual features of the images or by analyzing the accompanying content of the images like in search engines. However, for tagging to be useful, it is important to also capture the connotational aspects viz. how the image is perceived by the consumers by accounting for how the image is used in various places. It can be argued that analyzing the accompanying content captures the image usage, but existing systems use the entire set of tags from the accompanying content to tag the image, which fails to capture the exact usage.

In this work, we try to bridge the gap between denotational and connotational aspects of an image by using a combination of visual features of the image and textual features of accompanying content to improve upon the existing tagging engines. We propose a novel framework to combine the tags derived from usage content with the image tags based on the visual features. We thus achieve a balance between connotational and denotational aspects of an image. We further show that such a combination beats the state-of-the-art tagging engines in our subjective and objective evaluations.

The paper is organized as follows. In Section 2, we describe the existing state of image tagging and position our framework with respect to existing systems. Section 3 introduces the key components of the proposed system. In Section 4 evaluates the different parameters of the proposed system to arrive at the right system configuration. We then compare the performance of the proposed system against existing works via subjective and objective evaluations. Section 5 concludes the paper.

2. RELATED WORK

Tagging and understanding textual content has been widely studied. The most fundamental part of extracting and detecting named entities; the popular one here is the Stanford NLP parser [15]. Once the named entities are identified, they are disambiguated and resolved into various categories [13]. Finally, the inter relationships in the content or hierarchies are identified by a semantic understanding of the understanding text. In these works, the entities in the textual content are typically processed into a rich semantic representation (e.g. [2]) which is utilized to gain the deeper understanding of their inter-relationships.

Yang et al. [28] extract the textual tags based on a nearest-neighbor based approach and utilize the neighbors to extract

International Conference on World Wide Web 2017

DOI: 10.1145/1235

the relationships between entities. Nallapatti et al. [18] use "event threading" to join different pieces of text and identify the undercurrent events in the textual topics. Shahaf et al. estimate the importance and "jitteriness" of the entities in the text and use it to infer the connections between different parts of the textual content.

With the advent of knowledge bases like DBpedia [1], Freebase [3] and YAGO [24], relationships from these sources are used to further enhance the understanding of the textual content. Kuzey et al. [10] resolve temponym based on a YAGO based entity resolution to understand textual content with temporal scopes. They develop an Integer Linear Program that jointly optimizes the mappings to knowledge base for a rounded document representation. Tandon et al. [25] mine activity knowledge from Hollywood narratives to answer questions around these activities. They capture the spatio-temporal context of the topics by constructing multiple graphs to capture relationships among activity frames which is leveraged for effective understanding. However, none of these works aim at understanding content based on multiple cues which is the key challenge in our problem, where we have to combine the content and usage cues in tagging.

There exists a large body of literature in the space of image tagging. Li et al.[12] propose methods for assignment of tags from visual aspects and use them for effective retrieval of images. Once the image is tagged, the relationships with other images have also been used for further enhancing the tags [19] or propagating the tags to neighboring images [6]. Like in the textual tagging, the tags can be enhanced and disambiguated with the help of knowledge base and conceptnets [27]. With the successful emergence of deep learning for image understanding, convolution neural networks have been used to find an intermediary representation *Visual Word2Vec* [9] in order to generate the image tags from this latent space. However, all these works focus on tagging the image from their visual cues/content. In our problem, we capture the usage of the images along with the visual content in the image tags to have a rounded understanding of the image.

One work that is close to the proposed solution framework is by Leong et al. [11], which relies exclusively on accompanying content for mining information relevant to the image. They construct relationships among entities in form of graph edges weighted with multiple factors and the weights learned via an ensemble model. However, they do not use the visual tags of the images to align the accompanying content to the image.

3. USAGE BASED TAGGING

The proposed solution enriches the tags around an image which may not be initially contained in the set of image based tags based on the visual features. Our approach takes as input the image tags (author given and the auto tags) along with the usage content. The content is processed to extract key tag candidates which are then pruned with the help of a knowledge base to come up with a set of enriched candidate tags and their relationships. The relationships are leveraged to extract top tags for the image along with the confidence around the tag. Fig. 1 shows a schematic of the proposed solution framework.

3.1 Entity and Relationship Extraction

The proposed approach starts with disambiguating the accompanying content for ambiguous entities via Ambiverse [7] based off of YAGO and replacing each occurrence of the entity with their disambiguated version. This helps in reducing the ambiguities that can get into the candidate tags. The disambiguated content is extensively parsed to identify all named entities. Here, we use the Stanford NLP Parser [15] for simple entities. For complex entities, we leverage HeidelTime[23] and SUTime [4].

We then establish the relationships between these entities across the entire accompanying content. Note that the image may/may not be relevant to the entirety of the entities in the accompanying text and we address this in Section 3.4. At the end of this step, we have a set of all candidate tags which we shall be considering for our final tags.

We represent these tag candidates in a graph capturing the inter-tag relationships on the edges and the importance of the tags on the nodes of the graph.

3.2 Node Importance Scoring

For each tag candidate extracted in the previous step, a score is assigned based on their importance in the local context. We first calculate the total frequency count of the candidate in the usage content accounting for the co-reference of the candidates via proper nouns by an appropriate co-reference parsing.

For every tag candidate we also compute the average distance of the entity from the root node of the corresponding dependency tree (obtained by passing the accompanying content through a dependency parser[5]). This yields the local relevance of the entity in the subject that is being discussed in the content. The average of the two measures yields the final tag importance (n_i).

3.3 Inter-tag Relationship

We then build the relationships between each tag candidates based on a global knowledge base based on two measures.

In the first measure, we used the Word2Vec [16] model trained on a corpus of Google News dataset with 100 billion words resulting in a final corpus of about 3 million word representations. Word2Vec yields a 300 dimensional vector for every tag candidate identified in our previous step. To measure the relationship between a pair of words, we compute the cosine-similarity between the vectors which captures the semantic closeness of the words in the trained Word2Vec space as described in [16].

In our second measure, we calculate the point-wise mutual information [26] between two entities based on their co-occurrences in the Wikipedia articles as the "co-occurrence" score.

Our final edge weight (e_{ij}) is computed as the average of the two measures. This results in a graph representation of the candidate tags with the edge weights indicating the global relationship between the tags and the node weights indicating their local importance in the usage content.

3.4 Unifier

Often images are accompanied with author tags / tags from the visual features capturing the information contained in the images. To capture the usage tags from the accompanying content, it is important to understand how the tag candidates relate to the visual tags to extract the final set

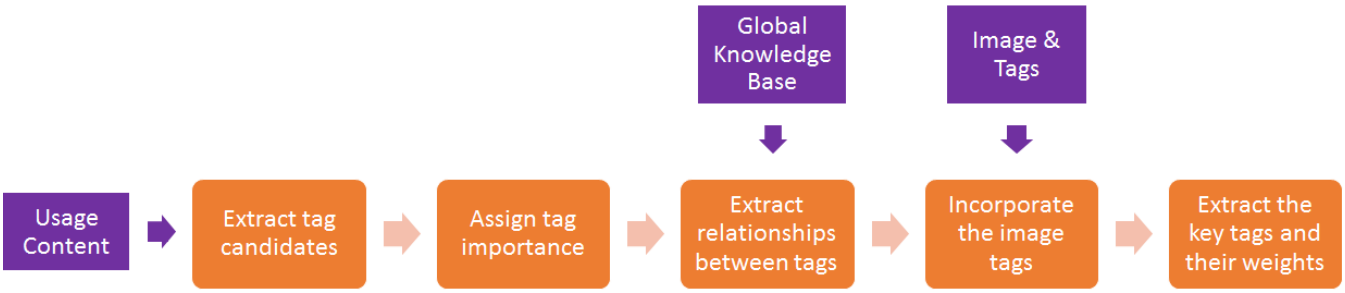


Figure 1: Framework for tag extraction and enhancement in the tagging engine

of tags.

We first add the tags derived from visual features as independent nodes in the tag graph extracted from the previous step. We retain the tag confidence from the visual tagger as the node importance of these visual tags. We calculate the similarity score between the visual tags and tag candidates in the graph based on Section 3.3.

Tag pairs with similarity is greater than a threshold (0.95 in our experiment) are merged into a single node. We further propagate the merged node importance to the adjacent nodes using an exponential decay over the edge weights upto 2 hops. This ensures the propagation of the strength of the merged nodes to its neighbors and thus emphasizing the relevant pieces of the tag graphs with respect to the visual tags.

For the tag pairs less than the matching threshold, an edge is added in our existing graph if this measure is significant ($>$ 0.1 in our experiments). The series of steps is summarized in Algorithm 1.

Algorithm 1 Tag Unifier

```
1: procedure UNIFY(tagsFromImage, TagGraph)
2:     for tag ∈ tagsFromImage do
3:         tag ← normalize(tag)
4:     end for
5:     for node ∈ graph do
6:         addNewNode(tag)
7:         val ← findSimilarity(tag,node)
8:         if val > σ1 then
9:             mergeNodes(tag,node)
10:            node.weight ← amplifiedWeight()
11:            propagateWeight(node)
12:        else if val > σ2 then
13:            edge ← createNewEdge(tag,node)
14:            edge.weight ← val
15:        else
16:            continue
17:        end if
18:    end for
19: end procedure
```

3.5 Tag Extraction

With the graph representation of the tags, the problem of identifying tags that capture the context around the image boils down to identifying the top nodes in the tag graph. For this we use a random walk based algorithm [20], starting the random walk from the visual tags, thus ensuring the node ranking relevant to the tag images and avoiding irrelevant tags from the accompanying text.

We define the probability of the random walk moving from a node i to another node j as,

$$P(tr_{i\to j}) = e_{ij} \times n_j \tag{1}$$

where, e_{ij} is the weight of the edge (from Section 3.3) between tags i and j and n_j is the node importance of tag j from Section 3.2. The probability of the node staying in the same node is defined as

$$P(tr_{i\to i}) = n_i \tag{2}$$

The probabilities above are normalized to conform to the requirements of a probability distribution. The final set of tags is then extracted by performing a random walk for a few iterations starting from the visual/author tag nodes. This ensures that the tags selected are not just based on their importance from the accompanying text but also emphasizes on a strong relationship with the visual tags. The set of tags with weights above a certain threshold is output as the final set of tags for the images.

4. EXPERIMENTAL EVALUATION

We utilized the dataset[1] curated in [11] which contains 300 image-text pairs collected by issuing a query to Google Image API and processing one of the query results that has a significant text around the images. The authors of [11] have created a gold standard tags based on manual annotations from 5 annotators via Amazon Mechanical Turk. We used the datasets along with the gold standard tags for our evaluations. We used the Clarifai API [22] to generate the visual tags for all our experiments.

To extensively evaluate our system, we compare it against the baseline algorithm in [11]. Leong et. al [11] propose 3 independent algorithms - "Wikipedia Salience", "Flickr Picturability" and "Topic Modeling" to extract tags for image from associated textual content. In their experiments, the Wikipedia Salience based tagger was best performing in terms of the precision and recall. We used this algorithm for our evaluations.

4.1 Survey-based evaluation

We first conducted a survey among 45 participants to rate the overall relevance and diversity of the tags on a scale of 010 for the tags from the proposed approach as well as the

[1]http://lit.csci.unt.edu/index.php/index.php?P=research/downloads#TEXT_MINIG"

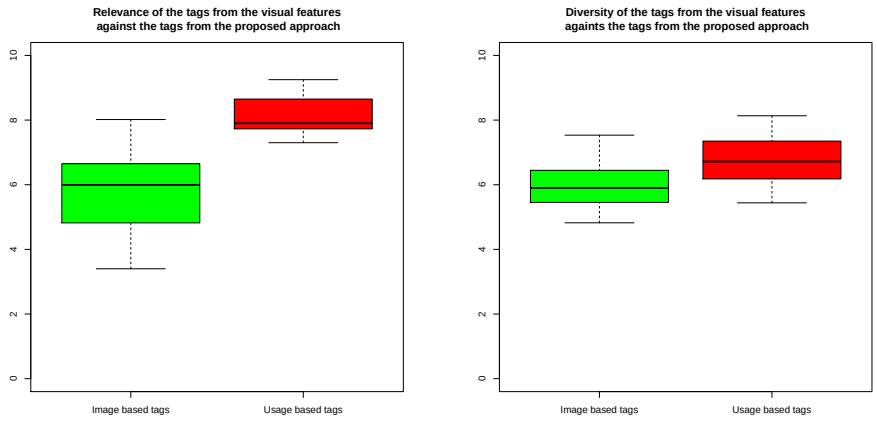


Figure 2: Relevance and diversity of tags based on annotations from 45 participants

tags provided by Clarifai [22] on a subset of 20 images. Fig 2 shows the relevance and diversity of the two tags based on the annotation. It is easy to see that the proposed approach increases the overall relevance of the tags to the image and also performs better in terms of the diversity of the tags indicating the viability of the proposed approach.

4.2 Metrics for evaluation

Human annotations cannot be extended for a comprehensive evaluation of the tags. We therefore extend several existing metrics to measure various aspects of the tags which are described below.

4.2.1 Term Significance [14]

The term-significance [14] is calculated as the significance of the tags to the textual content and is calculated by computing the Normalized Discounted Cumulative Gain(NDCG)[8] over the term frequency of the tags from the usage content normalized based on the tag's inverse document frequency in a global corpus. The intuition here is to compute how important a tag is to the given context (usage) and normalize it with its "commonness" across a bigger corpus (as computed by the idf). We use Wikipedia as the bigger corpus similar to Leong et al. [11].

4.2.2 Relevance Score

The term significance metric purely tests the relevance of the tags to the usage content and is biased towards the system proposed in [11]. It fails to capture the relevance of the tags to the gold standard tags or its overall diversity. We therefore propose two metrics to capture the tag relevance to the image and its overall diversity. To determines how relevant our tags are to the gold standard tags, we compute a weighted cosine similarity between the Word2Vec [16] representation of the extracted tags and the gold tags as given by,

$$sim = \frac{1}{N}\sum_{i}\frac{\sum_{a_j \in TopK(G_i,I_i)} cos(a_j, I_i)\gamma^j}{\sum_j \gamma^j}, \qquad (3)$$

where N is the number of tags generated for the images, I_i is the vector representation of the i^{th} image tag and G_i is the set of all vector representations of the gold standard tags. The inner sum above computes a weighted average of the similarity between the generated tag and the top gold-standard tags. The parameter γ, $(0 \leq \gamma \leq 1)$, penalizes the generation of image tags that is similar to only a small set of the gold standard tags via a decayed-weighted-average. The outer summation calculates the average weighted similarity between the generated image tags and the gold standard tags.

4.2.3 Cophenetic Correlation Coefficient

Finally, for measuring the diversity in the tags, we use the cophenet correlation coefficient [21] (which is a measure of how faithfully a dendogram preserves the pairwise distances between the original un-modeled data points). We perform a hierarchical clustering on the tags based on their Word2Vec representation [17] and compute the cophenet correlation coefficient as the diversity score. Cophenet correlation coefficient is then given by,

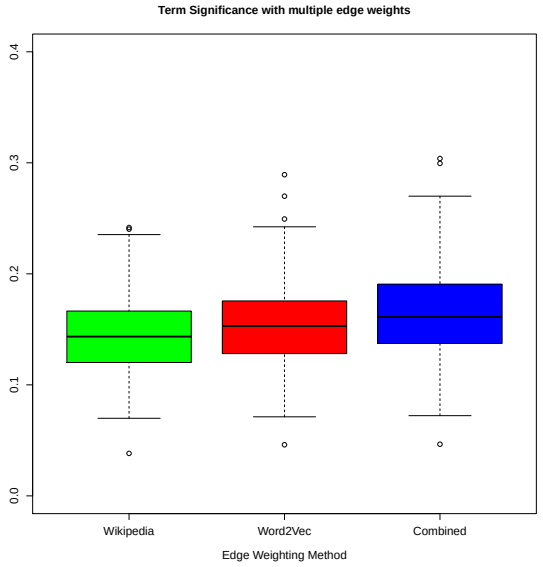
$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}} \qquad (4)$$

where, $x(i,j)$ is the distance between the i^{th} and j^{th} tag. $t(i,j)$ is the height of the node at which the clusters corresponding to ith and jth cluster are first joined together. A higher value of the cophenet correlation coefficient indicates the presence of more significant clusters and hence more tag diversity.

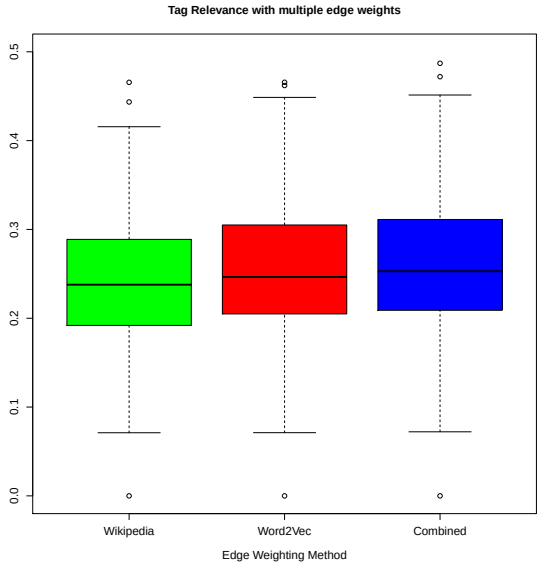
4.3 Effect of edge weights

To analyze the performance of various edge relationships (Section 3.3), we compare the term significance, tag relevance and tag diversity between the edge weighting mechanisms based on Word2Vec, Wikipedia and the combined metric.

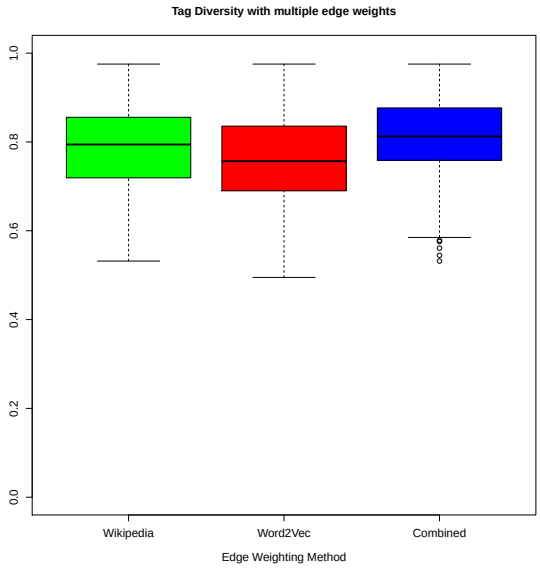
From Fig. 3, it can be seen that while Word2Vec performs marginally better than the Wikipedia based relationship on the scales of term significance ad relevance, Wikipedia based metric is marginally better than Word2Vec in terms of overall tag diversity perhaps because Wikipedia include more entities than the Google News Corpus on which the Word2Vec were trained.



(a) Term Significance



(b) Tag Relevance



(c) Tag Diversity

Figure 3: Term Significance, Tag Relevance (Eq. 3) and Diversity (Eq. 4) for tags based on Gold Annotated Tags from [11], Clarifai [22], Wikipedia Salience [11] and the proposed approach

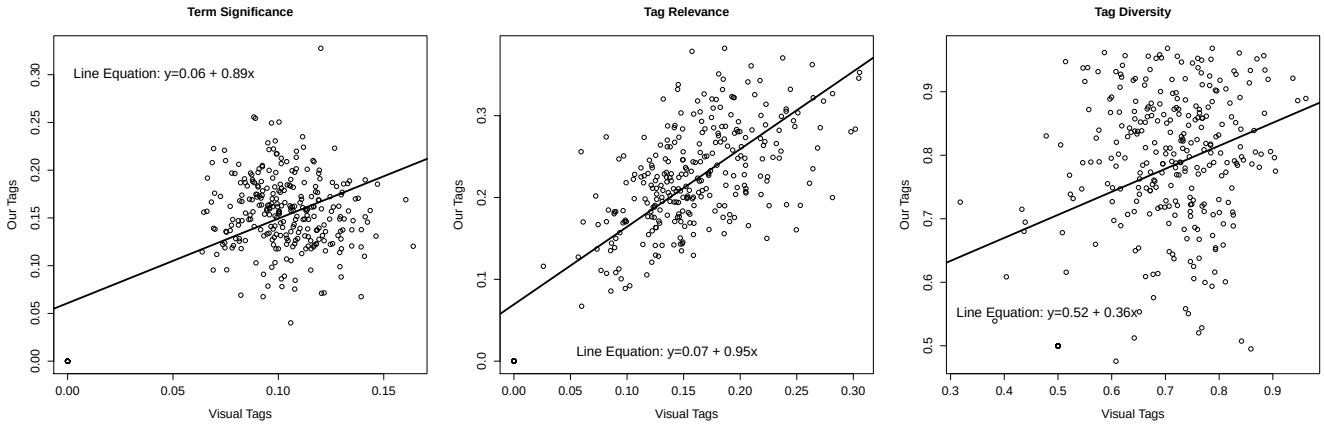


Figure 4: Correlation between the quality of visual tags and the tags from the proposed system

It can also be seen that the combined approach yields the best tags across all metrics and was used for our further comparisons.

4.4 Effect of visual tag quality

We then compared the correlation between the quality of the visual tags and the tags from the proposed system. Fig. 4 shows the correlation between the two set of tags on the scales of Term Significance, Tag Relevance and Tag Diversity.

It can be seen that there is a strong dependence of the term significance and relevance of our tags with the visual tags. This is expected since the algorithm performs the random walk starting from the visual tags and hence the output tag quality is pivoted on the quality of visual tags.

However the tag diversity is less dependent on the visual tags, since the diversity of the output tags is obtained more from the accompanying text than from the visual tags.

4.5 Tagging Performance

Finally, we compare the performance of the proposed tagging system against the visual tagger in [22] and the text based tagger in [11]. Fig. 5 shows the Term Significance, Tag Relevance and Tag Diversity for the tags from the system in [11] compared against our proposed system. As mentioned before, the data set included the gold-standard human annotated tags which were also included in our comparison below.

The term significance (Fig 5(a)) is the best for the proposed method indicating the superiority of the tags from the proposed system. The tags from the proposed system are also more relevant/close to the human annotated tags (Fig. 5(b)) again proving the superior performance. Finally, the tags from the proposed system are also more diverse as indicated by the cophenet correlation based diversity measure.

5. CONCLUSION

In this paper, we have proposed a novel graph based approach to enhance the tags of an image by capturing its usage. We compare the proposed approach against existing baselines in the lights of several quality metrics and improvement is observed. Such a tagging system will serve well to improve the image retrieval based on the user's context.

6. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *DBpedia: A Nucleus for a Web of Open Data*. 2007.

[2] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation (amr) 1.0 specification. In *ACL Conference on Empirical Methods in Natural Language Processing*. ACL, 2012.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD International Conference on Management of data*. ACM, 2008.

[4] A. X. Chang and C. D. Manning. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, 2012.

[5] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

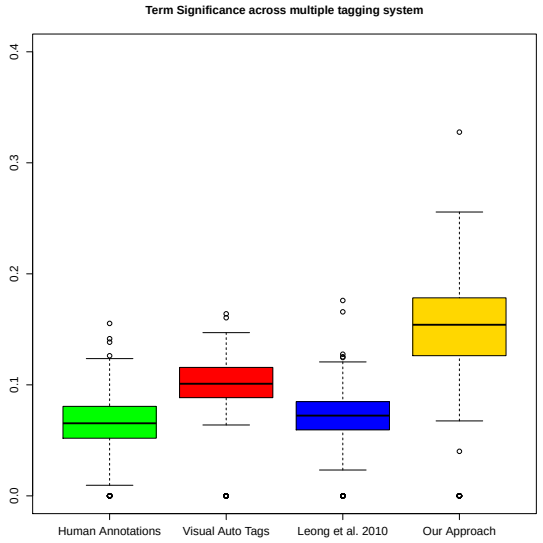
[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *IEEE International Conference on Computer Vision,*, 2009.

[7] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

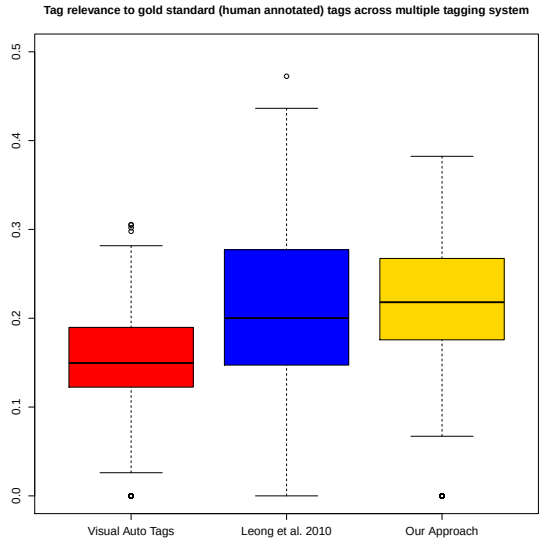
[8] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[9] S. Kottur, R. Vedantam, J. M. Moura, and D. Parikh. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. *arXiv preprint arXiv:1511.07067*, 2015.

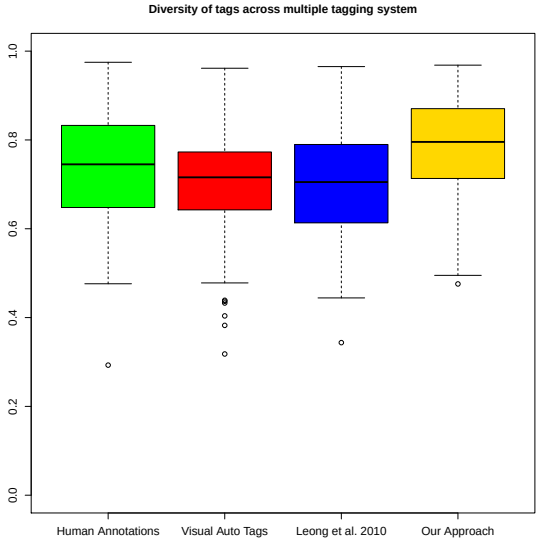
[10] E. Kuzey, V. Setty, J. Strötgen, and G. Weikum. As time goes by: comprehensive tagging of textual phrases with temporal scopes. In *Proceedings of the 25th International Conference on World Wide Web*,



(a) Term Significance



(b) Tag Relevance



(c) Tag Diversity

Figure 5: Term Significance, Tag Relevance (Eq. 3) and Diversity (Eq. 4) for tags based on Gold Annotated Tags from [11], Clarifai [22], Wikipedia Salience [11] and the proposed approach

pages 915–925, 2016.

[11] C. W. Leong, R. Mihalcea, and S. Hassan. Text mining for automatic image tagging. In *International Conference on Computational Linguistics: Posters*, 2010.

[12] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. Del Bimbo. Image tag assignment, refinement and retrieval. In *21st ACM International Conference on Multimedia*, 2015.

[13] M. D. Lieberman and H. Samet. Adaptive context features for toponym resolution in streaming news. In *35th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 731–740. ACM, 2012.

[14] Y.-T. Lu, S.-I. Yu, T.-C. Chang, and J. Y.-j. Hsu. A content-based method to enhance tag recommendation. In *IJCAI*, volume 9, pages 2064–2069, 2009.

[15] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60, 2014.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*. 2013.

[17] T. Mikolov, W. tau Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, May 2013.

[18] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *13th ACM International Conference on Information and Knowledge Management*. ACM, 2004.

[19] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei. Learning semantic relationships for better action retrieval in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015.

[20] Sanjay and D. Kumar. A review paper on page ranking algorithms. *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, Volume 4(6):pp. 2806–2811, 2015.

[21] R. R. Sokal and F. J. Rohlf. The comparison of dendrograms by objective methods. *Taxon*, pages 33–40, 1962.

[22] G. Sood. *clarifai: R Client for the Clarifai API*, 2016. R package version 0.4.0.

[23] J. Strötgen and M. Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 2013.

[24] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In *16th International Conference on World Wide Web*, pages 697–706. ACM, 2007.

[25] N. Tandon, G. de Melo, A. De, and G. Weikum. Knowlywood: Mining activity knowledge from hollywood narratives. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 223–232. ACM, 2015.

[26] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[27] L. Xie and X. He. Picture tags and world knowledge: learning tag relations from visual semantic sources. In *21st ACM International Conference on Multimedia*, pages 967–976. ACM, 2013.

[28] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer. Improving text categorization methods for event tracking. In *23rd ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.