Query Difficulty Prediction for Contextual Image Retrieval

Xing Xing[1], Yi Zhang[1], and Mei Han[2]

[1] School of Engineering, UC Santa Cruz, Santa Cruz, CA 95064
[2] Google Inc., Mountain View, CA 94043

Abstract. This paper explores how to predict query difficulty for contextual image retrieval. We reformulate the problem as the task of predicting how difficult to represent a query as images. We propose to use machine learning algorithms to learn the query difficulty prediction models based on the characteristics of the query words as well as the query context. More specifically, we focus on noun word/phrase queries and propose four features based on several assumptions. We created an evaluation data set by hand and compare several machine learning algorithms on the prediction task. Our preliminary experimental results show the effectiveness of our proposed features and the stable performance using different classification models.

Key words: Query difficulty, Contextual image retrieval

1 Introduction

Given a word/phrase in a document as a query, a contextual image retrieval system tries to return images that match the word/phrase in the given context. Psychological studies for decades [2] have justified the effectiveness of pictorial illustration on improving people's understanding and learning from texts. A contextual image retrieval system annotates the word/phrase in a document with appropriate images and can help readers learn new concepts. For example, a non-English speaker can easily understand the meaning of *panda* if she has seen a picture of it when reading an article that contains the word *panda*.

Although this idea sounds intriguing, image search engines also return useless or even misleading images, either because the image retrieval algorithm is not perfect, or because the query is inherently hard to be represented by images. For example, *honesty* is very difficult to be explained by an image. Some people may suggest a picture of little Washington with a chopped-down cherry tree, while others may disagree with it for the reason that it may be only a legend instead of fact or users may have no idea about this story. If the word in the context cannot be represented by an image, it is better not to annotate it with images, thus avoiding confusing the users with poor image retrieval results. To decide when to use a contextual image retrieval system to provide image annotations, we explore the task of predicting the inherent difficulty of describing the query as images in this paper. As a starting point to study this problem, we focus on

noun word/phrase queries and exploit the query context from the text to make the prediction. Each query is represented as a vector, where each dimension corresponds to a feature proposed based on our intuition about image query difficulty. Machine learning algorithms are used to train the query difficulty classification models using cross validation. The trained models are compared on a evaluation data set.

2 Representing contextual query as a vector of features

To make it possible to build a contextual image difficulty classifier, we first represent each query as a vector, where each dimension corresponds to one feature. Good features are critical for this application. To find features for all the noun queries, we explore linguistic features and heuristics. In our preliminary research, we start with the following four features, and each is based on one heuristic assumption.

Concreteness To capture the concreteness of a word, we use whether this word is physical or abstract given its context as a feature. The assumption is that a physical query usually corresponds to some physical existence and hence is easier to be illustrated with images than an abstract query. Since the same word can be either concrete or abstract given different context, we use word sense disambiguation (WSD) and WordNet [5] to compute the query concreteness. Semcor [7] is used as our training corpus to train the WSD models. First, we extract the context features for the training queries, including part-of-speech of neighboring words, single words in the surrounding context and local collocations [4]. Next, we use the maximum entropy modeling approach [3] to train the WSD models. Then, we extract the same feature for the noun queries in the document and obtain the sense-level disambiguation results using the trained WSD models. Finally, the sense level hypernymy in WordNet is used to get the query concreteness. If the disambiguated sense of the query in the context is traced back to a physical entity, its concreteness is 1; otherwise its concreteness is 0.

However, not all queries can be disambiguated this way since WSD training data semantically aligned with WordNet are still limited concerning the coverage of the whole vocabulary. For a word without WSD training data but still available through WordNet, the ratio of its physical senses to all the senses in WordNet is used as the concreteness measure. For a word absent from Wordnet, we look up its first dictionary entry in Dictionary.com and use the average concreteness value of all the nouns in the definitions and explanations as its concreteness measure. For example, *endotherm* is not available in WordNet, while its first dictionary result from Dictionary.com is *a warm-blooded animal*. In this definition, *animal* is the only noun and its concreteness is computed to be 1, therefore the concreteness for *endotherm* is considered to be 1. If the word is neither covered by Wordnet nor Dictionary.com, its concreteness is set to be 0.

Commonness To capture the commonness of a word, we use the word usage frequency on the web as a feature. More specifically, we use the Google unigram frequency count publicly available in [1] to approximate the commonness. The assumption is that the most frequently used nouns are usually simple words and might be easier to be illustrated by an image. Although this assumption is not always true, there might be a correlation between query commonness and query difficulty.

Also, we use this dataset to validate the correctness of the query. Since the token counts are generated from approximately 1 trillion word tokens of publicly accessible Web pages, we assume that all the nouns are included. If the query is not present in this dataset, we consider it to be an invalid word and assume it cannot be represented by an image.

Ambiguity To capture the ambiguity of a word, we use the number of noun senses for the word in WordNet as a feature. The assumption is that ambiguous words are more difficult to describe pictorially than unambiguous words. For words not covered by WordNet, the first dictionary result from Dictionary.com is used and the ambiguity is approximated by the number of entries in the result. If the word is absent from Dictionary.com, its ambiguity is set to 1.

Figurativeness To capture the figurativeness of a word, we send the word to Google image search and Google web search and then use the statistics of the retrieval results as a feature. The ratio of the total number of images retrieved from Google image search to the total number of webpages retrieved from Google web search is computed as the figurativeness measure. The idea is that current commercial image search engines mainly rely on text features to do image retrieval and therefore the percentage of the webpages containing the images for the word may indicate the difficulty to obtain an image to represent the word.

3 Experiments

We randomly select several paragraphs from four literature books and one science book: *The Wind in the Willows*, *The Story of My Life*, *A Connecticut Yankee in King Arthur's Court*, *Don Quixote* and *CPO Focus on Life Science*. Two evaluators manually label the 675 noun words in the texts into two classes: class 0 refers to difficult queries which are unable to be represented by images in the context and class 1 refers to the opposite. The evaluation for each query is based on the top 100 images crawled from Google image search. If the evaluator finds any images that can clearly explain the word in the context, it is considered as an easy query; otherwise it is a difficulty query. We use this approach because our contextual image retrieval system uses search engine to collect candidate images for all the noun words/phrases, and then applies our contextual image retrieval algorithm to select the most appropriate images for the query.

Evaluator 1 labels 357 easy queries and 318 difficult queries, while evaluator 2 labels 408 easy queries and 267 difficult queries. They agree on 81.5% (550) of the

queries, among which 230 are difficult queries, such as *absence*, *benefit* and *chatter*, and 320 are easy queries, such as *destruction*, *earth* and *face*. The Cohen's kappa coefficient between the evaluators is 0.6229. We use the mutually agreed 550 queries as the ground truth in our experiments. We first represent each query as a vector of 4 features described before and then compare three machine learning algorithms on the task of predicting difficulty level for these queries, which are C4.5 decision tree (J48), naive Bayes tree (NBTree) and bagging predictors (Bagging). 10-fold cross-validation is used in our experiments, and precision and recall are used to evaluate the performance of these algorithms.

Table 1. Experimental Results.

Classifier	Class	Precision			Recall		
		All	Concrete	C&F	All	Concrete	C&F
J48	0	0.717	0.706	0.52	0.739	0.761	0.278
	1	0.808	0.818	0.611	0.791	0.772	0.816
NBTree	0	0.664	0.655	0.503	0.8	0.809	0.37
	1	0.832	0.835	0.619	0.709	0.694	0.738
Bagging	0	0.679	0.684	0.519	0.743	0.791	0.413
	1	0.802	0.831	0.632	0.747	0.738	0.725
Average	0	0.687	0.682	0.514	0.761	0.787	0.354
	1	0.814	0.828	0.621	0.749	0.735	0.76

As shown in Table 1, the prediction performance doesn't vary much using different machine learning algorithms. We also compare the performance with various feature combinations: using all proposed features (All), using concreteness (Concrete), and using commonness and figurativeness (C&F). The result is also shown in Table 1. It is expected that concreteness is the most important feature to predict query difficulty, however it is unexpected that the other features are almost useless given concreteness. Further investigation shows that ambiguity is irrelevant and the other two features are useful only if concreteness is absent. The irrelevance of the ambiguity may be caused by several reasons. First, our objective is to predict the possibility to represent the queries as images, while the assumption that ambiguity is related to this possibility is wrong. Second, the measure of ambiguity may be poor, since a word with many senses or dictionary entries may not be ambiguous given the context. Commonness and figurativeness are inherently inferior to concreteness, because contextual information is not captured by these two features. Although common words may be easier to be illustrated with images, they are also likely to have more senses. Without word sense disambiguation, features derived from the whole web, as commonness and figurativeness, are different from commonness and figurativeness of the word in the context.

4 Discussions

Predicting query difficulty is an important problem for a contextual image retrieval system, since the queries are not explicitly provided by users, and thus automatically suggesting queries and returning poor retrieved images bother the users a lot. This paper is a first step towards solving this problem and the results are promising using the proposed features and machine learning algorithms. We find that features/classifiers based on the context of the query are better than those without contextual information. Although we study the query difficulty prediction for contextual image search, we expect the proposed context based features are useful for query difficulty prediction of other contextual search engines.

In our preliminary research, all the features are based on the text. How to incorporate visual features to supplement existing textual features still needs to be investigated, especially when certain queries only have one or two qualified images among all the candidate images. An assumption is that the image results of easy queries may be more homogeneous. Similarly, the context information may be critical when this assumption is used. For example, the query *bearing* means *the manner of Don Quixote* instead of *a machine part* in our experimental data, thus it is labeled as a difficult query in the context. However, all the images retrieved for *bearing* are images of *a machine part*, thus the image features developed based on this assumption may not be useful under this circumstance. Another direction for improvement may exploit the query relationships. For example, some abstract words related to *human beings*, such as *admire*, *worship* and *courage*, can be explained by a picture of body language, facial expression or human behavior, while other abstract words related to *measurement*, such as *weight*, *mile* and *hour*, are still difficulty queries. This may indicate that abstract queries, when closely related to some concrete entities, may be able to use some related pictures of these entities to represent.

References

1. T. Brants and A. Franz. Web 1t 5-gram version 1. 2006. `http://www.ldc.upenn.edu/`.
2. R. N. Carney and J. R. Levin. Pictorial illustrations still improve students?learning from text. *Educational Psychology Review*, 14(1):5–26, March 2002.
3. Z. Le. Maximum entropy modeling toolkit for c++. `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`.
4. Y. K. Lee and H. T. Ng. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP '02*, NJ, USA, 2002.
5. G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.
6. K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *SIGDAT '00*, NJ, USA, 2000.
7. G. A. Miller, M. Chodorow, S. Landes, C. Leacock and R. G. Thomas. Using a semantic concordance for sense identification. In *Proc. of Workshop on Human Language Technology*, 1994.