PREDICTION OF ADJECTIVES FOR GIVEN NOUNS USING PROBABILITY DISTRIBUTION OF ADJECTIVE-NOUN PAIRS AND ADJECTIVE-ADJECTIVE SIMILARITY

Chaitanya Ahuja,
IN THE GUIDANCE OF
Dr. Tsuhan Chen

IIT-Cornell ECE Program

ABSTRACT

Tagging system has become very convenient and popular due to the popularity of the social networking sites such as Flicker, Pinterest, Facebook and many such web-sites which involve people tagging images, videos and even text. Tagging involves a lot of effort and one would not prefer to tag any media item very extensively. Hence along with the randomness involved in classless based tagging, we have less amount of information for each media item.

The motivation for this work lies in expanding the scope of tagging, by increasing the vocabulary of the words used. The issue faced with the new words is its compatibility with the existing tags. 'Will this tag fit in with the old tags?', 'Does it change the sense of the portrayed meaning?' are a few questions that are tackled in this work.

***Index Terms*—** Tag prediction, Noun adjective compatibility, Adjective Extraction

1. INTRODUCTION

The system is designed as follows. The input is a set of tags (for simplicity take only nouns and adjectives). These tags are passed through a set of similarity tests, which gives a score to a set of new adjectives extracted from the sentence corpus. The final output is a new set of adjectives, arranged in a sorted order, for every noun in the input set. The predicted adjectives are synonymous to the input adjectives and compatible to the input nouns.

The rest of the report is arranged as follows: Sec.2 describes all the criteria for comparison in detail along with the proper explanation of the scoring system. Sec.3 refers to a method which can be incorporated to add new adjectives, in form of colour, to the input. In Sec.4, the corpus arrangement is explained. Some results are demonstrated in Sec.5. The report concludes in 6.

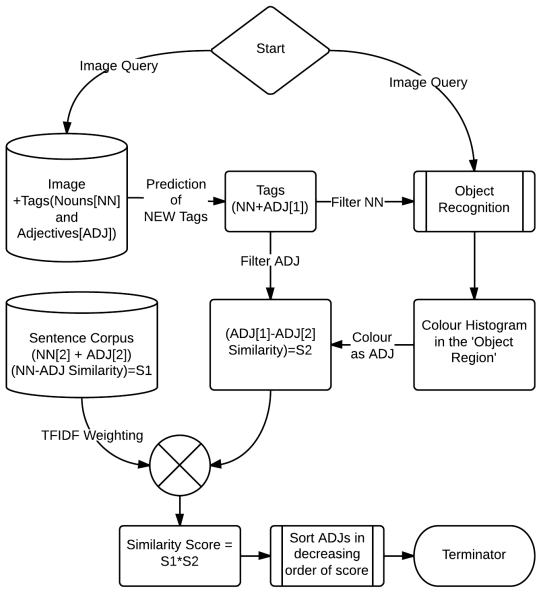

Fig. 1. Flow chart of the proposed algorithm as a combination of multiple subsystems

2. CRITERION FOR RANKING NEW SET OF ADJECTIVES

There are broadly 2 criteria for the calculation of the scores of the suggested adjectives. (1) Similarity between the suggested-adjective **(ADJ-2)** and given-adjective **(ADJ-1)** on basis of meaning. (2) Compatibility between suggested-adjective **(ADJ-2)** and given-noun **(NN)** based on probability of its occurrence in the corpus. Of course, the results highly depend on the data in case of probabilistic analysis and hence the a highly diverse data-set British-National-Corpus (BNC)[1] has been chosen for testing purpose of the proposed algorithm.

The rest of this section explains in detail about the simi-

larity criteria used as the part of the proposed algorithm.

2.1. Similarity between ADJ-1 and ADJ-2

Similarity between 2 words is a very old problem and has been tackled in may different ways in the past. The most popular method is known as 'Lesk Algorithm'[2]. This model indicates a similarity score based on direct overlap of the meanings of two given words. The score increases linearly with the number of overlaps. This algorithm might fail in scenarios where the meanings of two words might be same but is explained using different words. Hence alternative varieties to the same algorithm have been proposed over the years. This includes, but is not limited to, *lch* (Leacock & Chodorow), *path* (Wordnet)[3] and *wup* (Wu & Palmer)[4]. The method used to compare two adjectives is *path* in the proposed algorithm.

Wordnet arranges the words in a hierarchy based on word groups. For eg. Mammals and Reptiles are a subgroup of Animals, Cats and Dogs are a subgroup of Mammals. *path* is score, in the interval [0,1], calculated as the inverse of the path distance between 2 words. Hence this provides us with a way to work around with similarity between any given adjectives. Henceforth similarity between 2 adjectives will be referred to as the Sim(**ADJ-1** ,**ADJ-2**)

2.2. Compatibility between NN and ADJ-2

Probability that **ADJ-2** is a more frequent adjective occurring alongside of **NN** can be calculated using its occurrence in BNC Corpus. For the sentences in the corpus, the Part-of-Speech (POS) is already known. Hence we know all the adjectives and nouns beforehand. Although, in case of multiple adjectives and nouns we cannot determine which adjective corresponds to which noun. We propose a probability function which can determine this.

Let a_i and n_j be the adjectives and nouns in a given sentence. For each a_i we can calculate its probability $P_{a_i \Rightarrow n_j}$ of describing a noun n_j by

$$P_{a_i \Rightarrow n_j} = \frac{\frac{1}{x_{i \Rightarrow j}}}{\sum_{\forall k} \frac{1}{x_{i \Rightarrow k}}} \tag{1}$$

where $x_{i \Rightarrow j}$ is the absolute position difference between a_i and n_j. For example, we have a sentence: The grass is greener on the other side. a_i = greener,other and n_j = grass, side.

$$P_{a_1 \Rightarrow n_1} = \frac{\frac{1}{|1-3|}}{\frac{1}{|1-3|} + \frac{1}{|3-7|}} = 0.67$$

Similarly,

$$P_{a_1 \Rightarrow n_2} = 0.33$$

$$P_{a_2 \Rightarrow n_1} = 0.17$$

$$P_{a_2 \Rightarrow n_2} = 0.83$$

This method of compatibility is first applied on the complete corpus and made available in form of a lookup table.

2.3. TFIDF Weighting

It was observed from the preliminary results that there are a few words which subdued the effect of the similarity with the synonym by virtue of their very high frequency. This needs to be controlled so that the higher ranked adjectives must represent similarity both to **ADJ-1** and **NN** .

TFDIF, though a heuristic argument[5] has been shown to work well for these kind of conditions. We shall define the fundamental equations just for completeness sake.

$$\text{tf}(t, d) = 0.5 + \frac{0.5 \times \text{f}(t, d)}{\max\{\text{f}(w, d) : w \in d\}} \tag{2}$$

where $\text{f}(t, d)$ is the raw frequency of word t in document d,

$$\text{idf}(t, D) = \log \frac{N}{1 + |\{d \in D : t \in d\}|} \tag{3}$$

where N is the total number of documents in the corpus D. $|\{d \in D : t \in d\}|$ is the number of times a word appears in the corpus.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \tag{4}$$

The following algorithm lists all the steps incorporating the similarity scores described above. This algorithm is also incorporated in Fig.1

Algorithm 1 Rank Adjectives based on compatibility and similarity

1: **Initialization**: Take the given set of tags and filter out adjectives(**ADJ-1**) and nouns(**NN**)
2: Compare each word in the set **ADJ-1** with **ADJ-2** (corresponding adjectives to **NN** in the lookup-table) based on similarity test discussed in Sec:2.1
3: Apply TFIDF weighting on each score based on the formulae in Sec:2.3
4: Calculate final score based by multiplying scores from the loookup table and those calculated in Step 2 and 3
5: Repeat this step for every adjective in set **ADJ-1**
6: Order the adjectives in decreasing order of scores

3. EXTRACT MORE ADJECTIVES FROM THE PICTURE

This section provides a theoretical background on how to extract more adjectives out of a given image.

3.1. Colour Histogram

Given tags help recognize the object(s) in the image. Hence object-recognition techniques can be applied using the noun-tags as the classifiers. This will give a region of the object.

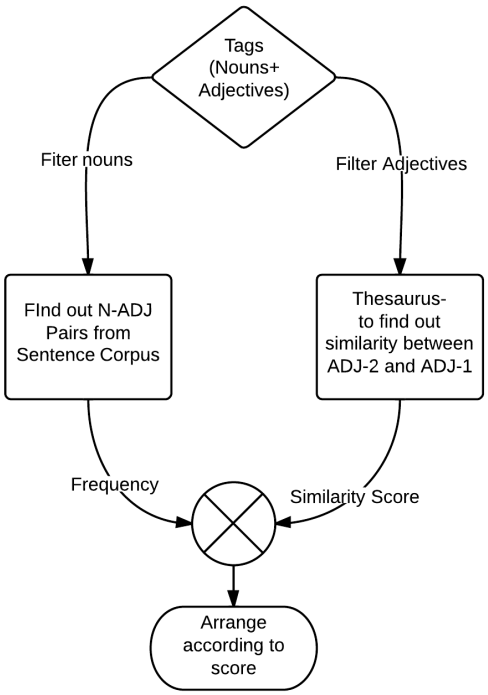

Fig. 2. Base-Line Flowchart which uses frequency instead of probability for **NN** - **ADJ-2** similarity

Taking the colour histogram of this area provides us with the dominant colour in the object. This colour can be directly used as an adjective for the prediction of new adjectives to describe the given nouns/subjects

4. SENTENCE CORPUS

Choice of data for the aforementioned purposes is a very crucial step. Hence, British National Corpus (BNC)[1] was chosen.

It contains written(90%) and oral(10%) English extracts. The written part includes subsets of regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. Whereas, the oral portion of the text is from orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.[1]

5. RESULTS

This section displays some of the results for the given inputs. The output is in the decreasing order of assigned scores.

Basic Algorithm as explained in Fig.2

- **Input:** elegant walk
  - **Output:** NO SUGGESTIONS
- **Input:** comfortable bed
  - **Output:** comfortable, comfy
- **Input:** huge house
  - **Output:** huge, vast

Results for the proposed algorithm as explained in Algorithm-1

- **Input:** elegant walk
  - **Output:** elegant, graceful
- **Input:** comfortable bed
  - **Output:** comfortable, easy, easier, comfy, comfier,easiest, prosperous
- **Input:** huge house
  - **Output:** huge, vast, immense, vaster

The effect of the additional similarity criteria has brought a change in prediction of extra words for a given noun. In the 1^{st} case the base-line does not provide any suggestion whereas the proposed Algorithm predicts 'graceful' as an alternative (or a suggestion).

6. CONCLUSION

In this study, an algorithm for prediction of new adjectives was proposed based on both nouns and adjectives in the given tags. This prediction model has applications in automatic tagging system. Any other model which requires adjective prediction based on input nouns or adverb prediction based on input verb can be incorporated in this model.

7. REFERENCES

[1] BNC Baby and BNC Sampler, "British national corpus," .

[2] Michael Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proceedings of the 5th Annual International Conference on Systems Documentation*, New York, NY, USA, 1986, SIGDOC '86, pp. 24–26, ACM.

[3] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi, "Wordnet:: Similarity: measuring the relatedness of concepts," in *Demonstration Papers at HLT-NAACL 2004*. Association for Computational Linguistics, 2004, pp. 38–41.

[4] Zhibiao Wu and Martha Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze, "Scoring, term weighting, and the vector space model," in *Introduction to Information Retrieval*, pp. 100–123. Cambridge University Press, 2008, Cambridge Books Online.