LANGUAGE FORUM
VOL. 34, NO. 1, JAN-JUN 2008

Image Specific Language Model: Comparing Language Models from Two Independent Distributions from FlickR and the Web

GREGORY GREFENSTETTE
GUILLAUME PITEL
Cea List, Fontenay Aux Roses, France

ABSTRACT

Here we present a study comparing vocabulary from two different language models: the language used by people to describe their pictures on Flickr, and unrestricted language found on the Web. We examine these two language models through the lens of the semantically typed vocabulary in the General Inquirer lexicon. Developed in the 1960s for computerized content analysis of text, it provides a number of semantic categories for each word it contains. We compare the relative frequencies of usage of these categories, and usage of individual word within these categories, in Flickr tags versus the Web, using the weirdness *metric from research in special languages. The distribution of words used in each independent set, image annotations and Web text, is found to be different as expected, and we describe the differences. We show, for example, that noun Flickr tags are much more preferred over adjectives and verbs, more than the Web would lead us to expect. Names for animals, humans, and weddings also appear more frequently in Flickr tag than on the Web. Exploring these two language models helps us to better understand the vocabulary specific to images, which should ultimately be helpful in determining vocabulary for automated image annotation.*

1. INTRODUCTION

Adam Kilgarriff (2005) has shown that language distributions over any two large text collections are always different, even if you just divide one corpus in two random parts. It would seem all the more obvious that distribution of words used by people use when they describe photos is different from that used when they write ordinary text. And though the existence of differences seems intuitively obvious, the