Topics can also be identified in machinegenerated

summaries using standard statistical inference

techniques (Asuncion et al., 2009).

Nenkova et al. (2007) describe in more detail the

incorporation of human variation in the Pyramid

method, the reliability of content annotation, and

the correlation of Pyramid scores with other evaluation

measures

methods

such as ROUGE that measure word n-gram

overlap (Lin and Hovy, 2003) Rogue(lin 2004)

the rankings from the automatic BLEU

method (Papineni et al., 2002) have been shown to correlate

well with human evaluation, and it has been wide

Applying the same data representation

and reliability formula (Krippendorff’s Alpha) as

in (Passonneau, 2004),

factoids with the number of summaries seems to follow a

Zipfian distribution