Comparing Word Relatedness Measures Based on Google n-grams

DALHOUSIE UNIVERSITY Inspiring Minds

Aminul Islam, Evangelos Milios, Vlado Kešelj

islam@cs.dal.ca, eem@cs.dal.ca, vlado@cs.dal.ca

Coling 2012
8-15 December 2012

1. Corpus-based Word Relatedness

- Estimating word relatedness is essential in NLP, and in many other related areas.
- Corpus-based measures have their advantages over knowledge-based supervised measures.
- Corpus-based measures are not *fairly* comparable when different corpora are used.

2. Motivation

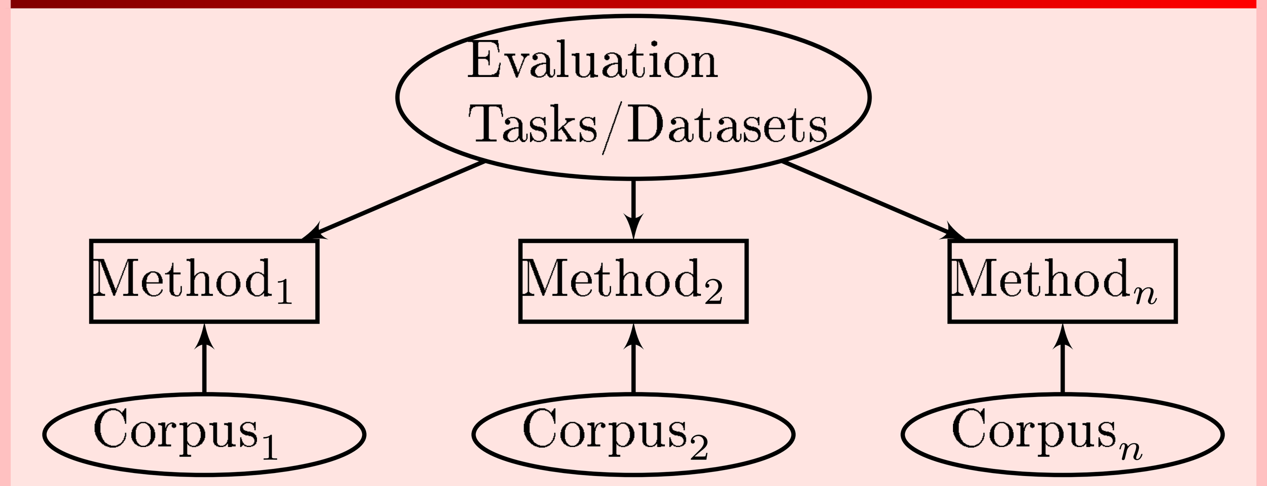

Figure 1: **Current evaluation approach**

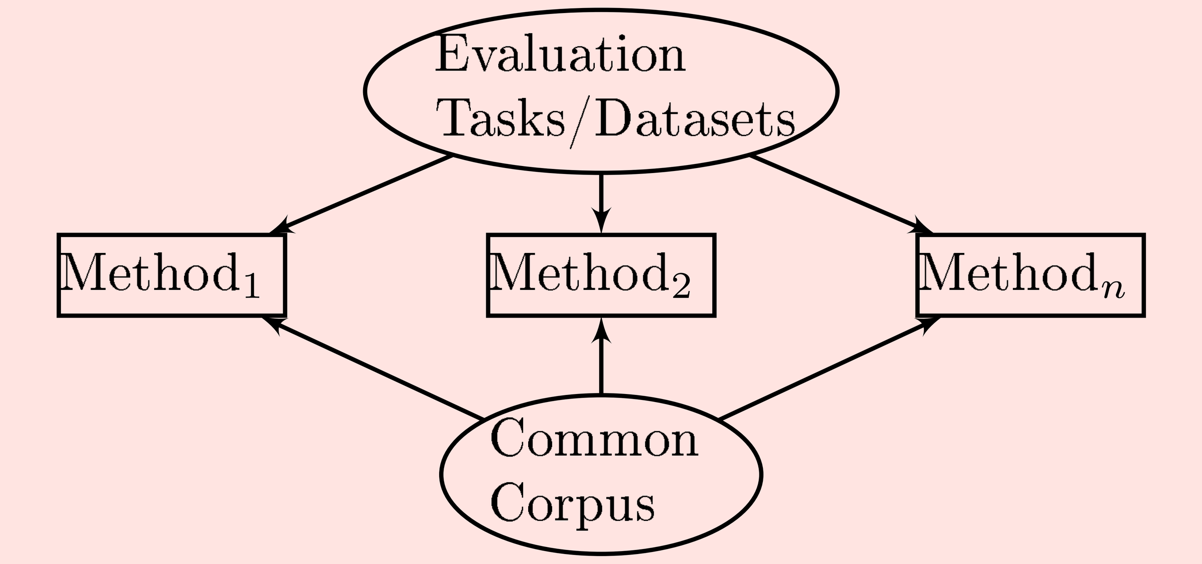

Figure 2: **Proposed evaluation approach**

3. Notation Used

Notation	Description
$C(w_1 \cdots w_n)$	frequency of the n-gram $w_1 \cdots w_n$, where $n \in \{1, \cdots, 5\}$
$D(w_1 \cdots w_n)$	number of web documents having n-gram, $w_1 \cdots w_n$, where $n \in \{1, \cdots, 5\}$
$\lvert V \rvert$	total number of uni-grams in Google n-grams
N	total number of web documents used in Google n-grams
$C_{\max}$	$\max(\{C(w_i)\}_{i=1}^{\lvert V \rvert})$
$M(w_1, w_2)$	number of tri-grams that start with w_1 and end with w_2 (say, M_1)
$M(w_2, w_1)$	M_2
$\mu_T(w_1, w_2)$	$\frac{1}{2}(\sum_{i=3}^{M_1+2} C(w_1 w_i w_2) + \sum_{i=3}^{M_2+2} C(w_2 w_i w_1))$
X	$= \frac{\mu_T(w_1,w_2) C_{\max}^2}{C(w_1)C(w_2)\min(C(w_1),C(w_2))}$
Y	$= \frac{\min(C(w_1),C(w_2))}{C_{\max}}$

4. Common Corpus Used

Google n-grams

Examples of Google tri-grams data:

w_1 w_2 w_3	$C(w_1 w_2 w_3)$
he was a	3,683,417
he was an	563,471
he was am	121
he was awesome	7,520

5. Mapping: web search ⇒ corpus

$D(w_1 \cdots w_n) \leq C(w_1 \cdots w_n)$ as an n-gram may occur multiple times in a single document

Considering the lower limits of $C(w_1)$ and $C(w_1 w_2)$

Two assumptions: (1) $D(w_1) \approx C(w_1)$ and (2) $D(w_1 w_2) \approx C(w_1 w_2)$

6. Corpus Based Measures

Simpson Coefficient$(w_1, w_2) =$

$$\frac{D(w_1 w_2)}{\min(D(w_1), D(w_2))} \approx \frac{C(w_1 w_2)}{\min(C(w_1), C(w_2))}$$

Jaccard Coefficient$(w_1, w_2) =$

$$\frac{D(w_1 w_2)}{D(w_1)+D(w_2)-D(w_1 w_2)} \approx \frac{C(w_1 w_2)}{C(w_1)+C(w_2)-C(w_1 w_2)}$$

Dice Coefficient$(w_1, w_2) =$

$$\frac{2D(w_1 w_2)}{D(w_1)+D(w_2)} \approx \frac{2C(w_1 w_2)}{C(w_1)+C(w_2)}$$

Normalized Google Distance (NGD)$(w_1, w_2) =$

$$\frac{\max(\log D(w_1), \log D(w_2)) - \log D(w_1 w_2)}{\log N - \min(\log D(w_1), \log D(w_2))}$$

$$\approx \frac{\max(\log C(w_1), \log C(w_2)) - \log C(w_1 w_2)}{\log N - \min(\log C(w_1), \log C(w_2))}$$

Pointwise Mutual Information (PMI)(w_1, w_2)

$$= \log_2\left(\frac{\frac{D(w_1 w_2)}{N}}{\frac{D(w_1)}{N}\frac{D(w_2)}{N}}\right) \approx \log_2\left(\frac{\frac{C(w_1 w_2)}{N}}{\frac{C(w_1)}{N}\frac{C(w_2)}{N}}\right)$$

Relatedness Based on Tri-grams (RT)(w_1, w_2)

$$= \begin{cases} \frac{\log X}{-2 \times \log Y} & \text{if } X > 1 \\ \frac{\log 1.01}{-2 \times \log Y} & \text{if } X <= 1 \\ 0 & \text{if } \mu_T(w_1, w_2) = 0 \end{cases}$$

7. Evaluation of Corpus Based Measures on Five Datasets

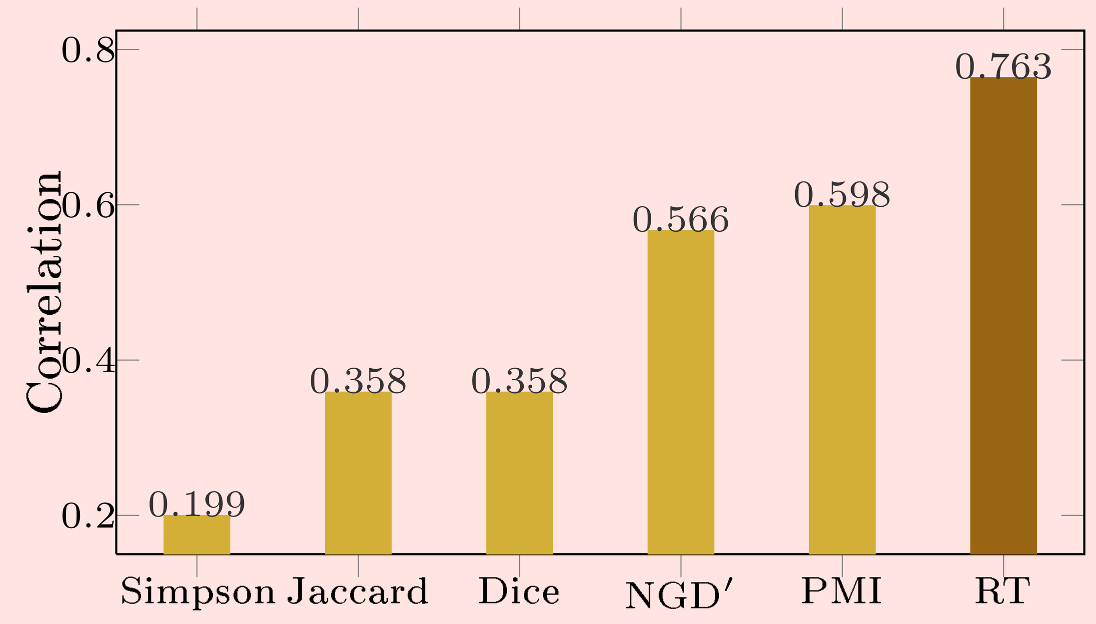

Figure 3: **R&G's 65 noun pairs** [1]

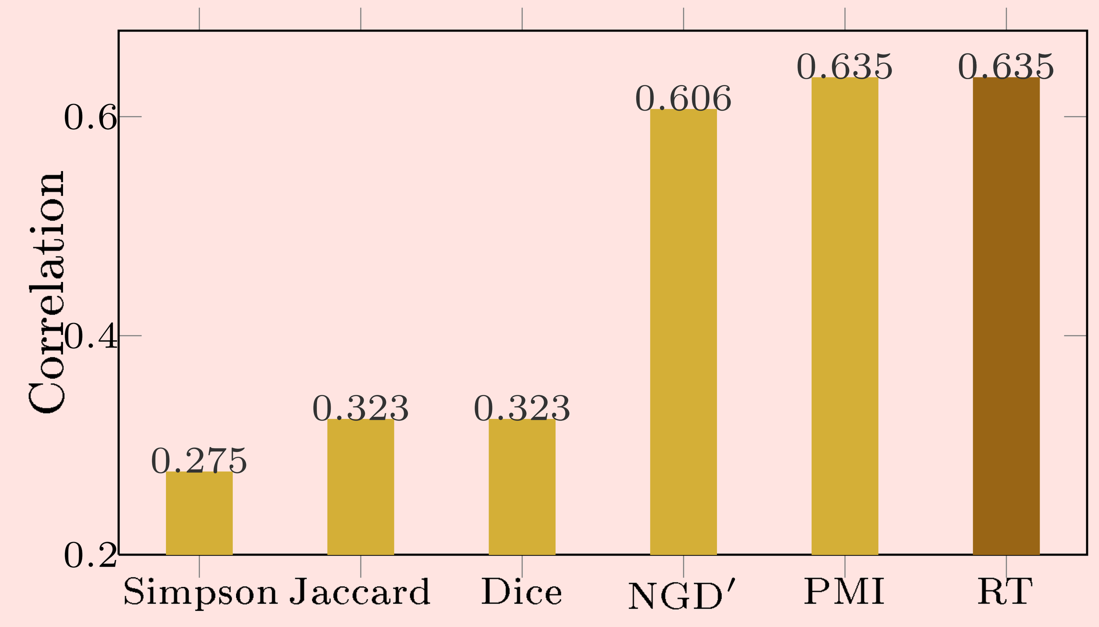

Figure 4: **M&C's 28 noun pairs** [2]

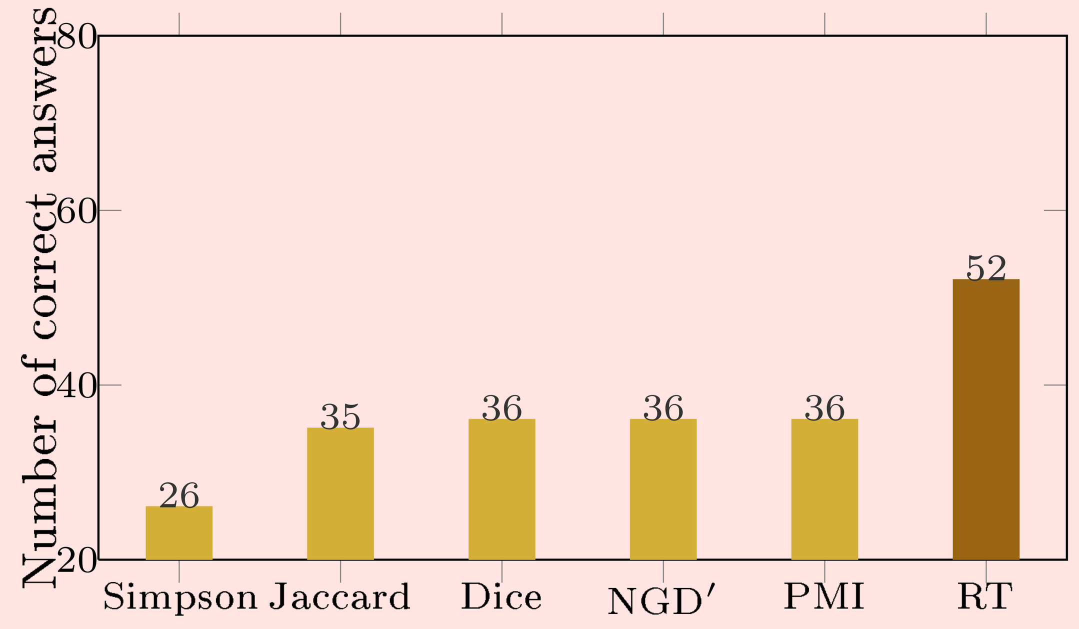

Figure 5: **TOEFL's 80 synonym questions** [3]

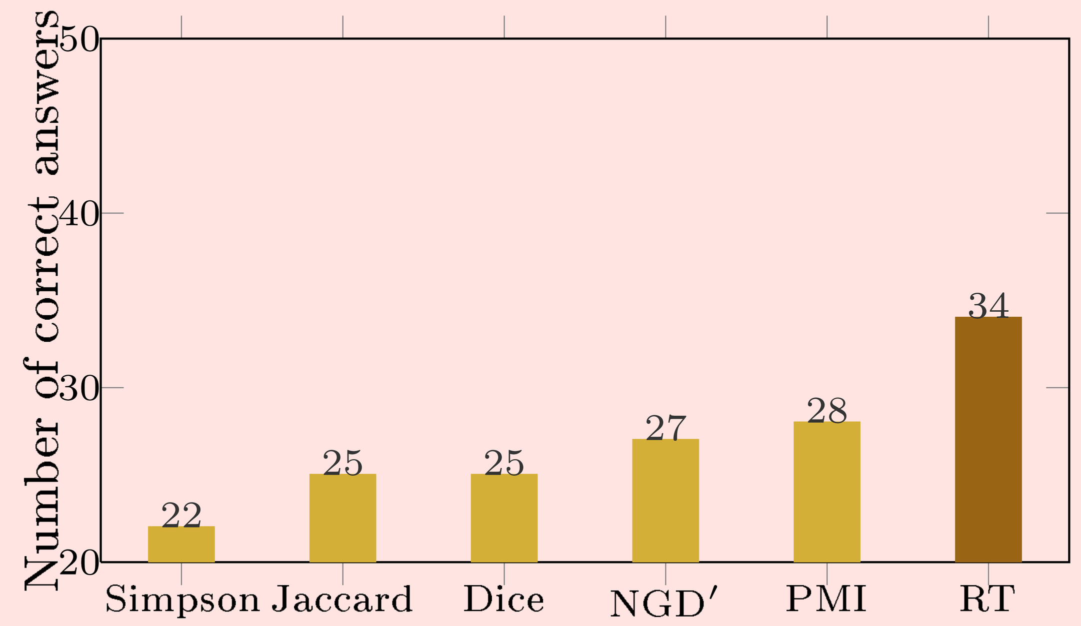

Figure 6: **ESL's 50 synonym questions**

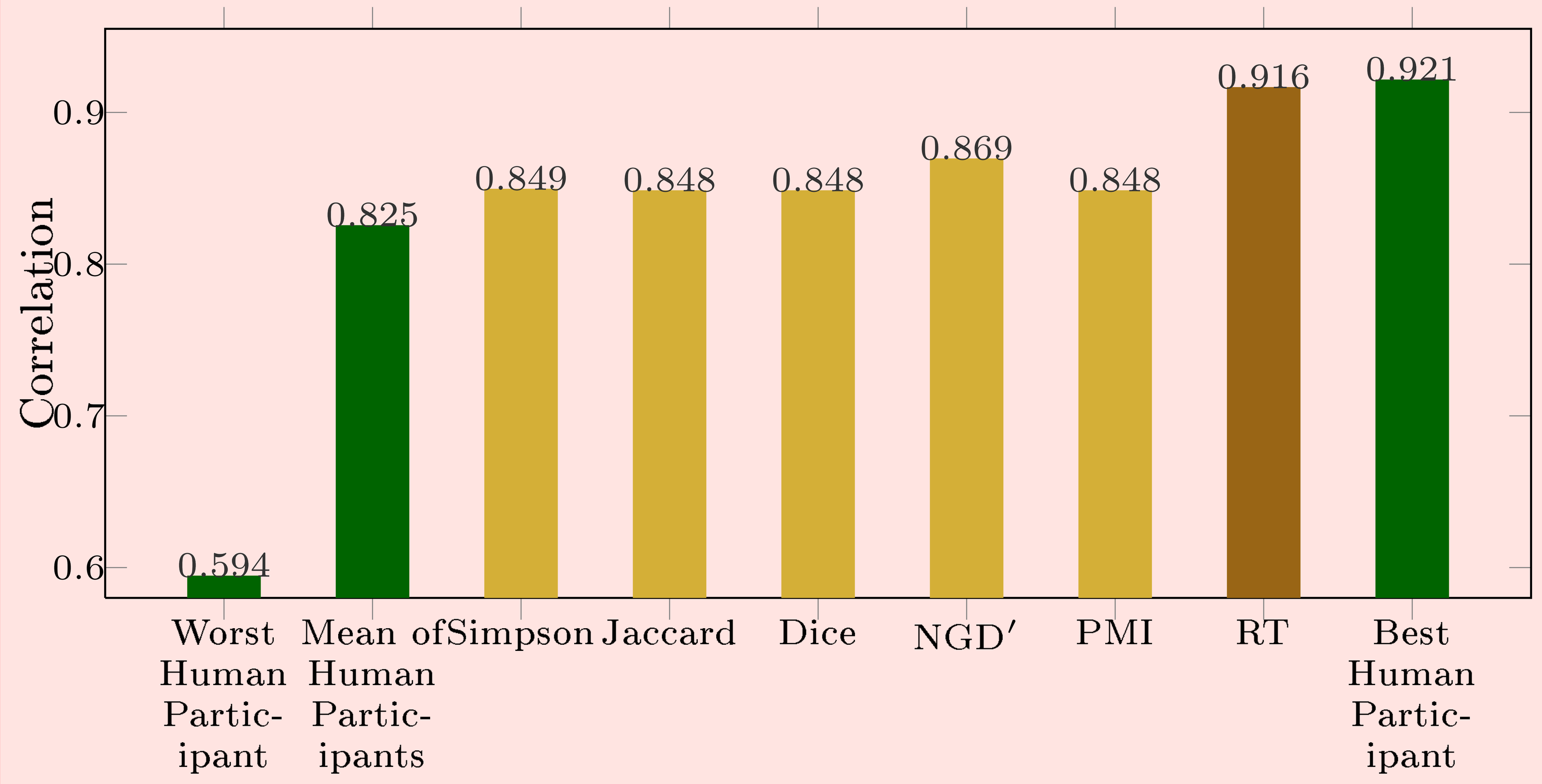

Figure 7: **Li's 30 sentence pairs** [4]

- Li's 30 sentence pairs [4] are used for text similarity task.
- Each sentence pair is rated by 32 human judges.
- The correlation coefficients of text similarity measure [5] (based on the discussed word relatedness measures) with the human judges are shown in Figure 7.

8. Conclusion

- Several corpus-based word relatedness measures have been implemented on the Google corpus and have been *fairly* evaluated on benchmark datasets.
- Mapping between a web search engine and the Google corpus using some assumptions.
- Discussed corpus-based measures are language and domain independent.

References

[1] Rubenstein et al. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

[2] Miller et al. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

[3] Landauer et al. (1997). A solution to plato's problem. *Psychological Review*, 104(2):211–240.

[4] Li et al. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE TKDE*, 18:1138–1150.

[5] Islam et al. (2012). Text Similarity using Google Tri-grams. In: Proc. of CAI'12, 312–317