Word2vec Enhanced BLEU

John Morgan
University of Maryland

MT Evaluation

What is the Problem?

- perfect translations get low BLEU scores

src	ha vueilto el miedo
ref	fear has returned
trans	fear is back

Table : Example of a source sentence in Spanish, one reference, and a perfect translation that receives a low BLEU score.

What is the Cause?

- paucity of references
- sentences can have billions of translations
- only 1 reference

BLEU Solutions

- designed assuming four references
- incorporates features to capture variations of references
- does not require an exact match for entire sentence
- allocates credit to sub-sentence segments
- matches of overlapping n-grams for $1 \leq n \leq 4$

Other Solutions

- METEOR
  - adds stem, synonym, and paraphrase matches
- TERP
  - uses minimum edit distance to score translations
  - compares paraphrases
  - grounded in human judgements
- HyTER
  - employs human annotators
  - write alternative translations
  - phrases compiled into Recursive Transition Networks
  - Encodes exponential number of references
  - disadvantage: requires prohibitively large amount of human labor

Goals

General

- improve BLEU
- attack paucity of references problem
- generate references equivalent to given reference

Desired features in improved BLEU metric

- assign credit to approximate matches
- assign credit to output similar in meaning to input
- adapt easily to new language pair
- require little or no human labor for annotation
- require few NLP resources
- assign reasonable scores to partial output
- measure interactive incremental translation quality
- measure quality of partial inputs

NGram Similarity

What We Did

- build three enhanced BLEU metrics
- one metric yields boost in scores relative to BLEU
- provide scores for $n$grams that do not have exact matches in reference

How did we do it?

- use similarity measure defined in semantic vector space
- vector space generated by word2vec
- word2vec embeds words into vector space
- similarities given by word2vec
- extend similarity function from words to phrases
- average vectors
- compute similarities between phrases
- used pre-trained vectors available on word2vec webpage
- **300** dimensional vectors
- CBOW embedding

3 similarity methods

Scalar BLEU:
operates on word pairs

Vector BLEU:
operates on pairs of n-grams for $1 \leq n \leq 4$

Cross BLEU:
operates on pairs of j and k-grams for $1 \leq j, k \leq 4$

V BLEU	X BLEU
sys → ref	sys → ref
1 → 1	1 → 1 2 3 4
2 → 2	2 → 1 2 3 4
3 → 3	3 → 1 2 3 4
4 → 4	4 → 1 2 3 4

Table : Comparisons of n-grams in V and S BLEU.

Thresholding

Why Threshold?

- many non-exact matches get non-zero similarity
- large portion of matches are spurious

metric	0	70	80	90	95
S BLEU	19.41	19.40	19.40	19.40	19.40
V BLEU	19.57	19.51	19.47	19.41	19.40
X BLEU	41.11	31.96	25.27	20.74	19.54
BLEU	19.40	19.40	19.40	19.40	19.40

Table : Scores by metrics at different thresholds.

BLEU Enhancements

BLEU and New Metrics

- loop over ngrams in system output and reference
- match count accumulated for four precision computations
- add **1** to count if match

BLEU

add **0** for non-match

New Metrics

- compute similarities
- between non-matches and remaining reference $n$grams
- add maximum similarity instead of **0**