Using Collocations to Assess MT Quality

Benjamin Han, Alon Lavie

May 10, 2005

Motivations

- Current MT evaluation metrics (NIST, BLEU, METEOR) are based on *n*-gram similarity between system translations and *reference* translations.
- But *n*-grams do not capture long-distance dependency!
- And most SMT systems already use n-gram language models.
- The requirement of reference translations make them less useful for decoding too.

Example

- Translation from Chinese:
  "*Dow Jones Industrial Average all dropped 97* ***dot worlds*** *on average...*"
- Reference translation:
  "*The average price for 30 main industrial stocks of* ***Dow Jones*** *was down 97 . 15* ***points*** *throughout the day...*"
- Even without the reference translation, we know something is wrong in the machine generated translation!

Goals

- We want to assess X-to-English MT quality
- We propose to use *collocations* obtained from independent English corpora
  - To tell humans and MT systems apart
  - To rank MT systems
  - To rank MT system outputs (decoding)

Collocations Defined

- Every sentence is tokenized and part-of-speech tagged
- Only *content words* are kept: nouns, verbs, adjectives and adverbs (of all forms)
- WordNet is used to canonicalize content words
  - dogs → dog
  - dogged → dog

Collocations Defined

- Collect only the *unique*, canonicalized content words.

  The/DT **Egyptian/NNP Prime/NNP Minister/NNP** ,/, **Atif/NNP Abeed/NNP** ,/, **also/RB met/VBD** the/DT **Sudanese/NNP Minister/NNP today/NN** to/TO **discuss/VB mutual/JJ** and/CC **trade/NN relations/NNS** between/IN **Egypt/NNP** and/CC **Sudan/NNP** ./.

  →

  egyptian/NNP prime/NNP minister/NNP atif/NNP abeed/NNP also/RB meet/VBD sudanese/NNP today/NN discuss/VB mutual/JJ trade/NN relation/NNS egypt/NNP sudan/NNP

- But collocations are not created equal - how to compute their 'strengths'?

Metric I: Dice Metric

$$s = \frac{2 \times c_{12}}{c_1 + c_2}$$

- c_1, c_2: counts of single words
- c_{12}: count of collocations
- When $c_1 = c_2 = c_{12} = 1$: $s = 1$

Metric II: T Score

$$s = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / N}}$$

- $\bar{x}$ is the observed frequency of a collocation
- μ is the null hypothesis (the occurrences of the two words are independent)
- σ^2 is the variance; N is the sample size
- For infinite degree of freedom, we can reject the null hypothesis with 99.5% confidence when $s \geq 2.576$.

Metric III: X^2

$$s = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

- O_{ij} is the corresponding count in the contingency table
- No assumption about the underlying distribution
- A form of null hypothesis testing too - 7.88 is the critical value for 99.5% confidence

	w_1	~ w_1
w_2	O_{11}	O_{12}
~ w_2	O_{21}	O_{22}

Metric IV: Log-likelihood Ratio

$$s = -2 \log \frac{L(H_1)}{L(H_2)}$$

- H_1 (independent hypothesis)
  - $p(w_2|w_1) = p(w_2|\sim w_1)$
- H_2 (dependent hypothesis)
  - $p(w_2|w_1) \neq p(w_2|\sim w_1)$

Training Data

- WSJ sections of Treebank
  - 49,722 sentences, 32,411 unique content words, 2,459,065 collocations
  - Collocation table ≈ 60 MB
- One file from the Gigaword Corpus (nyt199701)
  - 629,164 sentences, 44,713 unique content words, 11,307,512 collocations
  - Collocation table ≈ 780 MB (and we have 4!)

Example Collocations

zeitung zuercher 1.000000
ymca ywca 1.000000
yankee yastrzemski 1.000000
yaniv zvi 1.000000
...
hong kong 0.974504
mixte navigation 0.974359
freddie mac 0.967742
fulton prebon 0.965517
du pont 0.961039
...
jones point 0.118375
...

Dice Metric

T Score

do n't 35.729017
new york 32.843865
cent share 22.145876
earlier year 21.393558
exchange stock 20.737076
...
jones point 7.528200
...
airline transaction 2.576339
aid congress 2.576338
...
banks creditor 2.576163
positive very 2.576157
china reserves 2.576157
earthquake hit 2.576096
...

(from WSJ sections of Treebank)

Example Collocations

zeitung zuercher 49722.000000
ymca ywca 49722.000000
yankees' yastrzemski 49722.000000
yaniv zvi 49722.000000
...
hong kong 47241.054744
mixte navigation 47233.046901
freddie mac 46611.560122
fulton prebon 46406.266423
du pont 45990.073101
...
jones point 724.436690
...

X^2 Score

new york 10133.7
do n't 8124.73484344
street wall 4868.5440851
chief officer 4427.38280735
francisco san 4271.57303174
dow jones 4229.75863621
...
jones point 334.726841926
...
company statement 86.8237568468
democratic republican 86.8038966121
average banks 86.8003659115
book write 86.7912248281
herald newspaper 86.7908043688
...

Log-likelihood Ratio

(from WSJ sections of Treebank)

Outline of Experiments

(4 metrics available)

- Obtain scored collocations from the training data
- For each MT system
  (4 variations available)
  - For each sentence, find out the collocations appearing inside, and compute an average collocation score - call it the sentence score
  - Compute the average sentence score over the entire output - call it the system score

Find the best output (decoding)

System ranking, telling humans and systems apart etc

No reference translation is used

Testing Data

- Tides MT evaluation data 2002 and 2003 on Arabic and Chinese (source languages)
  - Human judgments are available only for Chinese in 2002 and 2003

	# of Humans	# of Systems	# of Sentences
2002 Arabic	4	3	728
2002 Chinese	4	7	878
2003 Arabic	4	6	338
2003 Chinese	4	11	919

Computing Sentence Scores

- Given a sentence we want to find the *interesting* collocations inside to compute the average collocation score
  - Method 1 (*simple*): all collocations found in the training data count!
  - Method 2 (*MST*): only collocations found in the maximum-spanning tree (MST) count.
  - Method 3 (*MST-NCB*): similar to 2 but no crossing branch is allowed in MST
  - Method 4 (*MST-NCB2*): similar to 3 but we add one initial branch when building the MST.

Method 2: Collocations in MST

- Why? If compositionally holds, words in different constituents should matter less

  "*Mussa 's one-day trip coincides with Sudanese Foreign Minister Mustafa Uthman Ismail 's visit who arrived in Tripoli today .*"

  - The strength of collocation "*minister today*" should not affect our judgment of the translation quality
  - But the strength of "*arrive today*" should!
- How to enforce this bias?

Method 2: Collocations in MST (cont'd)

- Answer: use *dependency structures*!

- But we don't have full parses.
- Fake it!
  - Every content word is a vertex in a graph; collocations are weighted edges (branches)
  - Find the graph's MST using Kruskal's algorithm

MST: Example

*"**Mussa** 's **one-day trip coincides** with **Sudanese Foreign Minister Mustafa Uthman Ismail** 's **visit** who **arrived** in **Tripoli today** ."*

Pretty bad hum?

MST (11 collocations, sentence t-score: 3.019)

Correct (14 collocations):

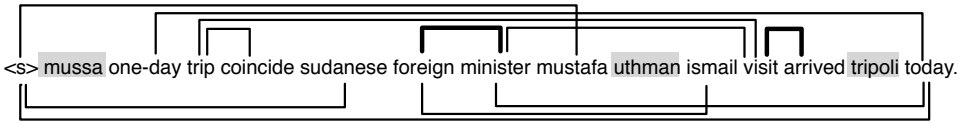
Method III: MST with No-Crossing Branches

- Modify Kruskal's algorithm so no crossing branch is allowed - result could be disconnected.

Hum a bit better now...

MST-NCB (8 collocations, sentence t-score: 3.764)

MST (11 collocations, sentence t-score: 3.019)



Correct (14 collocations):

Method IV: MST-NCB with One Initial Branch

- An obvious hack: always connect <s> to the first verb from left to right!

Even better now, but what next?

MST-NCB2 (8 collocations, sentence t-score: 3.356)

MST-NCB (8 collocations, sentence t-score: 3.764)

Correct (14 collocations):

And the Results are...

- We have 2 years of data, across 2 source languages.
- We have 4 different collocation metrics.
- We have 4 variations on how to compute sentence scores.
- In total we have 64 combinations of results!
- And we have 3 goals to fulfill (I'll only cover the first two).
- I'll try my best.

Goal 1: Telling Humans and Systems Apart

2002 Arabic: Different Collocation Metrics

MST (best performing one among the 4 variations)

	Dice Metric	T	X^2	LR
Human Average	0.0620049	7.5624950	11009.499	559.36625
System Average	0.0482311	5.6388967	8420.9759	337.50552
Separation Ratio	22.21%	25.44%	23.51%	39.66%

Separation ratio = (human avg - system avg) / human avg

2002 Chinese: Different Collocation Metrics

MST (best performing one among the 4 variations)

	Dice Metric	T	X^2	LR
Human Average	0.0633512	7.9221359	11604.130	625.73902
System Average	0.0595995	7.5500239	11242.706	600.59049
Separation Ratio	5.92%	4.70%	3.11%	4.02%

2003 Arabic: Different Collocation Metrics

MST-NCB2 (best performing one among the 4 variations)

	Dice Metric	T	X^2	LR
Human Average	0.0636795	6.8804258	12029.167	574.46616
System Average	0.0532616	6.0165137	9782.2846	456.69555
Separation Ratio	16.36%	12.56%	18.68%	20.50%

2003 Chinese: Different Collocation Metrics

MST (best performing one among the 4 variations)

	Dice Metric	T	X^2	LR
Human Average	0.0728313	9.3319466	13288.081	802.48052
System Average	0.0655111	8.8165155	11835.198	762.27504
Separation Ratio	10.05%	5.52%	10.93%	5.01%

2002 Arabic: Different Sentence Scoring Methods

LR (best performing one among the 4 metrics)

	Simple	MST	MST-NCB	MST-NCB2
Human Average	119.76790	559.36625	549.90046	549.90046
System Average	84.589610	337.50552	334.29914	334.25972
Separation Ratio	29.37%	39.66%	39.21%	39.21%

2002 Chinese: Different Sentence Scoring Methods

Dice metric (best performing one among the 4 metrics)

	Simple	MST	MST-NCB	MST-NCB2
Human Average	0.0127268	0.0633512	0.0609119	0.0597488
System Average	0.0119374	0.0595995	0.0576491	0.0566402
Separation Ratio	6.20%	5.92%	5.36%	5.20%

2003 Arabic: Different Sentence Scoring Methods

LR (best performing one among the 4 metrics)

	Simple	MST	MST-NCB	MST-NCB2
Human Average	120.27843	579.96496	574.47475	574.46616
System Average	97.302655	464.98311	456.69555	456.69555
Separation Ratio	19.10%	19.83%	20.50%	20.50%

2003 Chinese: Different Sentence Scoring Methods

X^2 (best performing one among the 4 metrics)

	Simple	MST	MST-NCB	MST-NCB2
Human Average	1027.9084	13288.081	13184.918	13108.019
System Average	1050.6777	11835.198	11756.706	11645.443
Separation Ratio	-2.22%	10.93%	10.83%	11.16%

Goal 1: Conclusions

- We can tell the difference between humans and systems using collocations.
- Dice metric and X^2 have similar behavior, while T and LR behave similarly.
- Arabic favors T/LR, but Chinese favors Dice metric/X^2.

Goal I: Conclusions

- In 2002, Simple and MST perform better, but in 2003, MST-NCB and MST-NCB2 perform better (system improved?).
- For some reason Chinese seems to be different from/harder than Arabic?
  - Conjecture: segmentation errors can alter word choices and grammatical structures of translations

Goal II: Ranking the Systems

Comparing Ranked Lists

- We want to give a "similarity" score to two fully-ordered ranked lists.
- Decouple each list into relational pairs, and calculate the accuracy according to pair overlaps.

Example:

Gold list: a > b > c ⇀ Relational pairs: (a,b), (a,c), **(b,c)**
Answer list: b > c > a ⇀ Relational pairs: **(b,c)**, (b,a), (c,a)

Accuracy of the answer list = 1 / 3

2002 Chinese: Collocation-based vs. Human Judgments

	Simple	MST	MST-NCB	MST-NCB2
Dice Metric	0.571429	0.714286	0.714286	0.714286
T	0.666667	0.857143	0.809524	0.809524
X2	0.52381	0.666667	0.666667	0.666667
LR	0.52381	0.714286	0.761905	0.761905

- BLEU vs. Human: 0.8095238
- NIST vs. Human: 0.9047619

2003 Chinese: Collocation-based vs. Human Judgments (Adequacy)

	Simple	MST	MST-NCB	MST-NCB2
Dice Metric	0.466667	0.466667	0.466667	0.466667
T	0.4	0.6	0.6	0.6
X2	0.6	0.466667	0.466667	0.466667
LR	0.4	0.533333	0.533333	0.533333

- METEOR vs. Adequacy: 0.8667
- BLEU vs. Adequacy: 0.7333
- NIST vs. Adequacy: 0.8667

2003 Chinese: Collocation-based vs. Human Judgments (Fluency)

	Simple	MST	MST-NCB	MST-NCB2
Dice Metric	0.6	0.6	0.6	0.6
T	0.533333	0.733333	0.733333	0.733333
X2	0.733333	0.6	0.6	0.6
LR	0.533333	0.666667	0.666667	0.666667

- METEOR vs. Fluency: 0.8667
- BLEU vs. Fluency: 0.8667
- NIST vs. Fluency: 0.8667

Goal II: Conclusions

- Our methods perform adequately comparing to other evaluation metrics, but ours do not require reference translations.
- Looks like T score + MST is the winning ticket for Chinese-to-English translations.

Future Work

- Exploring the usage in decoding.
- Revisit the assumption: strong collocations = dependency.
- More sophisticated collocation models, incorporating simple POS patterns and distance information.
- A systematic way of measuring the similarity of induced dependency structures vs. the real structures, and the correlation with MT rankings.