Combining Coherence Models and Machine Translation Evaluation Metrics for Summarization Evaluation

Ziheng Lin [1,2] Chang Liu [1] Hwee Tou Ng [1] Min-Yen Kan [1]

[1] National University of Singapore

[2] SAP Research

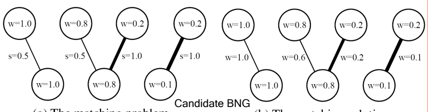
Introduction

- A good machine-generated summary should have high content coverage and linguistic quality
- State-of-the-art summarization systems:
  Extraction-based, focusing on content
- Current AESOP task focuses on:
  Content, readability, and overall responsiveness
- Lin et al. (2011) used a discourse model to discern original text from its permutation
  → Adapt the model to evaluate readability
- Parallel between evaluations of MT and summarization
  → Adapt a state-of-the-art MT evaluation metric to evaluate summary content
- Combine 2 models to evaluate responsiveness with a trained regression model

TESLA-S: Evaluating Summary Content

TESLA: MT Evaluation Metric (Liu et al. 2010, Dahlmeier et al. 2011)

- Extends BLEU with linear programming-based matching
- Uses linguistic resources
- Considers both precision and recall
- Align 2 BNGs to maximize overall similarity



(a) The matching problem (b) The matching solution

Adapting TESLA for summarization

- Mimic ROUGE-SU4: construct 1 matching problem between unigrams and 1 between skip bigrams with a window size of 4, average to give a final score
- Do not match synonyms and POS, since most systems are extraction-based
- Significance test: Koehn's bootstrap resampling
- Tested on AESOP 2011
- Evaluated against:
  Pearson's r, Spearman's p, Kendall's t

Experiments

- Initial summarization task: outperforms all metrics on all correlations
  Significantly better than R-2 on Pearson
- Update summarization task: ranks 2^{nd}, 1^{st}, and 2^{nd}
  Significantly better than R-SU4 on Pearson

	Initial			Update		
	P	S	K	P	S	K
R-2	0.9606	0.8943	0.7450	0.9029	0.8024	0.6323
R-SU4	0.9806	0.8935	0.7371	0.8847	0.8382	0.6654
BE	0.9388	0.9030	0.7456	0.9057	0.8385	0.6843
4	0.9672	0.9017	0.7351	0.8249	0.8035	0.6070
6	0.9678	0.8816	0.7229	0.9107	0.8370	0.6606
8	0.9555	0.8686	0.7024	0.8981	0.8251	0.6606
10	0.9501	0.8973	0.7550	0.7680	0.7149	0.5504
11	0.9617	0.8937	0.7450	0.9037	0.8018	0.6291
12	0.9739	0.8972	0.7466	0.8559	0.8249	0.6402
13	0.9648	0.9033	0.7582	0.8842	0.7961	0.6276
24	0.9509	0.8997	0.7535	0.8115	0.8199	0.6386
TESLA-S	**0.9807**	**0.9173**	**0.7734**	0.9072	**0.8457**	0.6811

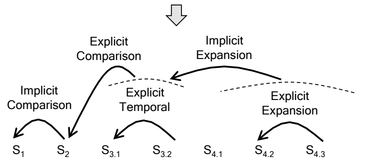
DICOMER: Evaluating Summary Readability

- A readable text should be coherent
- An incoherent text will result in low readability

→A coherence model can also measure readability

Lin et al. (2011)'s Coherence Model

S_1	Japan normally depends heavily on the Highland Valley and Cananea mines as well as the Bougainville mine in Papua New Guinea.
S_2	Recently, Japan has been buying copper elsewhere.
$S_{3.1}$	But as Highland Valley and Cananea begin operating,
$S_{3.2}$	they are expected to resume their roles as Japan's suppliers.
$S_{4.1}$	According to Fred Demler, metals economist for Drexel
$S_{4.3}$	Burnham Lambert, New York,
$S_{4.3}$	"Highland Valley has already started operating and Cananea is expected to do so soon."



	Terms				
	copper	cananea	operat	depend	...
S_1	nil	Comp.Arg1	nil	Comp.Arg1	
S_2	Comp.Arg2 Comp.Arg1	nil	nil	nil	
S_3	nil	Comp.Arg2 Temp.Arg1 Exp.Arg1	Comp.Arg2 Temp.Arg1 Exp.Arg1	nil	
S_4	nil	Exp.Arg2	Exp.Arg1 Exp.Arg2	nil	

Discourse role transition prob of length 2 and 3:

e.g., Comp.Arg2→Exp.Arg2 = 2/25 = 0.08

Predicting Readability Scores

- Human judges score each model/candidate summary with a readability score from 1 to 5
  → List of training instances
- SVM^{light} preference ranking
- Trained on AESOP 2009 - 2010, tested on 2011

Experiments

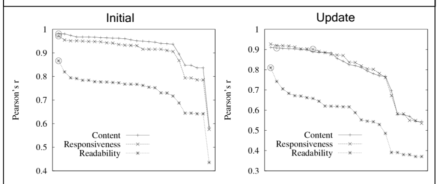
- LIN: outperforms all metrics on both tasks
  Better results on ranking-based Spearman and Kendall due to the ranking model
- Either new feature source improves all scores
- DICOMER: adding both gave the best performance for all scores

Koehn's significance test

		Initial			Update		
	vs.	P	S	K	P	S	K
LIN	4	*	**	**	**	**	**
LIN+C	4	**	**	**	**	**	**
LIN+E	4	**	**	**	*	**	**
DICOMER	4	**	**	**	**	**	**
DICOMER	LIN	–	*	*	*	–	–

	Initial			Update		
	P	S	K	P	S	K
R-2	0.7524	0.3975	0.2925	0.6580	0.3732	0.2635
R-SU4	0.7840	0.3953	0.2925	0.6716	0.3627	0.2540
BE	0.7171	0.4091	0.2911	0.5455	0.2445	0.1622
4	0.8194	0.4937	0.3658	0.7423	0.4819	0.3612
6	0.7840	0.4070	0.3036	0.6830	0.4263	0.3141
12	0.7944	0.4973	0.3589	0.6443	0.3991	0.3062
18	0.7914	0.4746	0.3510	0.6698	0.3941	0.2856
23	0.7677	0.4341	0.3162	0.7054	0.4223	0.3014
LIN	**0.8556**	**0.6593**	**0.4953**	**0.7850**	**0.6671**	**0.5008**
LIN+C	**0.8612**	**0.6703**	**0.4984**	**0.7879**	**0.6828**	**0.5135**
LIN+E	**0.8619**	**0.6855**	**0.5079**	**0.7928**	**0.6990**	**0.5309**
DICOMER	**0.8666**	**0.7122**	**0.5348**	**0.8100**	**0.7145**	**0.5435**

Discussion



- Initial task: correlations for content are consistently slightly higher than responsiveness
- Update task: correlations for content and responsiveness are overlapping
- Correlations for readability are much lower than those for content and readability: a gap of ~0.2
  → much room for improvement for readability
- Correlations are always better on initial task
  → eval metric needs to consider update factor

Two New Feature Sources

- Whether a relation is Explicit or Non-Explicit
  Explicit and Non-Explicit have different distribution on each relation, e.g.:
  Comp.Arg2 to E.Comp.Arg2
  Exp.Arg1 to N.Exp.Arg1
- Whether one relation is embedded in another
  Important to know how well-structured a summary is
  Represented by multiple discourse roles in each cell
  Introduce intra-cell bigrams to capture these:
  e.g., in $C_{cananea,S3}$, Comp.Arg2←→Exp.Arg1

CREMER: Evaluating Overall Responsiveness

We applied SVM^{light} to train a regression model with TESLA-S and DICOMER scores as features

- 3 kernels: linear, polynomial, radial basis
- Trained on AESOP 2009 - 2010, tested on 2011

Experiments

- Initial task: RBF outperforms all AESOP metrics:
  1.71%, 3.86%, 4.60% on Pearson, Spearman, and Kendall
- Update task: all 3 models do not perform as well
- Koehn's sig test: $CREMER_{RBF}$ significantly outperforms ROUGE-2 and -SU4 on initial task

	Initial			Update		
	P	S	K	P	S	K
R-2	0.9416	0.7897	0.6096	0.9169	0.8401	0.6778
R-SU4	0.9545	0.7902	0.6017	0.9123	0.8758	0.7065
BE	0.9155	0.7683	0.5673	0.8755	0.7964	0.6254
4	0.9498	0.8372	0.6662	0.8706	0.8674	0.7033
6	0.9512	0.7955	0.6112	0.9271	0.8769	0.7160
11	0.9427	0.7873	0.6064	0.9194	0.8432	0.6794
12	0.9469	0.8450	0.6746	0.8728	0.8611	0.6858
18	0.9480	0.8447	0.6715	0.8912	0.8377	0.6683
23	0.9317	0.7952	0.6080	0.9192	0.8664	0.6953
25	0.9512	0.7899	0.6033	0.9033	0.8139	0.6349
$CREMER_{LF}$	0.9381	0.8346	0.6635	0.8280	0.6860	0.5173
$CREMER_{PF}$	**0.9621**	**0.8567**	**0.6921**	0.8852	0.7863	0.6159
$CREMER_{RBF}$	**0.9716**	**0.8836**	**0.7206**	0.9018	0.8285	0.6588