Summarization Evaluation Under an N-Gram Graph Perspective. In View of Combined Evaluation Measures.

George Giannakopoulos [1,2]
Vangelis Karkaletsis [1] George Vouros [2]

[1]Institute of Informatics and Telecommunications – Software and Knowledge Engineering Lab – N.C.S.R. Demokritos
{ggianna|vangelis}@iit.demokritos.gr
[2]Department of Information and Communication Systems – University of the Aegean
georgev@aegean.gr

November 18, 2008

- Present AUTOmatic SUMMarization Evaluation using N-gram Graphs (AutoSummENG)
- Combinatory evaluation – Insight and Discussion
- Proposing Generic Algorithms and Methods for Evaluation and Summarization

Presentation Structure

Already Proposed Methods

- Rouge [Lin and Hovy, 2003, Lin, 2004]
- Basic Elements [Hovy et al., 2005]
- Pyramid [Passonneau et al., 2006]
- Other alternatives... [Steinberger and Jezek, 2004, Radev et al., 2000, Daume III and Marcu, 2005]

Overview[1]

- Statistical *i.e.* **Language-Neutral**
- Word N-gram or Character N-Gram (Q-Gram) Based
- Graph Based on Neighbourhood *i.e.* Includes Uncertainty / Fuzziness

[1] also see [Giannakopoulos et al., 2008]

Overview[1]

- Statistical *i.e.* **Language-Neutral**
- Word N-gram or Character N-Gram (Q-Gram) Based
- Graph Based on Neighbourhood *i.e.* Includes Uncertainty / Fuzziness
- **No Preprocessing**

[1] also see [Giannakopoulos et al., 2008]

Extraction Process

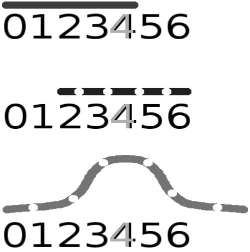
- Extract n-grams of ranks $[L_{\mathrm{min}}, L_{\mathrm{MAX}}]$
- Determine neighbourhood (window size D_{win})
- Assign weights to edges

Example

String:	*abcde*
Character N-grams (Rank 3):	*abc, bcd, cde*
Edges (Window Size 1):	*abc-bcd, bcd-cde*
Weights (Occurences):	*abc-bcd (1.0) , bcd-cde (1.0)*

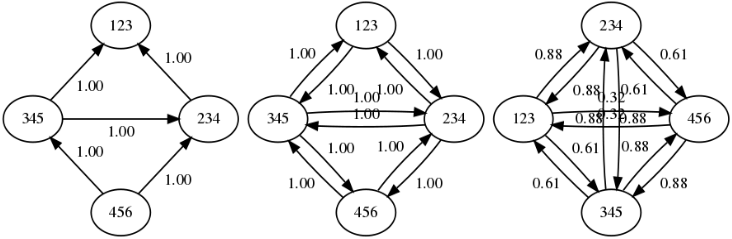
Window-based Extraction of Neighbourhood – Examples

Figure: N-gram Window Types (top to bottom): non-symmetric, symmetric and gauss-normalized symmetric. Each number represents either a word or a character n-gram



N-gram Graph – Representation Examples

Figure: Graphs Rerpesenting the String *123456* (from left to right): non-symmetric, symmetric and gauss-normalized symmetric. N-Grams of Rank 3.



N-gram Graph – Comparison Operator Process

- Size Similarity: Number of Edges
- Co-occurence Similarity: **Existence** of Edges
- Value Similarity: **Existence** and **Weight** of Edges

Notes

- Similarity measures are symmetric. Are they metrics? (Triangle Inequality)
- Derived Measures: Size-Normalized Value Similarity
- Overall similarity: Weighted Normalized Sum over All N-Gram Ranks

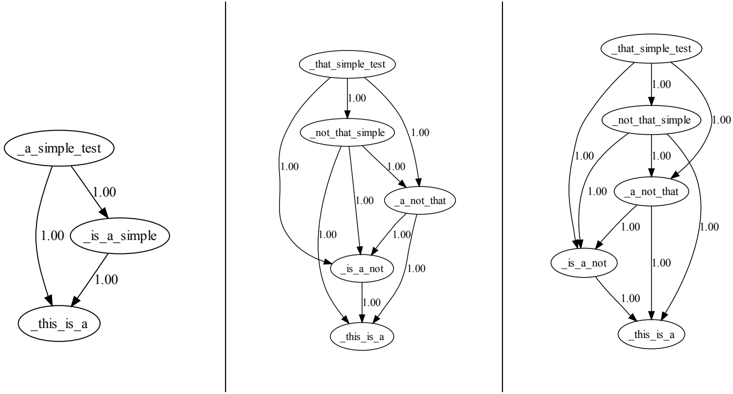
N-gram Graph – Comparison Example

Example

1. This is a simple test.
2. This is a, not that simple, test.
3. This is a not that simple test.

Graph Example – Word Graph

Example



Graph Example – Similarity Scores

Example

Operands	*Value*	*Co-occurence*	*Size*
Word 1-2	0.00%	0.00%	33.33%
Word 1-3	0.00%	0.00%	33.33%
Word 2-3	100.00%	100.00%	100.00%
Character 1-2	32.94%	53.85%	61.18%
Character 1-3	54.43%	82.69%	65.82%
Character 2-3	64.71%	69.62%	92.94%

TAC AutoSummENG System Score

Averaged score over all summaries of the average Value Similarity of the summary to the model summaries. Symmetric window, $(L_{\text{min}}, L_{\text{MAX}}, D_{\text{win}}) = (3, 3, 3)$.

AutoSummENG – Evaluation TAC 2008

AE to...	*Spearman*	*Kendall*	*Pearson*
Resp.	0.8953 (< 0.01)	0.7208 (< 0.01)	0.8945 (< 0.01)
Ling.	0.5390 (< 0.01)	0.3819 (< 0.01)	0.5307 (< 0.01)

Table: Correlation of the *system* AutoSummENG score to human judgement for peers only (p-value in parentheses)

AE to ...	*Spearman*	*Kendall*	*Pearson*
Resp.	0.3788 (< 0.01)	0.2896 (< 0.01)	0.3762 (< 0.01)
Ling.	0.1982 (< 0.01)	0.1492 (< 0.01)	0.1933 (< 0.01)

Table: Correlation of the *summary* AutoSummENG score to human judgement for peers only (p-value in parentheses)

AutoSummENG – Evaluation Over All DUC & TAC

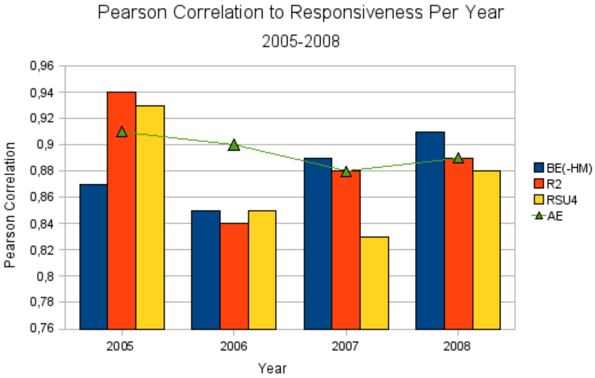


Figure: Pearson Correlation of Measures to the (Content) Responsiveness Metric of DUC 2005-2008 for Automatic Systems

AutoSummENG – Parameters

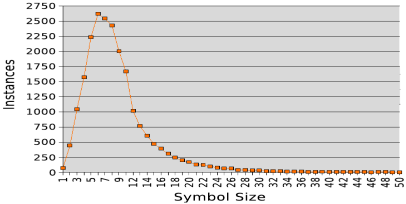
- Word or Character N-gram
- Neighbourhood Window Type
- Minimum N-gram length L_{min}.
- Maximum N-gram length L_{MAX}.
- Neighbourhood Window Size D_{win}.

Symbols – Non-Symbols

Symbols Sequences of characters (letters) that are not neighbours by mere chance.

Non-symbols Sequences of characters (letters) that simply happen to occur near each other.

Figure: The Distribution of Symbols per Rank (Symbol Size) in the DUC 2006 corpus



Parameter Estimation – Experiments

$L_{\text{min},0}, L_{\text{MAX},0}, D_{\text{win},0}$: Signal-to-Noise is maximized.

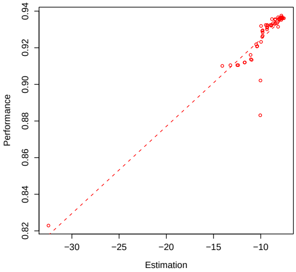


Figure: Correlation between Estimation (S/N) and Performance (Pearson: 0.912)

Experiments on Combined Evaluation – Setting

Regression Using All Eval Methods

- Features: Rouge-2, Rouge-SU4, BE, AutoSummENG (Char 3,5,7; Word 1,2,3)
- Target: Responsiveness / Linguistic Quality
- Platform WEKA [Witten and Frank, 2005] – 10-fold Cross-Validation

Experiments on Combined Evaluation – Results

Table: Pearson Correlation. Max Performances Indicated as **Bold**.

Method	*Resp.*			*Ling.*		
	All	*AE*	*Others*	*All*	*AE*	*Others*
Linear R.	*0.915*	0.915	0.903	*0.630*	**0.630**	**0.541**
SMO R.	*0.920*	0.914	0.880	*0.540*	0.567	0.471
Mult. Perc.	**0.928**	0.899	**0.905**	**0.704**	0.547	0.488
ϵ-SVR (LibSVM)	*0.924*	**0.923**	0.903	*0.409*	0.445	0.447

Measuring Feature Utility

PCA

- Gave a single complex feature
- Almost identical weights for features due to correlation

Need for orthogonal features (ideally).
See [Conroy and Dang, 2008]

N-Gram Graphs – Operators

Graph Operators

- Merging or Union $\cup$
- Intersection $\cap$
- Delta Operator (*All-Not-In* operator) $\triangle$
- Inverse Intersection Operator $\bigtriangledown$

N-Gram Graphs – Summarization Applications

- Content Selection (Chunking, Intersection, Comparison)
- Query Expansion (Semantic Annotation, Comparison)
- Redundancy Checking (Comparison)
- Summary Evaluation (Comparison)

N-Gram Graphs – Summarization Applications

- Content Selection (Chunking, Intersection, Comparison)
- Query Expansion (Semantic Annotation, Comparison)
- Redundancy Checking (Comparison)
- Summary Evaluation (Comparison)
- *Sequence Statistical Normality Estimation (Grammaticality)*

N-Gram Graphs – Summarization Applications

- Content Selection (Chunking, Intersection, Comparison)
- Query Expansion (Semantic Annotation, Comparison)
- Redundancy Checking (Comparison)
- Summary Evaluation (Comparison)
- *Sequence Statistical Normality Estimation (Grammaticality)*
- *Topic Clustering (Comparison)*

N-Gram Graphs – Summarization Applications

- Content Selection (Chunking, Intersection, Comparison)
- Query Expansion (Semantic Annotation, Comparison)
- Redundancy Checking (Comparison)
- Summary Evaluation (Comparison)
- *Sequence Statistical Normality Estimation (Grammaticality)*
- *Topic Clustering (Comparison)*
- *Multiple Granularity Evaluation (Comparison, Graph Cliques)*

N-Gram Graphs – Summarization Applications

- Content Selection (Chunking, Intersection, Comparison)
- Query Expansion (Semantic Annotation, Comparison)
- Redundancy Checking (Comparison)
- Summary Evaluation (Comparison)
- *Sequence Statistical Normality Estimation (Grammaticality)*
- *Topic Clustering (Comparison)*
- *Multiple Granularity Evaluation (Comparison, Graph Cliques)*
- *Probabilistic Topic Models on N-gram Graphs*

N-Gram Graphs – Other Applications

- Record Linkage
- Authorship Identification
- Text Classification
- Clustering and Indexing
- Text Stemmatic Analysis

Summary[2]

AutoSummENG

- Statistical
- Language-Neutral
- No Preprocessing Required
- Parametric (with Implemented Effective Parameter Estimation)

Combinatory Evaluation

- Better Results
- More Experiments Required
- Per Summary Evaluation
- Orthogonal Features for Regression

JInsect Toolkit containing AutoSummENG available under LGPL: `http://www.ontosum.org`. **Thank you**.

[2] also see [Giannakopoulos et al., 2008]

AutoSummENG – Evaluation 2005

Year – Evaluated Group	*Spearman*	*Pearson*	*Kendall*
2005 – Automatic peers	0.840 (0.0)	0.885 (0.0)	0.669 (0.0)
2005 – Human peers	0.936 (0.0)	0.878 (0.0)	0.854 (0.0)
2005 – All peers	0.929 (0.0)	0.977 (0.0)	0.803 (0.0)

Table: Correlation of AutoSummENG to the Responsiveness Metric of DUC 2005 for *Automatic peers, Human peers and All peers* using estimated parameters based on DUC 2005. Within parethenses the p-value of the corresponding test. Statistical importance lower than the 95% threshold are noted by *emphatic text* in the parentheses.

AutoSummENG - Evaluation 2006

Year – Evaluated Group	*Spearman*	*Pearson*	*Kendall*
2006 – Automatic peers	0.871 (0.0)	0.891 (0.0)	0.709 (0.0)
2006 – Human peers	0.759 (0.01)	0.715 (0.02)	0.566 (0.03)
2006 – All peers	0.937 (0.0)	0.967 (0.0)	0.806 (0.0)
2007 – Automatic peers	0.842 (0.0)	0.871 (0.0)	0.687 (0.0)
2007 – Human peers	0.659 (0.04)	0.673 (0.03)	0.442 (*0.08*)
2007 – All peers	0.925 (0.0)	0.966 (0.0)	0.792 (0.0)

Table: Correlation of AutoSummENG to the Content Responsiveness Metric of DUC 2006, 2007 for *Automatic peers, Human peers and All peers* using estimated parameters based on DUC 2005. Within parethenses the p-value of the corresponding test. Statistical importance lower than the 95% threshold are noted by *emphatic text* in the parentheses.

Textual Qualities

[Endres-Niggemeyer, 2000]:

- Cohesion (linguistic, syntactic and anaphoric integrity)
- Coherence (semantic and functional connectedness, which serves communication)
- Acceptability (the communicative ability of the text from the perspective of its addressees)
- Intentionality (ability of the text to contain the intention of the writer, *e.g.*exaggeration or question)
- Situationality (ability of the text to result into the expected interpretation within a specific context)
- Intertextuality (the ability of the text to link to other texts, preserving the presented information)
- Informativity (the novelty of the textual information)

AutoSummENG Detailed Settings for Experiments

Character: (3,3,3), (5,5,5), (7,7,7)
Word: (1,1,8), (2,2,8), (3,3,3)

Tools Devised and Implemented for General NLP Uses

- Statistical Chunker (Entropy of next character)
- Semantic Annotation (Dynamic Programming and Background Knowledge)
- Redundancy Removal

References

Conroy, J. M. and Dang, H. T. (2008).
Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality.
In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 145–152, Manchester, UK. Coling 2008 Organizing Committee.

Daume III, H. and Marcu, D. (2005).
Bayesian summarization at duc and a suggestion for extrinsic evaluation.
In *Proceedings of the Document Understanding Conf. Wksp. 2005 (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*.

Endres-Niggemeyer, B. (2000).
Human-style www summarization.

Giannakopoulos, G., Karkaletsis, V., Vouros, G., and Stamatopoulos, P. (2008).
Summarization system evaluation revisited: N-gram graphs.
ACM Trans. Speech Lang. Process., 5(3):1–39.

Hovy, E., Lin, C. Y., and Zhou, L. (2005).
Evaluating duc 2005 using basic elements.
Proceedings of DUC-2005.

Lin, C. Y. (2004).
Rouge: A package for automatic evaluation of summaries.
Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), pages 25–26.

Lin, C.-Y. and Hovy, E. (2003).
Automatic evaluation of summaries using n-gram co-occurrence statistics.
In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78, Morristown, NJ, USA. Association for Computational Linguistics.

Passonneau, R. J., McKeown, K., Sigelman, S., and Goodkind, A. (2006).
Applying the pyramid method in the 2006 document understanding conference.
In *Proceedings of Document Understanding Workshop (DUC)*.

Radev, D. R., Jing, H., and Budzikowska, M. (2000).
Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies.
ANLP/NAACL Workshop on Summarization.

Steinberger, J. and Jezek, K. (2004).
Using latent semantic analysis in text summarization and summary evaluation.
In *Proc. ISIM'A04*, pages 93–100.

Witten, I. and Frank, E. (2005).
Data Mining: Practical Machine Learning Tools and Techniques.