**1) Explain the differences between ROUGE scores**  
Evaluation of automatic text summarization is an active research question. Currently, ROUGE scores are the state-of-the art metrics used to evaluate automated text summarization (i.e DUC 2006[[1]](#footnote-1), DUC 2007[[2]](#footnote-2), TAC 2010[[3]](#footnote-3), TAC 2011[[4]](#footnote-4), and TAC 2014[[5]](#footnote-5)). It has been shown that they are very effective to measure document (Lin, 2004).

The ROUGE scores measure summary quality by counting the overlapping units such as n-gram between the candidate summary and the reference summary, as defined by the formula below:

Where stands for the length of n-gram and is the number of n-grams co-occurring in a candidate summary and a set of reference summaries.

We adopt ROUGE-1, ROUGE-2 and ROUGE-SU4 in our evaluation metrics, short for R1, R2, and R-SU4 respectively. The more the scores, the better the summarization.

* R1 is unigram (single word) based (N=1 in the formula above). It is used to evaluate the word coverage. If the automatic summary is highly overlapping the manual summary in terms of words, it will have a high R1 score. The word order is not considered.
* R2 is bigram (two continuous words) based (N=2). It is used to evaluate the bigram coverage.
* R-SU4 is an extension of R2. For R2, only bigram is considered and only two continuous words will count as a bigram. For R-SU4, it will count both the unigram and skip-bigram (any two words within a window of 4 words).

**2) Give examples by using “Teacher asked questions to Ashley”**  
Here is an example to illustrate how to compute R1, R2, R-SU4 for automatic summarization.

Here is the human summary (reference):

“teacher asked questions to Ashley”

For the reference summary: there are

5 unigrams: “teacher”, “asked”, “questions”, “to”, “Ashley”

4 bigrams: “teacher asked”, “asked questions”, “questions to” and “to Ashley”

10 skip-bigrams (up to a skip within 4 words): “teacher asked”, “teacher questions”, “teacher to”, “teacher Ashley”, “asked questions”, “asked to”, “asked Ashley”, “questions to”, “questions Ashley”, and “to Ashley”

And, there two candidates of automatic summarization.

S1: “teacher ask a question to Ashley”

S2: “Ashley ask questions to teacher”

The R1 score for S1 = 3/5 (“teacher”, “to”, “Ashley” are in S1 and there are total of 5 unigrams in the reference.)

The R1 score for S2 = 3/5 (“teacher”, “to”, “Ashley” are in S2)

The R2 score for S1 = 1/4 (only “to Ashley” is in S1 and there are 4 bigrams)

The R2 score for S2 = 1/4 (only “questions to” is in S1 and there are 4 bigrams)

3 out of the 5 unigrams appear in the S1; 3 of them appear in the S2

3 out of the 10 skip-bigram appear in S1 (“teacher to”, “teacher Ashley”, and “to Ashley”); Only 1 of them is in S2 (“questions to”)

So, the R-SU4 score for S1 is (3 + 3) / (5 + 10) = 0.4 (3 overlap unigrams + 5 overlap skip bigrams)/(total number of unigram and skip-bigram)

R-SU4 for S2 is (3+1)/(5 + 10) = 0.27

In this term, S1 is better than S2.

**3) Table with muddiest point only**  
The data set we use is introduced in (Menekse et al., 2011). The performances are shown in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Type of method | | R1 | R2 | R-SU4 |
| Extractive | Mead | 0.250 | 0.085 | 0.062 |
| Shallow | Unigram | 0.346 | 0.020 | 0.087 |
| Keyphrase | 0.239 | 0.087 | 0.056 |
| Phrase Clustering | **0.367** | **0.101** | **0.096** |

\* For the shallow summarization, we set R=30 for the experiments.

\* For phrase clustering, we used the toolkit Senna (<http://ml.nec-labs.com/senna/>) to extract the noun phrases, K-medoids as clustering algorithm (K = 0.5 \* total number of candidate phrases) and the semantic similarity metric (optimumComparerLSATasa) implemented in SEMILAR (Rus et al., 2013).

**4) Explain which method we select, why (phrase clustering)**   
For the table above, we can see the phrase clustering method performs best among all the three metrics. Therefore, we will pick up the “phrase clustering” as the summarization technique.

In addition, we believe phrase based shallow summarization is a better way for student reflects compared to other methods at least for three reasons. First, the limited screen size makes a summary with several sentences are hard to read especially when they are long. However, showing a summary of reflects back to students give them a chance to review peer feedback. Second, shallow summarization servers a dense compress of the student reflects which can provide more coverage compared to sentences summary with the same space limit. Third, human summaries of student reflects tend to be phrases and they could be easily understood by the instructors without providing full sentences.

**5) Which metric is best, why? (Explain these percentages by using some literature)**

From our perspective, R2 and R-SU4 are more important than R1 to evaluate the quality of the summarization of student reflects. We found even though the word output (unigram model) has a high R1 score but the instructors does not them because a list of individually words does not give enough context information. It means that a high R1 does not mean a high-quality summary.

From the ROUGE scores, the phrase clustering method performs very well. It means 10.1% of bigrams of the human reference are in the automatic summarization. It is worthy to note that the state-of-the-art performance of R2 score for DUC 2006 is around 0.086 ~ 0.092 (Wong et al., 2008; Wang et al., 2009; Haghighi and Vanderwende, 2009; Ouyang et al., 2010). In advance, the ROUGE-2 score for two human summaries on an opinion summarization tasks is 0.1069 (Ganesan et al., 2010).

**6) A better example**

A manually summary of muddiest point for one lecture:

[1] In class activities (Specially the first activity)

[2] Atomic Packing Factor and relation between a&r

[3] Which structure produce which materials' properties and how unit cell affects the processing

Extractive Summarization based on Mead

[1] Whethere we describe unit cell lengths in terms of a or c

[2] Unit cell transformation pictures in 1st activity, 2nd activity (characteristics) # of atomic radii along touch directions with in unice cell

[3] The atomic packing factor idea was very confusing and needed more explanation

Shallow Summarization

|  |  |  |
| --- | --- | --- |
| **Unigram** | **Keyphrase extraction** | **Phrase Clustering** |
| unit | Structure characteristics | how unit cell |
| atomic | atomic packing | our activities |
| cell | packing factor | the material |
| activity | unit cell | which structures |
| confusing | Whethere we describe | structure characteristics |
| factor | unit cell lengths | the math |
| activities | describe unit cell | atomic packing factor |
| processing | lengths in terms | apf |
| packing | describe unit | the properties |
| properties | Caculating structure characteristics | a little more explanation |
| length | atomic | the exact difference |
| apf | slides How unit | different crystal systems |
| structures |  |  |
| characteristics |  |  |
| fcc |  |  |

**Reference**

Li, C., Qian, X., & Yang, L. (2013). Using Supervised Bigram-based ILP for Extractive Summarization. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic Annual Meeting of the Association for Computational Linguistic*, *ACL*, 1004–1013. Retrieved from

Ganesan, K., Zhai, C., & Han, J. (2010). Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 340–348). Beijing, China: Association for Computational Linguistics.

Haghighi, A., & Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 362–370). Boulder, Colorado: Association for Computational Linguistics. Retrieved from <http://dl.acm.org/citation.cfm?id=1620754.1620807>

Ouyang, Y., Li, W., Lu, Q., & Zhang, R. (2010). A study on position information in document summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 919–927). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1944566.1944672

Wang, D., Zhu, S., Li, T., Chi, Y., & Gong, Y. (2008). Integrating clustering and multi-document summarization to improve document understanding. *Proceedings of the 17th ACM Conference on Information and Knowledge Management*.

Wang, D., Zhu, S., Li, T., & Gong, Y. (2009). Multi-document summarization using sentence-based topic models. *Proceedings of the ACL-IJCNLP 2009.*

1. http://www-nlpir.nist.gov/projects/duc/duc2006/tasks.html [↑](#footnote-ref-1)
2. http://duc.nist.gov/duc2007/tasks.html [↑](#footnote-ref-2)
3. http://www.nist.gov/tac/data/past/2010/Summ10.html [↑](#footnote-ref-3)
4. http://www.nist.gov/tac/data/past/2011/Summ11.html [↑](#footnote-ref-4)
5. http://www.nist.gov/tac/2014/BiomedSumm/index.html [↑](#footnote-ref-5)