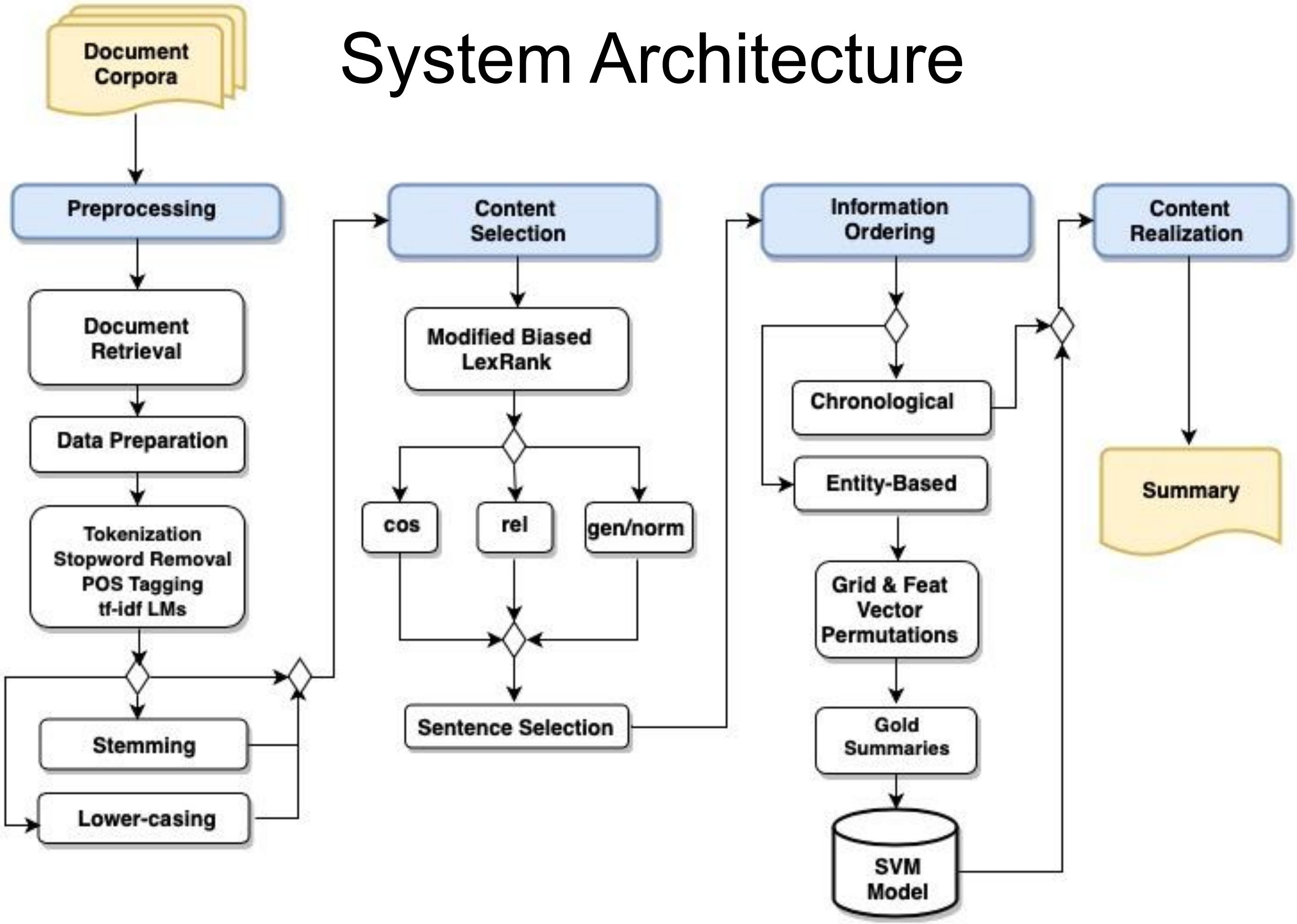
Nutshell

Shannon Ladymon, Haley Lepp, Ben Longwill, Amina Venton

System Architecture



Improvements in content selection

- Preprocessing:
  - Stemming: Bool
  - Lowercasing: Bool
  - TF: term frequency, log normalization
  - IDF: standard, smooth, or probabilistic
- Biased Lex-Rank:
  - Bias calculation method: cosine, relevance, generative
    - Cosine requires threshold similarity
  - Inter-sentential similarity weight: cosine, norm
  - Topic bias weight: float
  - Power Iteration Method average change in probabilities: float

- Sentence Selection:
  - Redundancy threshold: float

Biased LexRank (a refresher)

$$BLR(s|q) = d\frac{b(s|q)}{\sum\limits_{z\in C} b(z|q)} + (1-d)\sum_{v\in C}\frac{w(v,s)}{\sum\limits_{z\in C} w(z,v)}$$

Generative

$$P_{ML}(w|v) = \frac{tf_{w,v}}{|v|}$$

$$P_{norm}(u|v) = \left(\prod_{w\in u} P_{JM}(w|v)^{tf_{w,u}}\right)^{\frac{1}{|u|}}$$

$$P_{JM}(w|v) = (1-\lambda)P_{ML}(w|v) + \lambda P_{ML}(w|C)$$

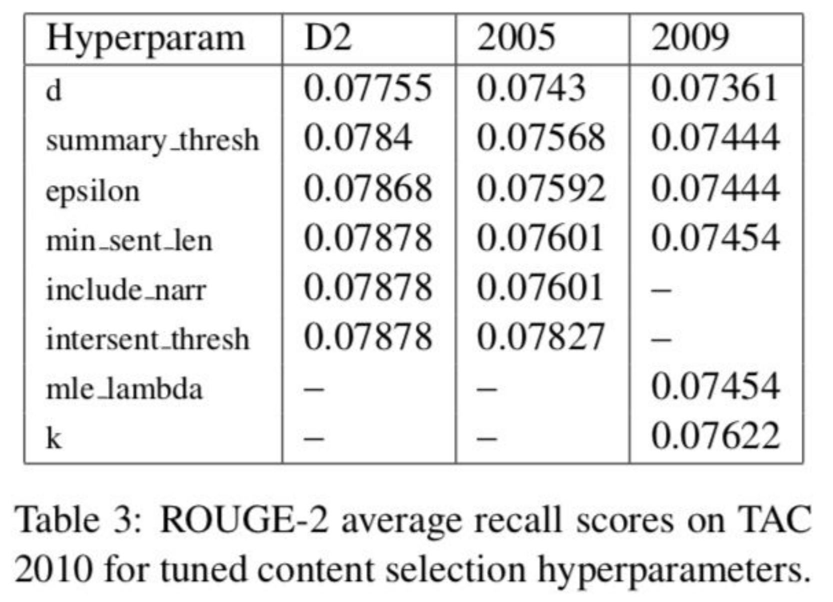
$$P_{gen}(u|v) = \prod_{w\in u} P_{JM}(w|v)^{tf_{w,u}}$$

Relevance

$$\mathrm{rel}(s|q) = \sum\nolimits_{w\in q} log(\mathrm{tf}_{w,s}+1) \times log(\mathrm{tf}_{w,q}+1) \times \mathrm{idf}_{w}$$

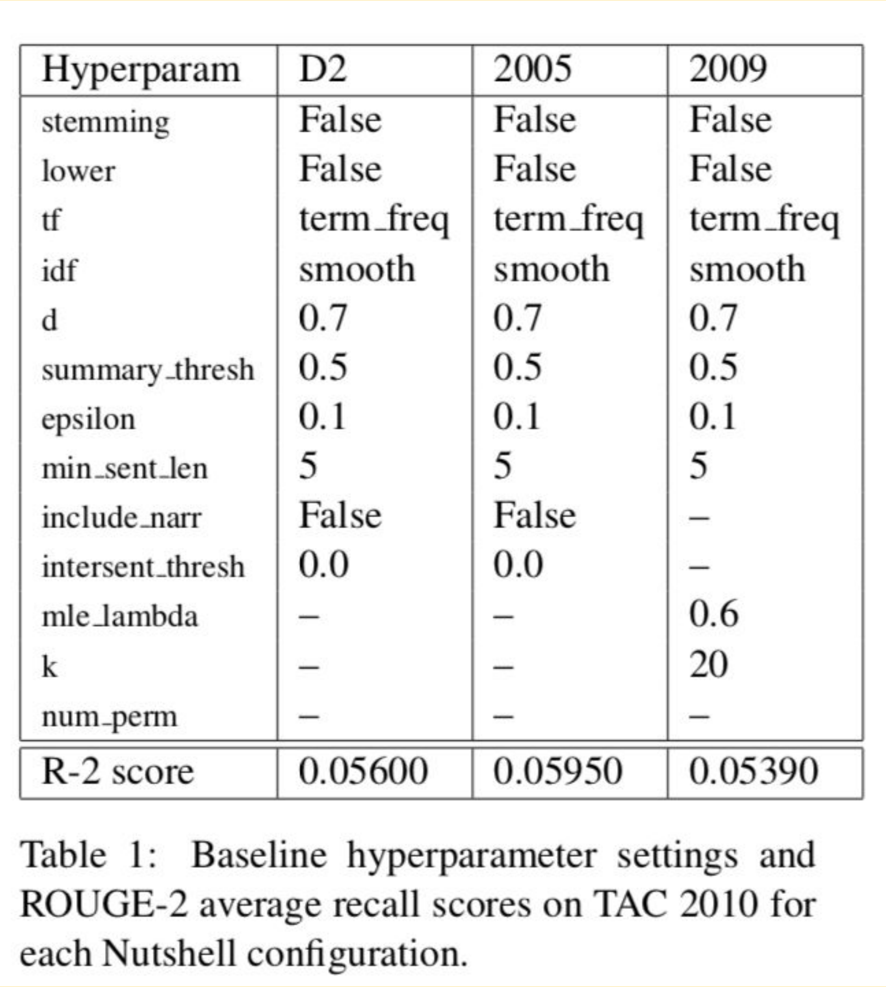
Hyperparameter Tuning

- Used 3 configurations:
  - D2 default (cos, cos)
  - Otterbacher et al. 2005 (rel, cos)
  - Otterbacher et al. 2009 (gen, norm)
- Tuned hyperparameters on each



Hyperparam	D2	2005	2009
d	0.07755	0.0743	0.07361
summary_thresh	0.0784	0.07568	0.07444
epsilon	0.07868	0.07592	0.07444
min_sent_len	0.07878	0.07601	0.07454
include_narr	0.07878	0.07601	–
intersent_thresh	0.07878	0.07827	–
mle_lambda	–	–	0.07454
k	–	–	0.07622

Table 3: ROUGE-2 average recall scores on TAC 2010 for tuned content selection hyperparameters.



Hyperparam	D2	2005	2009
stemming	False	False	False
lower	False	False	False
tf	term_freq	term_freq	term_freq
idf	smooth	smooth	smooth
d	0.7	0.7	0.7
summary_thresh	0.5	0.5	0.5
epsilon	0.1	0.1	0.1
min_sent_len	5	5	5
include_narr	False	False	–
intersent_thresh	0.0	0.0	–
mle_lambda	–	–	0.6
k	–	–	20
num_perm	–	–	–
R-2 score	0.05600	0.05950	0.05390

Table 1: Baseline hyperparameter settings and ROUGE-2 average recall scores on TAC 2010 for each Nutshell configuration.

Information ordering

Chronological Baseline:

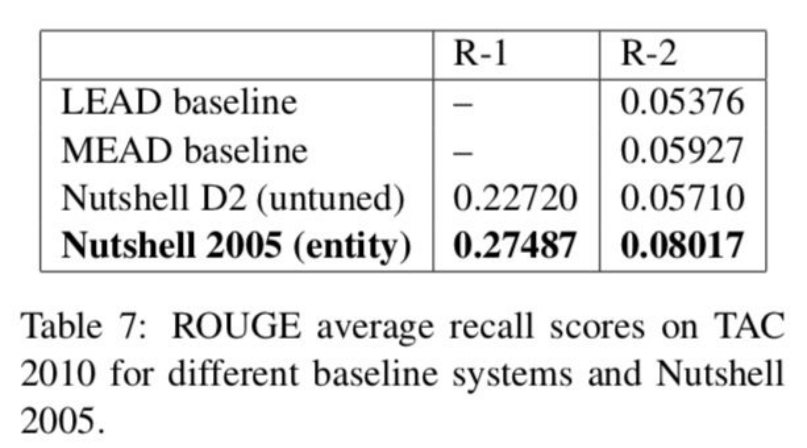
- Doc date and order within doc

Entity-based: Barzilay et. al (2008)

- Entity Grid
  - Coreference: exact match
  - Entities: either is present (X) or absent (–).
  - Transitions: length 2
  - Transition probability = sequence probability/total number of transitions
  - Permuted, these can be represented as feature vectors
- SVM
  - Joachims’(2006) SVMrank package

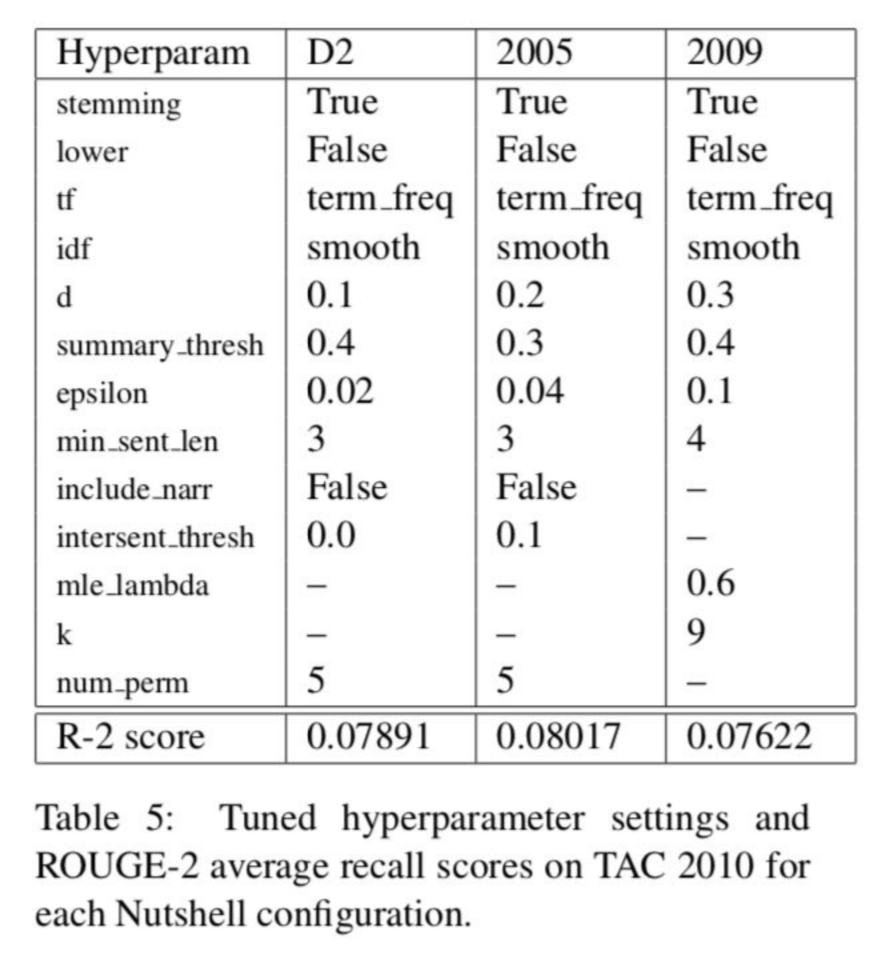
Issues and successes

- Tuned hyperparameters on three configurations
- Need to include Content Realization component
- Slight slow down but run time still seems reasonable



	R-1	R-2
LEAD baseline	–	0.05376
MEAD baseline	–	0.05927
Nutshell D2 (untuned)	0.22720	0.05710
Nutshell 2005 (entity)	**0.27487**	**0.08017**

Table 7: ROUGE average recall scores on TAC 2010 for different baseline systems and Nutshell 2005.



Hyperparam	D2	2005	2009
stemming	True	True	True
lower	False	False	False
tf	term_freq	term_freq	term_freq
idf	smooth	smooth	smooth
d	0.1	0.2	0.3
summary_thresh	0.4	0.3	0.4
epsilon	0.02	0.04	0.1
min_sent_len	3	3	4
include_narr	False	False	–
intersent_thresh	0.0	0.1	–
mle_lambda	–	–	0.6
k	–	–	9
num_perm	5	5	–
R-2 score	0.07891	0.08017	0.07622

Table 5: Tuned hyperparameter settings and ROUGE-2 average recall scores on TAC 2010 for each Nutshell configuration.

Future work clues

LITTLETON, Colo. (AP) – ...

“What’s the district attorney saying?

Debra Lafave, the former Tampa middle school teacher accused of having sex with a 14-year-old male student ... Debra Lafave, the former Tampa teacher accused of seducing a teenage boy.

But in Sichuan and in Shaanxi province, ...

By comparison, Floyd’s hurricane-force wind ...

Related reading which influenced your approach

- Barzilay and Lapata (2008)
- Otterbacher (2005)
- Joachims (2006)

References

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35– 55.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Gu¨nes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Annemarie Friedrich, Marina Valeeva, and Alexis Palmer. 2014. LQVSumm: A corpus of linguistic quality violations in multi-document summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1591–1599, Reykjavik, Iceland. European Language Resources Association (ELRA).

David Graff. 2002. The aquaint corpus of english news text. Web Download. Philadelphia: Linguistic Data Consortium.

T. Joachims. 2006. Training linear svms in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.* Philadelphia: Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schu˙tze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Vineeth G. Nair. 2014. *Getting Started with Beautiful Soup*. Packt Publishing.

Jahna Otterbacher, Gunes Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 915–922, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Jahna Otterbacher, Gunes Erkan, and Dragomir R. Radev. 2009. Biased lexrank: Passage retrieval using random walks with question-based priors. *Inf. Process. Manage.*, 45(1):42–54.

MF Porter. 1980. An algorithm for suffix stripping. *Program: Electronic Library and Information Systems*, 14.

Ellen Vorhees and David Graff. 2008. Aquaint-2 information-retrieval text research collection. Web Download. Philadelphia: Linguistic Data Consortium.