May 15, 2018

Extractive Summarization

A case study in tl;dr systems

Deliverable 3
Group 4

Ayushi Aggarwal, Elijah Rippeth, Seraphina Goldfarb-Tarrant

LING573, Spring 2018

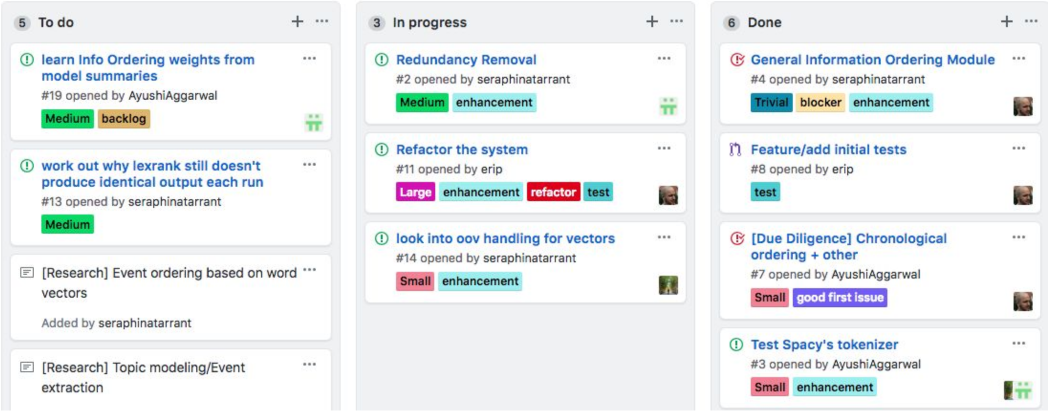
Walkthrough

1. Workflow Changes
2. Information Ordering
3. Improvements in Content Selection
4. Analyses
    a. Tokenization
    b. spaCy
    c. Hypothesis Testing
5. Observations - Successes and Issues
6. Scope for D4
7. Related reading which influenced your approach

Workflow Changes

- Git flow style workflow
  - feature/add-blah
- Github issue tracking/kanban

- At least one LGTM for a PR to be merged
- More communication
- Unit testing



5 To do

learn Info Ordering weights from model summaries
#19 opened by AyushiAggarwal
Medium backlog

work out why lexrank still doesn't produce identical output each run
#13 opened by seraphinatarrant
Medium

[Research] Event ordering based on word vectors
Added by seraphinatarrant

[Research] Topic modeling/Event extraction

3 In progress

Redundancy Removal
#2 opened by seraphinatarrant
Medium enhancement

Refactor the system
#11 opened by erip
Large enhancement refactor test

look into oov handling for vectors
#14 opened by seraphinatarrant
Small enhancement

6 Done

General Information Ordering Module
#4 opened by seraphinatarrant
Trivial blocker enhancement

Feature/add initial tests
#8 opened by erip
test

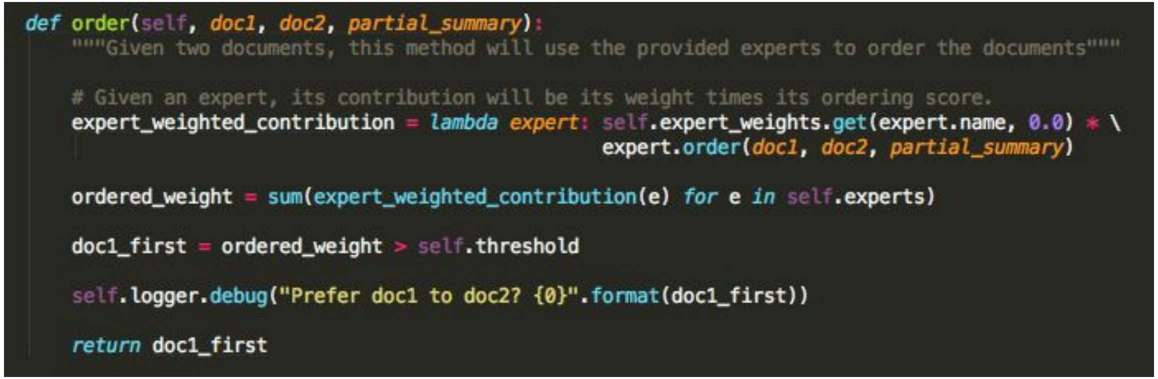
[Due Diligence] Chronological ordering + other
#7 opened by AyushiAggarwal
Small good first issue

Test Spacy's tokenizer
#3 opened by AyushiAggarwal
Small enhancement

Information Ordering (Proof of Concept)

- Adapted from information ordering experts
- Developed a chronological expert as a proof of concept for qualitative improvements

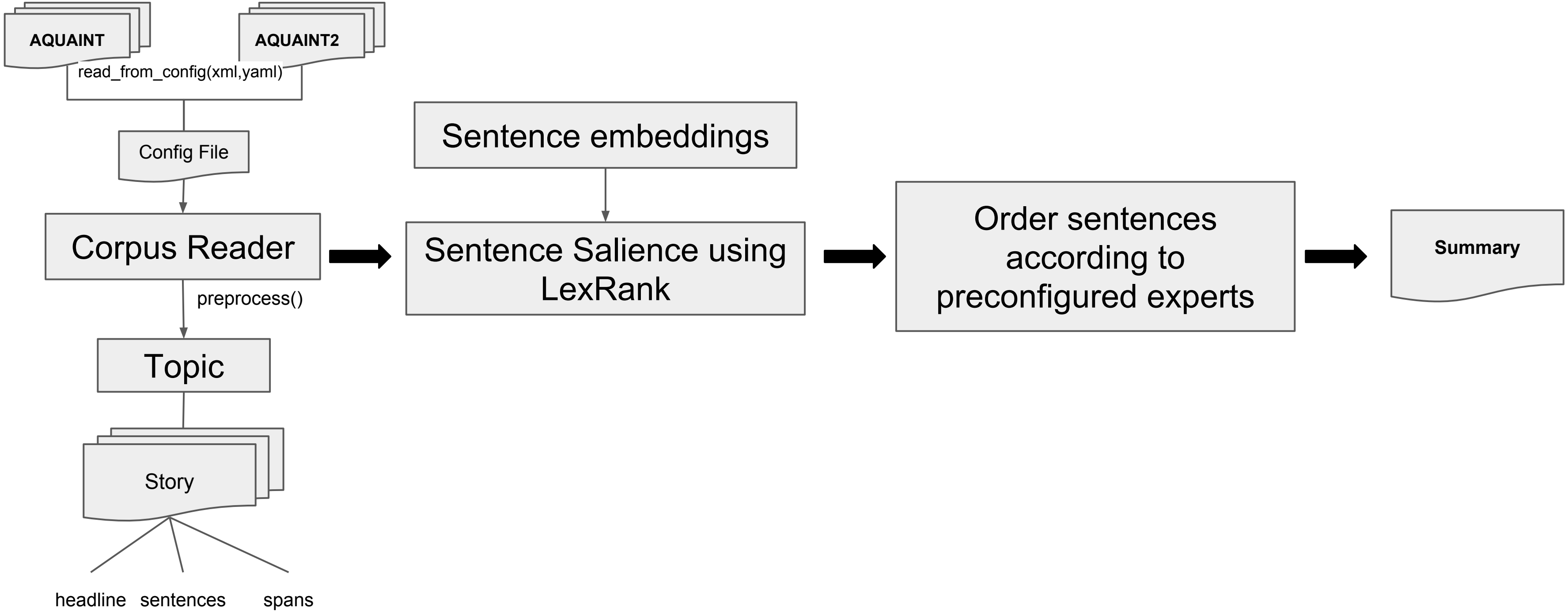
Information Ordering (Proof of Concept)



```
def order(self, doc1, doc2, partial_summary):
    """Given two documents, this method will use the provided experts to order the documents"""

    # Given an expert, its contribution will be its weight times its ordering score.
    expert_weighted_contribution = lambda expert: self.expert_weights.get(expert.name, 0.0) * \
                                                  expert.order(doc1, doc2, partial_summary)

    ordered_weight = sum(expert_weighted_contribution(e) for e in self.experts)

    doc1_first = ordered_weight > self.threshold

    self.logger.debug("Prefer doc1 to doc2? {0}".format(doc1_first))

    return doc1_first
```

- Extensibility in mind for the orderer
- Weights are manually determined currently

System Architecture Now



Recap: Content Selection in D2

Three baselines:

- Random sentence
- First sentence
- LexRank
    - Tf-idf of stemmed words

Problems we wanted to address:

- Tokenization
- Different sentence representations
    - To see if they do not overweight rare words and long sentences
    - To incorporate additional elements of sentence representation

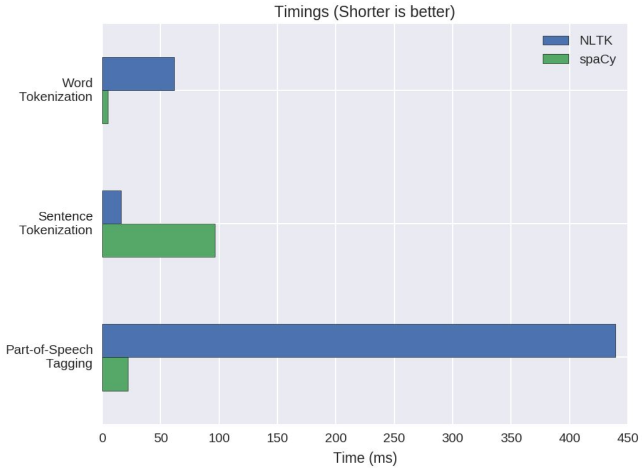
Tokenization

Punkt

- multilingual*
  - English, Brazilian Portuguese, Dutch, Estonian, French, German, Italian, Norwegian,
  - Spanish, Swedish, Turkish
  - equally(ish) good for all of them (~90 F-measures)!
- unsupervised *abbreviation classification* based on collocation and sentence position characteristics
  - Assumes that source of boundary detection issues is abbreviations, only applicable to languages that have this property
- fast

SpaCy

- slow!
- dependency parse
- good at fixing the ways that Punkt is weak: where collocational abbreviation detection is hard
    - *in.* for *inch* will always be detected as not an abbreviation
    - “helicopters were brought in.” vs. “15 in. with photo”



graph from https://blog.thedataincubator.com/2016/04/nltk-vs-spacy-natural-language-processing-in-python/, on 10kb of Wikipedia text in 2016.*

But what is a sentence?

Abstractly, LexRank operates on sentence similarity. We explored:

- tf-idf
    - Reminder: ‘term frequency’ is within sentence and ‘document frequency’ is within topic
- spaCy GloVe vectors (trained on Common Crawl)
- doc2vec (trained on Gigaword)

+ cosine similarity

Hypothesis 1: Extending beyond TF-IDF could help

- TF-IDF captures awareness of:
    - the context of the sentence
    - the context of the document cluster (or *Topic*)

Maybe making use of more data will be helpful!

- GloVe captures global collocational information from *the internet*.
- doc2vec captures a narrow window of context, but includes the parent paragraph and is trained on matching domain (Gigaword, superset of AQUAINT).

Hypothesis 2:

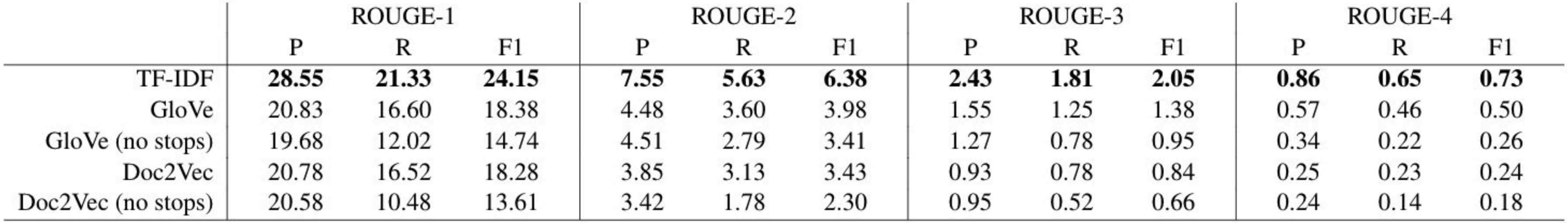
Information Ordering will have a negligible effect on ROUGE scores

- The ROUGE window we use is too small (max 4)
- The sentences are too long (and thus there are not enough of them in summary)

ROUGE Reality:

H1 ✗ H2 ✓

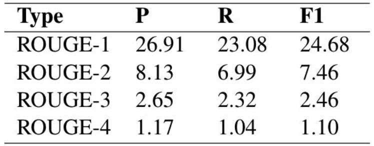
- TF-IDF beats everything
- Ordering doesn't matter



	ROUGE-1			ROUGE-2			ROUGE-3			ROUGE-4		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
TF-IDF	**28.55**	**21.33**	**24.15**	**7.55**	**5.63**	**6.38**	**2.43**	**1.81**	**2.05**	**0.86**	**0.65**	**0.73**
GloVe	20.83	16.60	18.38	4.48	3.60	3.98	1.55	1.25	1.38	0.57	0.46	0.50
GloVe (no stops)	19.68	12.02	14.74	4.51	2.79	3.41	1.27	0.78	0.95	0.34	0.22	0.26
Doc2Vec	20.78	16.52	18.28	3.85	3.13	3.43	0.93	0.78	0.84	0.25	0.23	0.24
Doc2Vec (no stops)	20.58	10.48	13.61	3.42	1.78	2.30	0.95	0.52	0.66	0.24	0.14	0.18

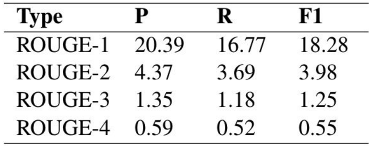
Table 5: ROUGE results with LexRank for different sentence embeddings.

D2 Reminder:



Type	P	R	F1
ROUGE-1	26.91	23.08	24.68
ROUGE-2	8.13	6.99	7.46
ROUGE-3	2.65	2.32	2.46
ROUGE-4	1.17	1.04	1.10

Table 2: ROUGE scores for first-sentence baseline run on DevTest data



Type	P	R	F1
ROUGE-1	20.39	16.77	18.28
ROUGE-2	4.37	3.69	3.98
ROUGE-3	1.35	1.18	1.25
ROUGE-4	0.59	0.52	0.55

Table 1: ROUGE scores for random-sentence baseline run on DevTest data

Observed Successes

- Sentence ordering improved quality of the summaries
  - Some coherence visible
- Doc2Vec shows improvement in summary quality
  - Qualitatively: doc2Vec without stopwords > doc2Vec with stopwords >> spacy without stopwords > spacy with stopwords
    - Less quoted statements
    - Sentences in summaries are more informative - Major win for salience module!

SpaCy vs Doc2Vec - Example I

Doc2vec *without* stopword removal

When they came in February, shortly after their son's killing, they attained near-celebrity status _ trailed by television cameras, greeted by crowds and transported from place to place by motorcade.
What happens next could intensify, rather than quell, the racial tensions that have been simmering in the two months since the shooting.

SpaCy *without* stopword removal

That case is set to begin on Tuesday.
and another falling down some steps as if he had been shot.
Friday, acting Justice Patricia Anne Williams, who refused defense requests to put off scheduling a date for the case, chose the Jan. 3 date after the defense lawyers said they would need time to go through the prosecution's evidence and discuss possible pretrial motions.
It was a point that Worth disputed outside court by saying, ``I don't think I have a conflict of interest, though it is obvious that Kornberg would like for me to go.''

Doc2Vec summaries end with a concluding statement, often taking a leap in discourse

SpaCy grants higher salience to quoted sentences

Understandably, summaries could benefit from(As seen in example outputs):
- Discourse coherence
- Coreference resolution
- Anaphora resolution
- Sentence length normalization

SpaCy vs Doc2Vec - Example II

Doc2vec *with* stopword removal

The lawyers also said there could be pretrial scheduling conflicts because of other trials some of them are involved in, like the Abner Louima case, in which officers are accused of brutalizing a Haitian immigrant.
Police officials declined to say where the officers _ Kenneth Boss, Richard Murphy, Carroll and McMellon _ would be working, adding only that they were on modified duty and carrying no guns or badges.
They left the courtroom to the cheers of nearly 200 other officers, gathered outside the Bronx County Building.

SpaCy *with* stopword removal

``We are waiting for this, for justice,'' he said, ``and we feel strong.''
The four officers fired 41 shots, hitting Diallo 19 times.
The Rev. Al Sharpton, who organized the daily protests over the Diallo shooting and is an adviser to the family, also is expected to attend the meeting, said Steven Reed, Johnson's spokesman.

Possible solutions:

- Use discourse features
- Semantic features
- Include more informative shorter sentences by normalizing sentence scores for sentence length

Foreseeable Challenges

Using Discourse Features:

- Primarily supervised method; possibly semi-supervised
- Human annotated discourse features are not scalable for a large dataset

Using Semantic features: feasibility of extracting these features and the payoff of the process is to be explored;
What we know:

- They are less helpful towards selecting summaries
- Usually indicative of non-summary sentences

Observed Issues

- Information Ordering does not improve ROUGE scores since content does not change...so it’s hard to tune!
- ROUGE scores do **not** seem correlated with qualitative improvements
- Lack of apple-to-apples comparison makes drawing conclusions about the suitability of different sentence representations difficult
    - Control for training corpora, and for method of representing sentence from word vectors
- No upper bound for LexRank, or even for Extractive summaries
- Different sentence representations made LexRank *worse* than random

Scope for D4

- Refactoring is becoming more important as the size and complexity of the codebase grows.
  - Huge inroads made in pursuit of D3!
- Explore compression strategies
  - Considering thresholded embedding proximity within content selection.
  - With additional sentence compression, ordering will help more
- Normalize sentence salience score to remove bias to longer sentences
- Topic relevance
- Profit

Related Reading

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 238-247).

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. Computational Linguistics, 32(4), 485-525.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Q&A