May 29, 2018

Extractive Summarization

A case study in tl;dr systems

Deliverable 4
Group 4
Ayushi Aggarwal, Elijah Rippeth, Seraphina Goldfarb-Tarrant
LING573, Spring 2018

Walkthrough

1. Refactor
2. Redundancy removal
3. Other content selection
4. Sentence normalization
5. Related reading which influenced your approach

Refactor

- Add abstractions so future work would be more extensible
- Drive completely with configs
  - Command-line arguments are painful and unmaintainable
- Use more open-source software for battle-tested warm-and-fuzzies.
- Use Python 3.5+ `typing` module for lenting and type-hinting
  - Easier to spot-check the method's inputs and outputs.
  - IDE warnings if types don't match.
- Make code easier to profile
  - Came in handy in this assignment. :-)

Refactor (the hard parts)

```
Scanning for model files...
Scanning for system output files...
Opening "../results/rouge_config_exp1.xml" for writing...
Success.

real    111m40.922s
user    111m30.079s
sys     0m19.924s
```

erip 3:24 PM
loool

Seraphina 3:24 PM
ahahahahahaa

Refactor (and profiling)

After the first 10 min of running, the culprit was easily identified:

```
ncalls  tottime  percall  cumtime  percall filename:lineno(function)
26805235   68.063    0.000  128.222    0.000 porter.py:248(_apply_rule_list)
```

Redundancy Removal

Removing duplicate/near-duplicate sentences before generating candidate sentences

Expectations:

- Precision should remain similar, if not same
- Recall should improve

Redundancy Removal: Detecting Paraphrase

- SemEval task 2017: rank sentence pairs: 0-5 exact paraphrases to entirely dissimilar
  - 4 and 5 should be removed, 3 is too far
- Multi-lingual! (task set up with MT in mind)
- Supervised, a lot of data overlap (Newswire training, SNLI with custom pairs eval)
- Baseline: cosine binary sentence vectors. Strategies from weighted word vectors to DL.

5	*The two sentences are completely equivalent, as they mean the same thing.*
	The bird is bathing in the sink. Birdie is washing itself in the water basin.
4	*The two sentences are mostly equivalent, but some unimportant details differ.*
	Two boys on a couch are playing video games. Two boys are playing a video game.
3	*The two sentences are roughly equivalent, but some important information differs/missing.*
	John said he is considered a witness but not a suspect. "He is not a suspect anymore." John said.
2	*The two sentences are not equivalent, but share some details.*
	They flew out of the nest in groups. They flew into the nest together.
1	*The two sentences are not equivalent, but are on the same topic.*
	The woman is playing the violin. The young lady enjoys listening to the guitar.
0	*The two sentences are completely dissimilar.*
	The black dog is running through the snow. A race car driver is driving his car through the mud.

Results

- Winner: ensemble of 4 deep learning models and 3 feature engineered ML models
- Just as good for English: WordNet (vanilla) Sentence IC!

STS 2017 Participants on STS Benchmark			
Name	Description	Dev	Test
ECNU	Ensembles well performing feature eng. models with deep neural networks each using sent. emb. from either LSTM, DAN, prj. word emb. or avg. word emb. (Tian et al., 2017)	84.7	81.0
BIT	Ensembles sent. information content (IC) with cosine of sent. emb. derived from summed word emb. with IDF weighting scheme (Wu et al., 2017)	82.9	80.9
DT_TEAM	Ensembles feature eng. and deep learning signals using sent. emb. from DSSM, CDSSM and skip-thought models (Maharjan et al., 2017)	83.0	79.2
UdL	Feature eng. model using cosine of tf-idf weighted char n-grams, num. match, sent. length and avg. word emb. cosine over PoS and NER based alignments (Al-Natsheh et al., 2017)	72.4	79.0
HCTI	Deep learning model with sent. emb. computed using paired convolutional neural networks (CNN) and then compared using fully connected layers (Shao, 2017)	83.4	78.4
RTM	Referential translation machines (RTM) use a feature eng. model with transductive learning and parallel feature decay algorithm (ParFDA) training instance selection (Biçici, 2017b,a)	73.2*	70.6
SEF@UHH	Cosine of paragraph vector (PV-DBOW) sent. emb. (Duma and Menzel, 2017)	61.6	59.2
Sentence Level Baselines			
InferSent	Sent. emb. from bi-directional LSTM trained on SNLI (Conneau et al., 2017)	80.1	75.8
Sent2Vec	Word & bigram emb. sum from sent. spanning CBOW (Pagliardini et al., 2017)	78.7	75.5
SIF	Weighted word emb. sum with principle component removal (Arora et al., 2017)	80.1	72.0
PV-DBOW	Paragraph vectors (PV-DBOW) (Le and Mikolov, 2014; Lau and Baldwin, 2016)	72.2	64.9
C-PHRASE	Word emb. sum from model of syntactic constituent context words (Pham et al., 2015)	74.3	63.9
Averaged Word Embedding Baselines			
LexVec	Weighted matrix factorization of PPMI (Salle et al., 2016a,b)	68.9	55.8
FastText	Skip-gram with sub-word character n-grams (Joulin et al., 2016)	65.2	53.9
Paragram	Paraphrase Database (PPDB) fit word embeddings (Wieting et al., 2015)	63.0	50.1
GloVe	Word co-occurrence count fit embeddings (Pennington et al., 2014)	52.4	40.6
Word2vec	Skip-gram prediction of words in a context window (Mikolov et al., 2013a,b)	70.0	56.5

* 10-fold cross-validation on combination of dev and training data.

Table 14: STS Benchmark. Pearson's $r \times 100$ results for select participants and baseline models.

Can we use these systems?

Pairs	Human	DT_Team	ECNU	BIT	FCICU	ITNLP-AiKF
There is a cook preparing food. A cook is making food.	5.0	4.1	4.1	3.7	3.9	4.5
The man is in a deserted field. The man is outside in the field.	4.0	3.0	3.1	3.6	3.1	2.8
A girl in water without goggles or a swimming cap. A girl in water, with goggles and swimming cap.	3.0	4.8	4.6	4.0	4.7	0.1
A man is carrying a canoe with a dog. A dog is carrying a man in a canoe.	1.8	3.2	4.7	4.9	5.0	4.6
There is a young girl. There is a young boy with the woman.	1.0	2.6	3.3	3.9	1.9	3.1
The kids are at the theater watching a movie. it is picture day for the boys	0.2	1.0	2.3	2.0	0.8	1.7

Table 12: Difficult English sentence pairs (Track 5) and scores assigned by top performing systems.[20]

*Simpler paragraph vector or weighted embedding similarity (by POS and IDF) performed well on other languages but wasn't tried in any overall systems.

Other options?

Data	Model	MSRP (Acc / F1)	MR	CR	SUBJ	MPQA	TREC
Unordered Sentences (Toronto Books: 70m sents, 0.9B words)	SAE	74.3 / 81.7	62.6	68.0	86.1	76.8	80.2
	SAE+embs.	70.6 / 77.9	73.2	75.3	89.8	86.2	80.4
	SDAE	**76.4 / 83.4**	67.6	74.0	89.3	81.3	77.6
	SDAE+embs.	73.7 / 80.7	**74.6**	**78.0**	**90.8**	**86.9**	78.4
	ParagraphVec DBOW	72.9 / 81.1	60.2	66.9	76.3	70.7	59.4
	ParagraphVec DM	73.6 / 81.9	61.5	68.6	76.4	78.1	55.8
	Skipgram	69.3 / 77.2	73.6	77.3	89.2	85.0	82.2
	CBOW	67.6 / 76.1	73.6	7730	89.1	85.0	82.2
	Unigram TFIDF	**73.6 / 81.7**	73.7	79.2	90.3	82.4	**85.0**
Ordered Sentences (Toronto Books)	SkipThought	**73.0 / 82.0**	**76.5**	**80.1**	**93.6**	**87.1**	**92.2**
	FastSent	72.2 / 80.3	70.8	78.4	88.7	80.6	76.8
	FastSent+AE	71.2 / 79.1	71.8	76.7	88.8	81.5	80.4

Further experimentation with Tokenization

- Oracle experiment: improvements in tokenization improved Precision
  - TF-IDF - 13%
  - GloVe - 7%
  - Doc2vec - 6%
- But hurt recall (more longer sentences) so no change in F1.

Hypothesis:

- if redundancy reduction or sentence normalization can help Recall, perhaps the two in combination would help.

Sentence length normalization

- Attempted to normalize sentences by length to prevent profoundly long sentences.
  - Many shorter sentences may pay off more than a single long sentence.
- Was it successful?
  - No; hyper short sentences selected instead and ROUGE scores worsened

ROUGE before and after

Before:

4 ROUGE-1 Average_R: 0.17299 (95%-conf.int. 0.15361 - 0.19216)
4 ROUGE-1 Average_P: 0.33135 (95%-conf.int. 0.30489 - 0.35619)
4 ROUGE-1 Average_F: 0.22301 (95%-conf.int. 0.20040 - 0.24439)
4 ROUGE-2 Average_R: 0.04579 (95%-conf.int. 0.03704 - 0.05539)
4 ROUGE-2 Average_P: 0.08476 (95%-conf.int. 0.07047 - 0.09909)
4 ROUGE-2 Average_F: 0.05874 (95%-conf.int. 0.04795 - 0.07003)
4 ROUGE-3 Average_R: 0.01537 (95%-conf.int. 0.00954 - 0.02239)
4 ROUGE-3 Average_P: 0.02774 (95%-conf.int. 0.01820 - 0.03884)
4 ROUGE-3 Average_F: 0.01958 (95%-conf.int. 0.01253 - 0.02798)
4 ROUGE-4 Average_R: 0.00620 (95%-conf.int. 0.00209 - 0.01181)
4 ROUGE-4 Average_P: 0.01052 (95%-conf.int. 0.00377 - 0.01958)
4 ROUGE-4 Average_F: 0.00777 (95%-conf.int. 0.00269 - 0.01468)

After:

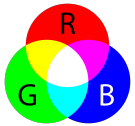
4 ROUGE-1 Average_R: 0.14384 (95%-conf.int. 0.12455 - 0.16346)
4 ROUGE-1 Average_P: 0.18709 (95%-conf.int. 0.16409 - 0.21154)
4 ROUGE-1 Average_F: 0.16176 (95%-conf.int. 0.14057 - 0.18366)
4 ROUGE-2 Average_R: 0.02554 (95%-conf.int. 0.01783 - 0.03389)
4 ROUGE-2 Average_P: 0.03256 (95%-conf.int. 0.02277 - 0.04337)
4 ROUGE-2 Average_F: 0.02849 (95%-conf.int. 0.02018 - 0.03778)
4 ROUGE-3 Average_R: 0.00743 (95%-conf.int. 0.00389 - 0.01179)
4 ROUGE-3 Average_P: 0.00924 (95%-conf.int. 0.00479 - 0.01448)
4 ROUGE-3 Average_F: 0.00821 (95%-conf.int. 0.00429 - 0.01300)
4 ROUGE-4 Average_R: 0.00262 (95%-conf.int. 0.00082 - 0.00508)
4 ROUGE-4 Average_P: 0.00326 (95%-conf.int. 0.00100 - 0.00632)
4 ROUGE-4 Average_F: 0.00290 (95%-conf.int. 0.00091 - 0.00559)

What about the quality?

- Faulkneresque.
- Long sentences penalized.
  - Short sentences rewarded!

```
The 911 tapes are released.
By Sam Howe Verhovek.
a man shouted.
``God Bless you guys!''
``They wanted death and destruction.
Hundreds lined the procession route.
thousands replied.
``They can't kill the message.
``Columbine!''
Religious leaders urged trust in God.
They can't kill the hope.''
The community outpouring has touched some Columbine students.
And that somehow, life keeps moving.
``Please comfort them and be with them, and comfort the
people of this town.
``Nobody knows where to go from here.
Many wonder what to do now.
Littleton needs comfort.
Littleton needs comfort.
It wasn't.
they sang.
she yelled.
```

What we would like to do differently



VERT or BLANC

- ROUGE results from D3 made us wary of tuning to it
- More useful than fluency: event coverage
- Also a SemEval task: TempEval 2013 Task B on TBAQ
- Good enough! Maxent/CRF/SVM

	F1	P	R	class F1
ATT-1	**81.05**	81.44	80.67	**71.88**
ATT-2	80.91	81.02	**80.81**	71.10
KUL	79.32	80.69	77.99	70.17
ATT-3	78.63	**81.95**	75.57	69.55
KUL-TE3RunABC	77.11	77.58	76.64	68.74
ClearTK-3,4	78.81	81.40	76.38	67.87
NavyTime-1	80.30	80.73	79.87	67.48
ClearTK-1,2	77.34	81.86	73.29	65.44
NavyTime-2	79.37	80.52	78.26	64.81
Temp:ESAfeature	68.97	78.33	61.61	54.55
JU-CSE	78.62	80.85	76.51	52.69
Temp:WordNetfeature	63.90	78.90	53.69	50.00
FSS-TimEx	65.06	63.13	67.11	42.94
TIPSem (TE2)	82.89	83.51	82.28	75.59

Table 6: Task B - Event Extraction Performance.

New Related Reading

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity-Multilingual and Cross-lingual Focused Evaluation. arXiv preprint arXiv:1708.00055.

De Boom, C., Van Canneyt, S., Demeester, T., & Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. Pattern Recognition Letters, 80, 150-156.

Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (Vol. 2, pp. 1-9).

Q&A