Detecting (Un)Important Content
for Single-Document News Summarization

Yinfei Yang
Redfin Inc.
Seattle, WA 98101
yangyin7@gmail.com

Forrest Sheng Bao
University of Akron
Akron, OH 44325
forrest.bao@gmail.com

Ani Nenkova
University of Pennsylvania
Philadelphia, PA 19104
nenkova@seas.upenn.edu

Abstract

We present a robust approach for detecting intrinsic sentence importance in news, by training on two corpora of document-summary pairs. When used for single-document summarization, our approach, combined with the "beginning of document" heuristic, outperforms a state-of-the-art summarizer and the beginning-of-article baseline in both automatic and manual evaluations. These results represent an important advance because in the absence of cross-document repetition, single document summarizers for news have not been able to consistently outperform the strong beginning-of-article baseline.

1 Introduction

To summarize a text, one has to decide what content is important and what can be omitted. With a handful of exceptions (Svore et al., 2007; Berg-Kirkpatrick et al., 2011; Kulesza and Taskar, 2011; Cao et al., 2015; Cheng and Lapata, 2016), modern summarization methods are unsupervised, relying on on-the-fly analysis of the input text to generate the summary, without using indicators of intrinsic importance learned from previously seen document-summary pairs. This state of the art is highly unintuitive, as it stands to reason that some aspects of importance are learnable. Recent work has demonstrated that indeed supervised systems can perform well without sophisticated features when sufficient training data is available (Cheng and Lapata, 2016).

In this paper we demonstrate that in the context of news it is possible to learn an accurate predictor to decide if a sentence contains content that is *summary-worthy*. We show that the predictors built in our approach are remarkably consistent, providing almost identical predictions on a held out test set, regardless of the source of training data. Finally we demonstrate that in single-document summarization task our predictor, combined with preference for content that appears at the beginning of the news article, results in a summarizer significantly better than a state-of-the-art global optimization summarizer. The results hold for both manual and automatic evaluations.

In applications, the detector of unimportance that we have developed can potentially improve snippet generation for news stories, detecting if the sentences at the beginning of the article are likely to form a good summary or not. This line of investigation was motivated by our previous work showing that in many news sub-domains the beginning of the article is often an uninformative teaser which is not suitable as an indicative summary of the article (Yang and Nenkova, 2014).

2 Corpora

One of the most cited difficulties in using supervised methods for summarization has been the lack of suitable corpora of document-summary pairs where each sentence is clearly labeled as either important or not (Zhou and Hovy, 2003). We take advantage of two currently available resources: archival data from the Document Understanding Conferences (DUC) (Over et al., 2007) and the New York Times (NYT) corpus (https://catalog.ldc.upenn.edu/LDC2008T19). The DUC data contains document-summary pairs in which the summaries were produced for research purposes during the preparation of a shared task for summarization. The NYT dataset contains thousands such pairs and the summaries were written by information scientists working for the newspaper.

DUC2002 is the latest dataset from the DUC series in which annotators produced extractive summaries, consisting of sentences taken directly from the input. DUC2002 contains 64 document sets.

The annotators created two extractive summaries for two summary lengths (200 and 400 words), for a total of four extracts per document set. In this work, a sentence from the original article that appears in at least one of the human extracts is labeled as *important* (summary-worthy). All other sentences in the document are treated as *unlabeled*. Unlabeled sentences could be truly not summary-worthy but also may be included into a summary by a different annotator (Nenkova et al., 2007). We address this possibility in Section 3, treating the data as partially labeled.

For the NYT corpus, we work with 19,086 document-summary pairs published between 1987 and 2006 from the Business section.

Table 3 in Section 5 shows a summary from the NYT corpus. These are abstractive, containing a mix of informative sentences from the original article along with abstractive re-telling of the main points of the article, as well as some meta-information such as the type of article and a list of the photos accompanying the article. It also shows the example of lead (opening) paragraph along with the summary created by the system we propose, InfoFilter, with the unimportant sentence removed.

In order to label sentences in the input, we employee Jacana (Yao et al., 2013) for word alignment in mono-lingual setting for all pairs of article-summary sentences. A sentence from the input is labeled as *important* (summary-worthy) if the alignment score between the sentence and a summary sentence is above a threshold, which we empirically set as 14 based on preliminary experiments. All other sentences in the input are treated as *unlabeled*. Again, an unlabeled sentence could be positive or negative.

3 Method

As mentioned earlier, existing datasets contain clear labels only for positive sentences. Due to the variability of human choices in composing a summary, unlabeled sentences cannot be simply treated as negative. For our supervised approach to sentence importance detection, a semi-supervised approach is first employed to establish labels.

3.1 Learning from Positive and Unlabeled Samples

Learning from positive (e.g., *important* in this paper) and unlabeled samples can be achieved by the

methods proposed in (Lee and Liu, 2003; Elkan and Noto, 2008). Following (Elkan and Noto, 2008), we use a two-stage approach to train a detector of sentence importance from positive and unlabeled examples.

Let y be the importance prediction for a sample, where $y = 1$ is expected for any positive sample and $y = 0$ for any negative sample. Let o be the ground-truth labels obtained by the method described in Section 2, where $o = 1$ means that the sentence is labeled as positive (important) and $o = 0$ means unlabeled.

In the first stage, we build an estimator e, equal to the probability that a sample is predicted as positive given that it is indeed positive, $p(o = 1|y = 1)$. We first train a logistic regression (LR) classier with positive and unlabeled samples, treating the unlabeled samples as negative. Then e can be estimated as $\Sigma_{x \in P}(LR(x)/|P|)$, where P is the set of all labeled positive samples, and $LR(x)$ is the probability of a sample x being positive, as predicted by the LR classifier. We then calculate $p(y = 1|o = 0)$ using the estimator e, the probability for an unlabeled sample to be positive as: $w = \frac{LR(x)}{e} / \frac{1-LR(x)}{1-e}$. A large w means an unlabeled sample is likely to be positive, whereas a small w means the sample is likely to be negative.

In the second stage, a new dataset is constructed from the original dataset. We first make two copies of every unlabeled sample, assigning the label 1 with weight w to one copy and the label 0 with weight $1 - w$ to the other. Positive samples remain the same and the weight for each positive sample is 1. We call this dataset the *relabeled data*.

We train a SVM classifier with linear kernel on the relabeled data. This is our final detector of important/unimportant sentences.

3.2 Features

The classifiers for both stages use dictionary-derived features which indicate the types / properties of a word, along with several general features.

MRC The MRC Psycholinguistic Database (Wilson, 1988) is a collection of word lists with associated word attributes according to judgements by multiple people. The degree to which a word is associated with an attribute is given as a score within a range. We divide the score range into 230 intervals. The number of intervals was decided empirically on a small development set and was inspired by prior work of feature engineering

for real valued scores (Beigman Klebanov et al., 2013). Each interval corresponds to a feature; the value of the feature is the fraction of words in a sentence whose score belongs to this interval. Six attributes are selected: imagery, concreteness, familiarity, age-of-acquisition, and two meaningfulness attributes. In total, there are 1,380 MRC features.

LIWC LIWC is a dictionary that groups words in different categories, such as positive or negative emotions, self-reference etc. and other language dimensions relevant in the analysis of psychological states. Sentences are represented by a histogram of categories, indicating the percentage of words in the sentence associated with each category. We employ LIWC2007 English dictionary which contains 4,553 words with 64 categories.

INQUIRER The General Inquirer (Stone et al., 1962) is another dictionary of 7,444 words, grouped in 182 general semantic categories. For instance, the word *absurd* is mapped to tags NEG and VICE. Again, a sentence is represented with the histogram of categories occurring in the sentence.

General We also include features that capture general attributes of sentences including: *total number of tokens, number of punctuation marks, if it contains exclamation marks, if it contains question marks, if it contains colons, if it contains double quotations.*

4 Experiments on Importance Detection

We train a classifier separately for the DUC2002 and the NYT 1986-2006 corpora. The DUC model is trained using the articles and summaries from DUC2002 dataset, where 1,833 sentences in total appear in the summaries. We also randomly sample 2,200 non-summary sentences as unlabeled samples to balance the training set. According to the criteria described in NYT corpus section, there are 22,459 (14.1%) positive sentences selected from total of 158,892 sentences. Sentences with Jacana alignment scores less than or equal to 10 form the unlabeled set, including 20,653 (12.9%) unlabeled sentences in total. Liblinear (Fan et al., 2008) is used for training the two-stage classifiers.

4.1 Test Set

The test set consists of 1,000 sentences randomly selected from NYT dataset for the year 2007. Half

of the sentences are from the Business section, where the training data was drawn. The rest are from the U.S. International Relations section (*Politics* for short), to test the stability of prediction across topic domains. Three students from the University of Akron annotated if the test sentences contain important summary-worthy information.

For each test (source) sentence from the original article, we first apply Jacana to align it with every sentence in the corresponding summary. The summary sentence with the highest matching score is picked as the target sentence for the source sentence. Each pair of source and target sentences is presented to students and they are asked to mark if the sentences share information. Sentences from the original article that contribute content to the most similar summary sentence are marked as positive; those that do not are marked as negative. The pairwise annotator agreements are all above 80% and the pairwise Kappa ranges from 0.73 to 0.79.

The majority vote becomes the label of the source (article) sentence. Table 1 presents the distribution of final labels. The classes are almost balanced, with slightly more negative pairs overall.

Table 1: The distribution of the annotated labels

Section	Positive	Negative
Business	232 (46.4%)	268 (53.6%)
Politics	219 (43.8%)	281 (56.2%)
Total	451 (45.1%)	549 (54.9%)

4.2 Evaluation Results

In the process above, we have obtained a set of article sentences that contribute to the summary (positive class) or not (negative class)[1].

Table 2 shows the evaluation results on the human-annotated test set. The baseline is assuming that all sentences are summary-worthy. Although the unimportant class is the majority (see Table 1), predicting all test samples as not summary-worthy is less useful in real applications because we cannot output an empty text as a summary.

Each row in Table 2 corresponds to a model trained with one training set. We use dictionary features to build the models, i.e., NYT Model and DUC Model. We also evaluate the effectiveness of

---

[1] We assume that an article sentence not contributing to the summary does not contribute any content to the summary sentence that is closest to the article sentence.

the general features by excluding it from the dictionary features, i.e. NYT w/o general and DUC w/o general. Precision, recall and F-1 score are presented for all models. Models trained on the NYT corpus and DUC corpus are both significantly better than the baseline, with $p < 0.0001$ for McNemara's test. The NYT model is better than DUC model overall according to F-1. The results also show a noticeable performance drop when general features are removed.

We also trained classifiers with bag of words (BOW) features for NYT and DUC respectively, i.e. BOW-NYT and BOW-DUC. The classifiers trained on BOW features still outperform the baseline but are not as good as the dictionary and general sentence properties models.

Table 2: Evaluation results on human annotations

	Precision	Recall	F-1
NYT Model	0.582	0.846	0.689
DUC Model	0.541	0.903	0.676
NYT w/o General	0.547	0.847	0.664
DUC w/o General	0.508	0.906	0.651
BOW-NYT	0.520	0.852	0.645
BOW-DUC	0.501	0.828	0.623
Baseline	0.464	1.000	0.621

4.3 NYT Model vs. DUC Model

Further, we study the agreement between the two models in terms of prediction outcome. First, we compare the prediction outcome from the two models using NYT2007 test set. The Spearman's correlation coefficients between the outputs from the two models is around 0.90, showing that our model is very robust and independent of the training set.

Then we repeat the study on a much larger dataset, using articles from the DUC 2004 multi-document summarization task. There are no single document summaries in that year but this is not a problem, because we use the data simply to study the agreement between the two models, i.e., whether they predict the same summary-worthy status for sentences, not to measure the accuracy of prediction. There are 12,444 sentences in this dataset. The agreement between the two models is very high (87%) for both test sets. Consistent with the observation above, the DUC model is predicting intrinsic importance more aggressively. Only for a handful of sentences the NYT model predicts positive (important) while the DUC model predicts negative (not important).

We compute Spearman's correlation coefficients between the posterior probability for sentences from the two models. The correlation is around 0.90, indicating a great similarity in the predictions of the two models.

5 Summarization

We propose two importance-based approaches to improving single-document summarization.

In the first approach, **InfoRank**, the summary is constructed solely from the predictions of the sentence importance classifier. Given a document, we first apply the sentence importance detector on each sentence to get the probability of this sentence being intrinsically important. Then we rank the sentences by the probability score to form a summary within the required length.

The second approach, **InfoFilter**, uses the sentence importance detector as a pre-processing step. We first apply the sentence importance detector on each sentence, in the order they appear in the article. We keep only sentences predicted to be summary-worthy as the summary till the length restriction. This combines the preference for sentences that appear at the beginning of the article but filters out sentences that appear early but are not informative.

5.1 Results on Automatic Evaluation

The model trained on the NYT corpus is used in the experiments here. Business and politics articles (100 each) with human-generated summaries from NYT2007 are used for evaluation. Summaries generated by summarizers are restricted to 100 words. Summarizer performance is measured by ROUGE-1 (R-1) and ROUGE-2 (R-2) scores (Lin, 2004).

Several summarization systems are used for comparison here, including LeadWords, which picks the first 100 words as the summary; RandomRank, which ranks the sentences randomly and then picks the most highly ranked sentences to form a 100-word summary; and Icsisumm (Gillick et al., 2009), a state-of-the-art multi-document summarizer (Hong et al., 2014).

Table 4 shows the ROUGE scores for all summarizers. InfoRank significantly outperforms Icsisumm on R-1 score and is on par with it on R-2 score. Both InfoRank and Icsisumm outperform RandomRank by a large margin. These results show that the sentence importance detector

Table 3: Example of unimportant content in the opening paragraph of an article. The detected unimportant sentences are italicized. The third panel shows a new summary, with unimportant content skipped.

Human Summary: Pres Bush and his aides insist United States is committed to diplomatic path in efforts to stop Iran's suspected nuclear weapons program and support for terrorism, but effort is haunted by similar charges made against Iraq four years ago. Democrats see seizure of Iranians in Iraq and attempts to starve Iran of money to revitalize its oil industry as hallmarks of administration spoiling for fight. some analysts see attempt to divert attention from troubles in Iraq . administration insiders fear Bush's credibility has been deeply damaged. Bush's advisors debate how forcefully to push confrontation with Iran.
Lead paragraph: *This time, they insist, it is different.* As President Bush and his aides calibrate how directly to confront Iran, they are discovering that both their words and their strategy are haunted by the echoes of four years ago, when their warnings of terrorist activity and nuclear ambitions were clearly a prelude to war. *"We're not looking for a fight with Iran," R. Nicholas Burns, the under secretary of state for policy and the chief negotiator on Iranian issues, said in an interview, just a few hours after Mr. Bush had repeated his warnings to Iran to halt "killing our soldiers" ...*
New summary; unimportant sentences removed: As President Bush and his aides calibrate how directly to confront Iran, they are discovering that both their words and their strategy are haunted by the echoes of four years ago, when their warnings of terrorist activity and nuclear ambitions were clearly a prelude to war. Mr. Burns, citing the president's words, insisted that Washington was committed to "a diplomatic path", even as it executed a far more aggressive strategy, seizing Iranians in Iraq and attempting to starve Iran of the money it needs to revitalize a precious asset, its oil industry. Mr. Burns argues that those are defensive steps ...

is capable of identifying the summary-worthy sentences.

LeadWords is still a very strong baseline single-document summarizer. InfoFilter achieves the best result and greatly outperforms the LeadWords in both R-1 and R-2 scores. The p value of Wilcoxon signed-rank test is less than 0.001, indicating that the improvement is significant. Table 3 shows the example of lead paragraph along with the InfoFilter summary with the unimportant sentence removed.

Table 4: Performance comparison on single-document summarization (%)

System	R-1	R-2	System	R-1	R-2
InfoRank	**37.6**	15.9	**InfoFilter**	**50.7**	**30.2**
Icsisumm	33.3	**16.0**	LeadWords	48.0	27.5
RandomRank	31.9	8.7			

The InfoFilter summarizer is similar to the LeadWords summarizer, but it removes any sentence predicted to be unimportant and replaces it with the next sentence in the original article that is predicted to be summary-worthy. Among the 200 articles, 116 have at least one uninformative sentence removed. The most frequent number is two removed sentences. There are 17 articles for which more than three sentences are removed.

5.2 Results on Human Evaluation

We also carry out human evaluation, to better compare the relative performance of the LeadWords and InfoFilter summarizers. Judgements are made for each of the 116 articles in which at least one sentence had been filtered out by InfoFilter. For each article, we first let annotators read the summary from the NYT2007 dataset and then the two summaries generated by LeadWords and InfoFilter respectively. Then we ask annotators if one of the summary covers more of the information presented in the NYT2007 summary. The annotators are given the option to indicate that the two summaries are equally informative with respect to the content of the NYT summary. We randomize the order of sentences in both LeadWords and InfoFilter summaries when presenting to annotators.

The tasks are published on Amazon Mechanical Turk (AMT) and each summary pair is assigned to 8 annotators. The majority vote is used as the final label. According to human judgement, InfoFilter generates better summaries for 55 of the 116 inputs; for 39 inputs, the LeadWords summary is judged better. The result is consistent with the ROUGE scores, showing that InfoFilter is the better summarizer.

6 Conclusion

In this paper, we presented a detector for sentence importance and demonstrated that it is robust regardless of the training data. The importance detector greatly outperforms the baseline. Moreover, we tested the predictors on several datasets for summarization. In single-document summarization, the ability to identify unimportant content allows us to significantly outperform the strong lead baseline.

References

Beata Beigman Klebanov, Nitin Madnani, and Jill Burstein. 2013. Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics*, 1:99–110.

Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 481–490, Portland, Oregon, USA, June. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015. Ranking with recursive neural networks and its application to multi-document summarization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2153–2159, Austin, TX, USA. AAAI Press.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany, August. Association for Computational Linguistics.

Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 213–220, New York, NY, USA. ACM.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, June.

Dan Gillick, Benoit Favre, Dilek Hakkani-tr, Berndt Bohnet, Yang Liu, and Shasha Xie. 2009. The icsi/utd summarization system at tac 2009. In *Proceedings of the Second Text Analysis Conference (TAC 2009)*, November.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Alex Kulesza and Ben Taskar. 2011. Learning determinantal point processes. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 419–427.

Wee Sun Lee and Bing Liu. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 448–455, Washington, DC, USA. AAAI Press.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2), May.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing Management*, 43(6):1506–1520, November.

Philip J. Stone, Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498.

Krysta Svore, Lucy Vanderwende, and Christopher Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 448–457, Prague, Czech Republic, June. Association for Computational Linguistics.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Yinfei Yang and Ani Nenkova. 2014. Detecting information-dense texts in multiple news domains. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1650–1656, Quebec City, Quebec, Canada. AAAI Press.

Xuchen Yao, Benjamin Van Durme, Chris Callison-Burch, and Peter Clark. 2013. A lightweight and high performance monolingual word aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 702–707, Sofia, Bulgaria, August. Association for Computational Linguistics.

Liang Zhou and Eduard Hovy. 2003. A web-trained extraction summarization system. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 205–211, Stroudsburg, PA, USA. Association for Computational Linguistics.