УДК 621.391

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО ОБОБЩЕНИЯ ТЕКСТА

Е.К. ВАШКЕВИЧ, И.А. БОРИСКЕВИЧ

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь

Поступила в редакцию 8 марта 2021

Аннотация. Проведен сравнительный анализ различных алгоритмов извлечения ключевых предложений для автоматического реферирования текста на наборах текстовых данных новостных статей на английском языке. Рассмотрено тринадцать различных алгоритмов реферирования, а именно TextRank, LexRank, Luhn, LSA, Edmundson, ChunkRank, TGraph, UniRank, NN-ED, NN-SE, FE-SE, SummaRuNNer и MMR-SE, и произведена оценка их эффективности с использованием нескольких показателей производительности, таких как точность, отзыв, F_1 на пяти различных уровнях отсечения суммарной длины для разных *n*-грамм.

Ключевые слова: обобщение текста, алгоритмы на основе графиков, алгоритмы на основе нейронных сетей, алгоритмы на основе метаэвристики, скрытый семантический анализ, ROUGE.

Введение

Задача суммаризации текстов (автореферирование) – одна из ключевых, широко обсуждаемых задач NLP (Natural Language Processing – Обработка естественного языка). Она состоит в сжатии больших объемов текста до связного краткого содержания, отражающего только основные идеи.

Сейчас доступно огромное количество текстовой информации по любой теме. Чтобы сократить время на ознакомление с интересующей информацией используют алгоритмы суммаризации текстов. Их задача – выделить из потока текстовых данных главные идеи и создать на их основе сокращенный читаемый текст. Так, суммаризация может помочь понять содержание той или иной научной статьи, получить свежие выдержки из новостей или облегчить понимание юридического заключения или финансового отчета. Автореферирование актуально практически во всех областях, так как существенно сокращает время чтения.

Алгоритмы обобщения текста на основе графов, метаэвристики и нейронной сети

Алгоритмы LSA, NN-ED, SummaRuNNer и NN-SE – основаны на семантическом анализе [1–2]. Среди них LSA – единственная модель, не основанная на нейронной сети, в которой семантическое сходство между двумя векторами слов было вычислено путем реализации инструмента уменьшения размерности SVD, который используется для проецирования каждого слова в подпространстве заранее определенного размера. Другие подходы были обработаны с помощью вложений, которые используют статистические свойства текстовой структуры для встраивания слов в векторное пространство. Внедрение слова – это модель, основанная на прогнозировании, тогда как LSA – это модель на основе счетчиков. Подходы, основанные на встраивании, превосходят LSA из-за того, что они лучше разбираются в словах, чем LSA (например, слово "Obama" тесно связано со словом "president" при встраивании слов). Производительность встраивания постепенно меняется в зависимости от размера обучающих данных. Если размер обучающих данных невелик, то он не укладывается во все параметры встраивания и имеет тенденцию к ухудшению производительности. Результат работы алгоритма NN-SE содержит слишком много предложений, что часто приводит к включению незначительных предложений, так же включение несущественных предложений снижает производительность системы. Алгоритм SummaRuNNer учитывает четыре особенности текста

хорошего резюме. Новизна предложений может удалить значимые предложения и вызвать снижение производительности системы. Алгоритм NN-DE просто основан на подобии уровня документа с использованием встроенных распределений. Он генерирует резюме с более похожими предложениями и имеет тенденцию уменьшать заметность и новизну резюме.

TextRank, LexRank, ChunkRank, TGRaph и UniRank – это подходы, основанные на графах [3–5]. TextRank использует функцию сходства косинусов на основе перекрытия слов, чтобы найти сходство между двумя предложениями, в то время как LexRank использует функцию сходства косинусов на основе векторов TF-IDF, чтобы найти сходство между двумя предложениями. В LexRank подобие учитывает фактор значимости, что приводит к увеличению его производительности по сравнению с TextRank. ChunkRank создает чанки, используя описанные правила, которые вызывают чанки с разными темами. Он использует расстояние Левенштейна, чтобы найти сходство между двумя предложениями. Затем текст ранжируется на основе схожести предложений и веса предложения. TGRaph обрабатывает тематическое моделирование с использованием двудольного графа, а предложения ранжируются с использованием алгоритма HITS. Помимо этого, он также опирается на два основных фактора: согласованность и отсутствие избыточности в резюме, что делает его более подходящим. UniRank резюмирует текст, основываясь на локальной и глобальной значимости. Он использует матрицу аффинности косинусного сходства, чтобы найти взаимосвязь между предложениями. Концепции локальной и глобальной значимости превосходят другие алгоритмы.

MMR-SE [6] резюмирует документ на основе содержания темы и новизны, которые достигаются максимальной предельной релевантностью. Алгоритм извлекает предложения с минимальным сходством из предыдущего предложения, что снижает удобочитаемость и последовательность в резюме. Фактор релевантности решает проблему низкой согласованности, но низкая читаемость снижает ее производительность. Алгоритм Луна [7] просто основан на термине "частота". Он рассматривает наиболее часто встречающийся термин как наиболее значимый термин. Алгоритм Эдмундсона [8] основан исключительно на четырех характеристиках текста: ключевой фразе, ключевой фразе, заголовке и позиции. Вес текстовых функций помогает назначить приоритет одной функции над другими. Алгоритм не находит оптимальных весов текстовых функций, что снижает его эффективность. Последний алгоритм, FE-SE, представляет собой нечеткий эволюционный подход, который анализирует особенности текста на уровне слова и предложения. Признак сходства предложений оценивается с использованием метаэвристического подхода. Данный алгоритм также находит оптимальные веса текстовых функций с использованием гибридного метаэвристического подхода и извлекает предложения в соответствии с описанными функциями. Данные функции делают алгоритм FE-SE лучшим среди всех, но у него также есть некоторые недостатки, которые могут снизить производительность. Во-первых, метаэвристический подход не гарантирует получения оптимального решения. Во-вторых, производительность алгоритма FE-SE зависит от набора обучающих данных. Эти два недостатка снижают эффективность алгоритма.

Наборы данных для сравнительной оценки эффективности алгоритмов

Эффективность алгоритмов обобщения текста зачастую зависит от типа данных. В данной работе используются наборы данных, представляющие собой новости на английском языке. Каждый набор данных, используемый для экспериментов, содержит документы из различного типа новостей, таких как политика, спорт, криминал, награды, конференции и т.д.

Для сравнения из каждого набора данных случайным образом выбрано 500 документов примерно одинаковой длины (100 предложений в каждом документе). В табл. 1 содержится подробная информация о наборах данных.

Табл. 1. **Набор данных**

Набор данных	Размер	Язык	Тип	Количество документов	Ср. кол-во предложений в документе	Ср. длина предложения (слова)
FIRE data (2019)	397,8 MB	English	News article	500	100	22

Оценка эффективности алгоритмов автоматического обобщения текста на основе графов, метаэвристики и нейронной сети

Анализ результатов проводился на основе следующих методов: анализ на основе алгоритма обобщения, анализ по *n*-граммам, анализ на основе суммарной длины.

Табл. 2. **Оценка точности по ROUGE-1**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,8312	0,7734	0,7137	0,6599	0,5899
LexRank	0,9134	0,7639	0,6905	0,6498	0,5859
LSA	0,7543	0,7250	0,6897	0,6405	0,5876
Luhn	0,7254	0,7146	0,6803	0,6516	0,5901
Edmundson	0,9122	0,76621	0,6993	0,6428	0,5925
ChunkRank	0,7316	0,7190	0,6837	0,6513	0,5911
TGRAPH	0,9523	0,8678	0,7813	0,6956	0,6582
UniRank	0,9562	0,8697	0,7884	0,7003	0,6632
SummaRuNNer	0,9015	0,7851	0,7351	0,6195	0,5464
NN-ED	0,9112	0,7987	0,7615	0,6465	0,5591
FE-SE	0,9769	0,8938	0,8027	0,7419	0,6830
NN-SE	0,9187	0,8021	0,7843	0,6501	0,5637
MMR-SE	0,9219	0,8106	0,7414	0,6771	0,6234

В области лингвистической обработки *n*-грамма – это последовательность из n элементов любого текста. Оценка производилась до 4 граммов. При использовании 1-граммового метода (табл. 2) каждый член предложения сравнивается со справочной сводкой. В 2-граммовом методе (табл. 3) одновременно рассматриваются два термина, которые должны соответствовать справочной сводке. Если обнаруживается точная последовательность из двух терминов, то говорят, что они совпадают и показатели ROUGE-1 лучше, как показано в табл. 2, 6, 8. С увеличением *n*-значения ROUGE скорость перекрывающихся членов уменьшается. Таким образом, оценка постепенно снижается с ROUGE-1 до ROUGE-4.

Оценка точности: показатели точности постоянно снижаются с постепенным увеличением уровня отсечения. Из табл. 2–5 видно, что существуют лишь незначительные различия в производительности различных алгоритмов. Некоторые алгоритмы хорошо работают при малой краткой длине, другие – при большой краткой длине. В случае сравнения оценок точности LexRank и LSA при 10 % LexRank работает лучше, чем LSA, а при 50 % LSA работает лучше, чем LexRank. Это также указывает на то, что производительность системы зависит от суммарной длины.

Табл. 3. **Оценка точности по ROUGE-2**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,7361	0,6462	0,5936	0,5621	0,5120
LexRank	0,8911	0,6397	0,5587	0,5339	0,4819
LSA	0,6142	0,5809	0,5582	0,5304	0,4998
Luhn	0,5755	0,5585	0,5425	0,5396	0,5001
Edmundson	0,8919	0,6492	0,5689	0,5234	0,4881
ChunkRank	0,5963	0,5610	0,5467	0,5401	0,5014
TGRAPH	0,9122	0,8321	0,7455	0,6627	0,6192
UniRank	0,9173	0,8388	0,7504	0,6664	0,6220
SummaRuNNer	0,8662	0,7442	0,6665	0,5794	0,5053
NN-ED	0,8710	0,7668	0,6741	0,6035	0,5180
FE-SE	0,9564	0,8791	0,7914	0,7046	0,6690
NN-SE	0,8791	0,7680	0,6775	0,6119	0,5185
MMR-SE	0,9035	0,7514	0,6931	0,6047	0,5737

Табл. 4. **Оценка точности по ROUGE-3**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,7076	0,6228	0,5750	0,5481	0,5021
LexRank	0,8787	0,6217	0,5402	0,5176	0,4678
LSA	0,5881	0,5568	0,5375	0,5120	0,4856
Luhn	0,5372	0,5324	0,5224	0,5226	0,4873
Edmundson	0,8793	0,6287	0,5508	0,5072	0,4738
ChunkRank	0,5468	0,5433	0,5281	0,5230	0,4888
TGRAPH	0,8906	0,7913	0,7220	0,6315	0,5926
UniRank	0,8961	0,7959	0,7272	0,6338	0,6013
SummaRuNNer	0,8460	0,7089	0,6452	0,5417	0,4782
NN-ED	0,8472	0,7355	0,6401	0,5884	0,4812
FE-SE	0,9445	0,8386	0,7625	0,6893	0,6438
NN-SE	0,8486	0,7367	0,6450	0,5890	0,4942
MMR-SE	0,8912	0,7714	0,6763	0,5914	0,5222

Табл. 5. **Оценка точности по ROUGE-4**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,7002	0,6179	0,5717	0,5461	0,5008
LexRank	0,8753	0,6187	0,5385	0,5167	0,4667
LSA	0,5839	0,5521	0,5321	0,5077	0,4821
Luhn	0,5260	0,5269	0,5179	0,5193	0,4854
Edmundson	0,8785	0,6255	0,5485	0,5064	0,4727
ChunkRank	0,5452	0,5374	0,5241	0,5231	0,4862
TGRAPH	0,8891	0,7881	0,7013	0,6253	0,5915
UniRank	0,8943	0,7915	0,7086	0,6321	0,5958
SummaRuNNer	0,8432	0,7043	0,6445	0,5408	0,4771
NN-ED	0,8458	0,7324	0,6319	0,5848	0,4806
FE-SE	0,9381	0,8227	0,7339	0,6738	0,6260
NN-SE	0,8461	0,7357	0,6431	0,5870	0,4931
MMR-SE	0,8831	0,7236	0,6416	0,5723	0,5241

Оценка отзыва: оценка отзыва постепенно увеличивается по мере увеличения порогового уровня. Также скорость извлечения релевантных терминов уменьшается по мере увеличения суммарной длины. Из табл. 6–7 видно, что, как и при оценке точности, некоторые алгоритмы лучше работают при малой суммарной длине, а другие – при большой суммарной длине. Во всех этих случаях показана зависимость от суммарной длины.

Табл. 6. **Оценка отзыва по ROUGE-1**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,1979	0,4100	0,5603	0,6803	0,7608
LexRank	0,1974	0,3578	0,4800	0,5914	0,6845
LSA	0,1980	0,4068	0,5538	0,6797	0,7706
Luhn	0,1569	0,3436	0,4919	0,6463	0,7409
Edmundson	0,1854	0,3512	0,4770	0,5919	0,6815
ChunkRank	0,1824	0,3717	0,5143	0,6567	0,7524
TGRAPH	0,2434	0,4768	0,5917	0,6943	0,7995
UniRank	0,2461	0,4782	0,5952	0,7011	0,8030
SummaRuNNer	0,1962	0,4079	0,5481	0,6147	0,7212
NN-ED	0,2019	0,4124	0,5490	0,6329	0,7381
FE-SE	0,27672	0,5073	0,6262	0,7470	0,8321
NN-SE	0,2118	0,4315	0,5543	0,6580	0,7441
MMR-SE	0,2119	0,4135	0,5412	0,6633	0,7218

Табл. 7. **Оценка отзыва по ROUGE-4**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,1446	0,2946	0,4079	0,5097	0,5844
LexRank	0,1676	0,2552	0,3306	0,4168	0,4838
LSA	0,1405	0,2862	0,3930	0,4917	0,5763
Luhn	0,1048	0,2275	0,3365	0,465982	0,5520
Edmundson	0,1585	0,2523	0,3291	0,4118	0,4809
ChunkRank	0,1227	0,2558	0,3636	0,4741	0,5631
TGRAPH	0,1949	0,3019	0,4258	0,5479	0,6211
UniRank	0,1972	0,3042	0,4288	0,5494	0,6240
SummaRuNNer	0,1638	0,2730	0,3926	0,5059	0,5726
NN-ED	0,1720	0,2821	0,4017	0,5102	0,5801
FE-SE	0,2367	0,3382	0,4562	0,5672	0,6321
NN-SE	0,1721	0,2910	0,4018	0,5153	0,5910
MMR-SE	0,1764	0,2819	0,3761	0,4928	0,5712

Табл. 8. **Оценка F_1 по ROUGE-1**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,3134	0,5299	0,6215	0,6657	0,6613
LexRank	0,3185	0,4817	0,5612	0,6157	0,6286
LSA	0,3076	0,5145	0,6068	0,6545	0,6617
Luhn	0,2532	0,4595	0,5656	0,6433	0,6525
Edmundson	0,3042	0,4756	0,5622	0,6134	0,6317
ChunkRank	0,2919	0,4901	0,5870	0,6539	0,6621
TGRAPH	0,3877	0,6154	0,6735	0,6949	0,7220
UniRank	0,3914	0,6171	0,6783	0,7007	0,7264
SummaRuNNer	0,3223	0,5369	0,6279	0,6170	0,6217
NN-ED	0,3306	0,5439	0,6380	0,6396	0,6362
FE-SE	0,4312	0,6472	0,7035	0,7444	0,7502
NN-SE	0,3442	0,5611	0,6495	0,6540	0,6415
MMR-SE	0,3446	0,5476	0,6257	0,6701	0,6690

Табл. 9. **Оценка F_1 по ROUGE-4**

Алгоритмы	Длина резюме, %				
	10	20	30	40	50
TextRank	0,2353	0,3948	0,4713	0,5239	0,5365
LexRank	0,2763	0,3569	0,4059	0,4588	0,4728
LSA	0,2224	0,3721	0,4463	0,4955	0,5207
Luhn	0,1719	0,3152	0,4038	0,4866	0,5128
Edmundson	0,2652	0,3546	0,4073	0,4521	0,4750
ChunkRank	0,2003	0,3466	0,4293	0,4974	0,5218
TGRAPH	0,3197	0,4365	0,5298	0,5840	0,6059
UniRank	0,3231	0,4394	0,5343	0,5878	0,6095
SummaRuNNer	0,2743	0,3935	0,4879	0,5227	0,5205
NN-ED	0,2859	0,4073	0,4911	0,5449	0,5256
FE-SE	0,3780	0,4793	0,5626	0,6159	0,6290
NN-SE	0,2860	0,4170	0,4945	0,5488	0,5376
MMR-SE	0,2941	0,4057	0,4742	0,5296	0,5466

Оценка английского языка F_1: из оценки английского языка F_1 в табл. 8–9 видно, что производительность некоторых алгоритмов, таких как TextRank, SummaRuNNer, NN-ED и NN-SE, снижается на 50 %, тогда как производительность других алгоритмов плавно увеличивается по мере увеличения суммарной длины.

Заключение

Исследованы тринадцать алгоритмов автоматического реферирования с аналогичными настройками для новостных наборов данных на английском языке. Произведена оценка производительности с использованием показателей точности, отзыва и показателей F_1 на пяти различных уровнях отсечения суммарной длины для разных *n*-грамм. Обнаружено, что показатель точности уменьшается с увеличением длины сводки, а также с увеличением значений слова *n*-грамм. Оценки запоминания увеличиваются с увеличением длины сводки, но уменьшаются по сравнению с *n*-значениями в *n*-граммах слов, максимальные значения наблюдались при длине сводки 40 %. Ограничение данных алгоритмов обусловлено тем, что их эффективность зависит от суммарной длины. Непосредственная близость значений F_1 для алгоритмов SummaRuNNer, NN-SE и NN-ED обусловлена тем, что они являются нейронными сетями, основанными на инструменте word2vec. При этом алгоритмы на основе нейронных сетей не показали лучшей производительности по сравнению с алгоритмами на основе графов. Также было показано, что почти все алгоритмы генерируют неизбыточные, удобочитаемые, значимые резюме.

COMPARATIVE ANALYSIS OF TEXT SUMMARIZATION ALGORITHMS IN ENGLISH LANGUAGE

E.K. VASHKEVICH, I.A. BARYSKIEVIC

Abstract. A detailed comparative study of various extraction algorithms for automatic text summarization on text data sets of news articles in English was carried out. Thirteen different summarization algorithms were considered, namely TextRank, LexRank, Luhn, LSA, Edmundson, ChunkRank, TGraph, UniRank, NN-ED, NN-SE, FE-SE, SummaRuNNer and MMR-SE, and their effectiveness was assessed using several performance metrics such as Accuracy, Recall, F_1 at five different levels of total length cutoff for different *n*-grams.

Keywords: text summarization, graph based techniques, neural networks based techniques, meta-heuristic based techniques, latent semantic analysis, ROUGE

Список литературы

1. Hayato Kobayashi, Masaki Noguchi, Taichi Yatsuka. // Summarization Based on Embedding Distributions, 2015. In EMNLP. 1984–1989.
2. Ramesh Nallapati, Feifei Zhai, Bowen Zhou. // SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents, 2017
3. Rada Mihalcea, Paul Tarau. // TextRank: Bringing order into texts. Association for Computational Linguistics, 2004.
4. Günes Erkan, Dragomir R Radev. // LexRank: Graph-based lexical centrality as salience in text summarization. Journal of Artificial Intelligence Research, 2004, P. 457–479.
5. Daraksha Parveen, Mohsen Mesgar, Michael Strube. // Generating Coherent Summaries of Scientific Articles Using Coherence Patterns, 2016. In EMNLP. P. 772–783.
6. Jade Goldstein, Jaime Carbonell. // Summarization: using MMR for diversity-based reranking and evaluating summaries. In Proceedings of a workshop on held at Baltimore, Maryland: October 13–15, 1998. Association for Computational Linguistics, P. 181–195.
7. Hans Peter Luhn. // The automatic creation of literature abstracts. IBM Journal of research and development, 1958, P. 159–165.
8. Harold P Edmundson. // New methods in automatic extracting. Journal of the ACM (JACM) 16, 1969, P. 264–285.