CNewSum: A Large-scale Chinese News Summarization Dataset with Human-annotated Adequacy and Deducibility Level

Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, Lei Li†

ByteDance AI Lab

†University of California Santa Barbara

ByteDance AI Lab

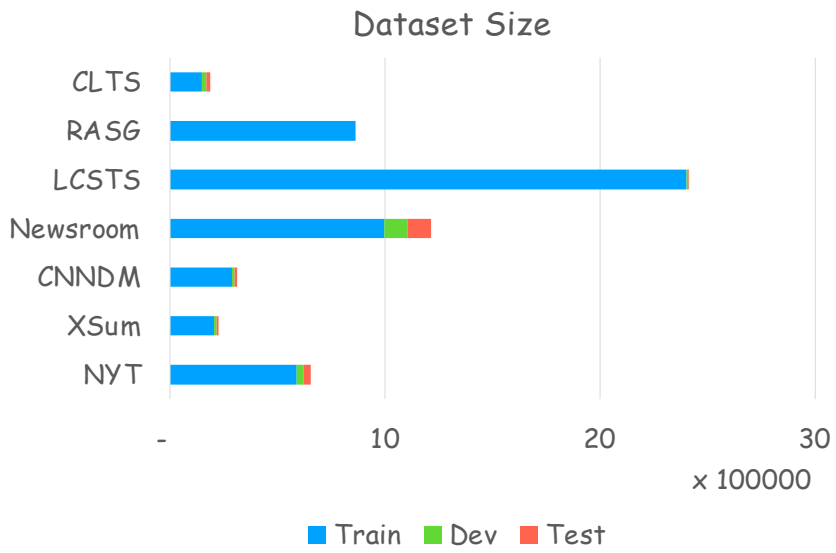
字节跳动人工智能实验室

Current Text Summarization: English >>> Chinese

Goal: A brief but crucial text for the long document

- input: a single long document / cluster of documents
- output: the summary

Dataset	Language	Source	Year	Citation
CNN/DailyMail	English	News	2015	1087
NYT / NYT50	English	News	2008	773 / 109
Newsroom	English	News	2018	130
XSum	English	News	2018	143
arXiv & PubMed	English	Academic paper	2018	121
...	...	...	...	...
LCSTS	Chinese	Social media	2015	168
RASG	Chinese	Social media	2019	16
CLTS	Chinese	News	2020	1



LCSTS: quite short

Article	本文总结了四个可穿戴产品的设计原则，而这些原则，同样也是笔者认为是这个行业最吸引人的地方：1.为人们解决重复性问题；2.从人开始，而不是从机器开始；3.要引起注意，但不要刻意；4.提升用户能力，而不是取代人。 This article summarizes ten design principles of wearable products, and these principles are also what I believe are the most attractive in this industry: 1. Solve repetitive problems for people; 2. Start with people, not machines; 3. Pay attention, but don't deliberately; 4. Improve user capabilities, not replace people.
Summary	可穿戴技术四大设计原则。 Ten design principles of wearable technology.

RASG: little information in the article

Article	5月7日，一毕业班55人，仅1名男生。同班女生说他是暖男，只要有力气活的地方他都会出现，一个可顶十个。可惜的是，男生至今未脱单。 On May 7, there were 55 students in the first graduating class, with only one boy. The girl in the same class said he was a warm guy, and he would show up wherever he had the strength to work, and one could be up to ten. It is a pity that the boys have not left the singles so far.
Summary	全班55人仅1名男生，被称为暖男却至今单身. There is only one boy in the class of 55. He is known as a warm man but is still single.
Comments	[0]其实主要还是颜值。[1]估计太娘了。[2]看了他的牙我得到了答案。[3]为什么感觉还是和颜值什么的有关。[4]哪都有他，这单身理由还不够。[5]主要看颜值。[6]还是因为现在是看脸的时代，09年学会计的时候班上五十个人，3个男生，一直单身。[7]我们全班五十六个人只有一个男生。[8]讨厌，大家都是姐妹。 [0] In fact, the main thing is the appearance. [1] I think it's too damn good. [2] I got the answer after looking at his teeth. [3] Why the feeling is still related to appearance. [4] He is everywhere, this reason for being single is not enough. [5] Mainly depends on the appearance. [6] It's because now is the age of looking at faces. When I was studying accounting in 2009, there were 50 people in my class, 3 boys, and they were single all the time. [7] There is only one boy in our class of fifty-six members. [8] Everyone is sister.

CLTS: extract from the article directly

Article	编者按：如果你“不想睡”或者“睡不着”，欢迎继续阅读。这里或许有个文艺片，这里或许有个惊悚片。距离1967年《雌雄大盗》（BonnieandClyde）过去已经超过40年，..... 显然为了能够对标《雌雄大盗》，《辣手骑警》在演员方面找来了两位好莱坞知名的硬汉凯文·科斯特纳和伍迪·哈里森.....从电影风格的角度来看，讲述同一事件的《雌雄大盗》与《辣手骑警》可谓是一体两面，前者浪漫又疯狂，后者则平淡和苍凉。.....如果不是《辣手骑警》这部电影，大概鲜有人会知道让“雌雄大盗”丧命的竟然是两位德州退休老头。 Editor‘s note: If you “don’t want to sleep” or “can‘t sleep”, welcome to continue reading. There may be a literary film here, and there may be a thriller here. It has been more than 40 years since "Bonnie and Clyde" (Bonnie and Clyde) in 1967,... Obviously, in order to be able to compete with "Bonnie and Clyde", "Hot Mounted Police" has recruited two famous Hollywood tough guys Kevin Coster in terms of actors. Na and Woody Harrison...From the point of view of movie style, "The Male and Female" and "Hot Horseman" about the same incident can be described as two sides, the former is romantic and crazy, the latter is plain and desolate. ...If it weren't for the movie "Hot Hand Mounted Police", probably few people would have known that it was the two retired Texas old men who killed the "male and female robbers".
Summary	从电影风格的角度来看，讲述同一事件的《雌雄大盗》与《辣手骑警》可谓是一体两面，前者浪漫又疯狂，后者则平淡和苍凉。 From the perspective of movie styles, "The Male and Female Thief" and "Hot Horseman" about the same incident can be described as two sides. The former is romantic and crazy, and the latter is plain and desolate.

CNewSum

- Large-scale and plenty of publishers (275.6k/14.4k/14.4k)
- Long article and human-written summary (730.4/35.1)
- High abstraction
- Adequacy & Deducibility

CNewSum : Large-scale and plenty of publishers

- Toutiao.com

- Hundreds of thousands publishers: thepaper, wallstreetcn, cankaoxiaoxi, yicai …

“看空”中国经济者何以频频“落空”

新华社 6451评论 2天前

解决急难愁盼 不怕鸡毛蒜皮

光明日报 0评论 13小时前

习近平：中国将构建起碳达峰、碳中和“1+N”政策体系

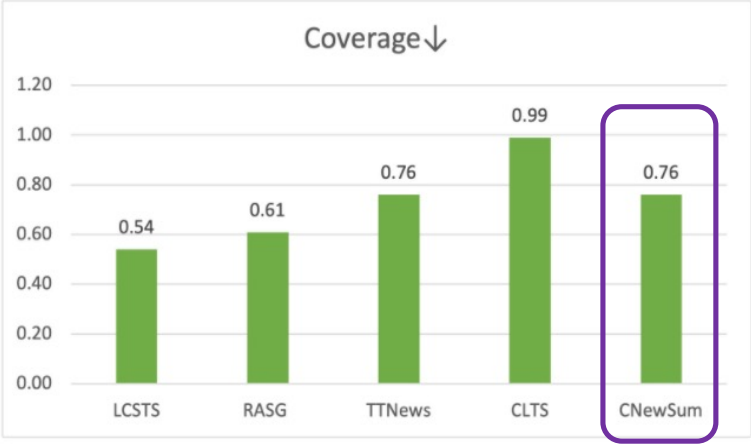
央广网 0评论 1小时前

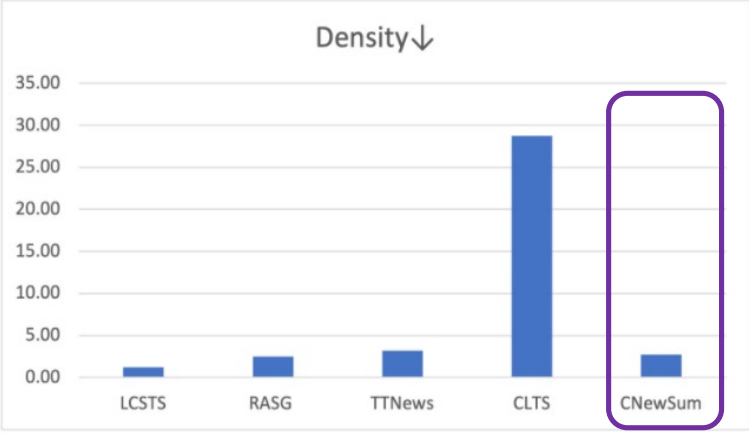
Dataet	Source
NYT	New York Times
CNNDM	CNN & Daily Mail
Newsroom	38 publishers
LCSTS	Weibo
RASG	Weibo
TTNews	Toutiao
CLTS	ThePaper

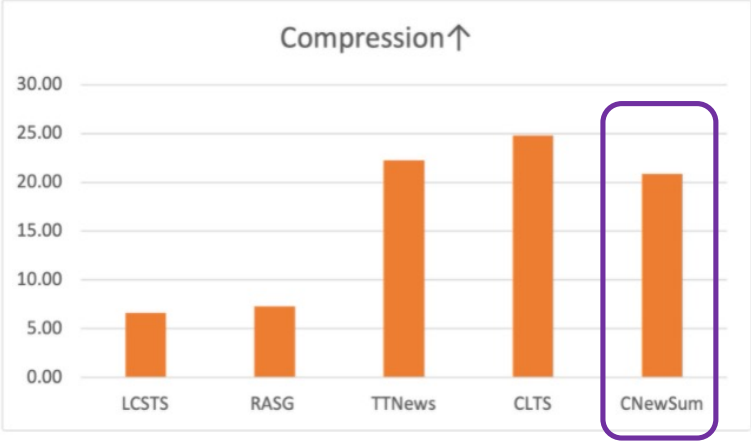
CNewSum: Long article and human-written summary

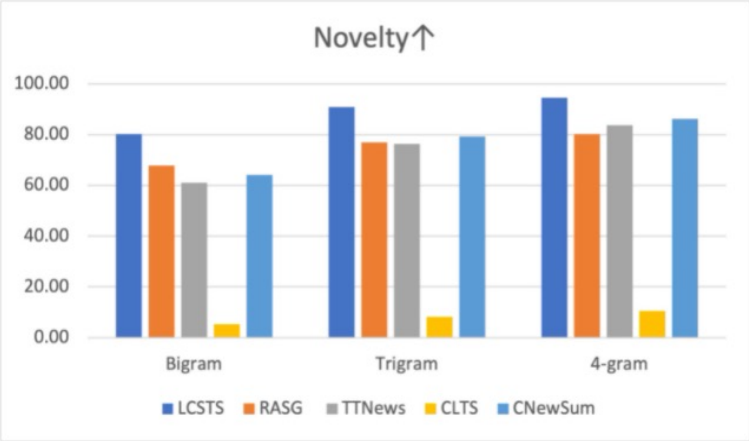
Article	[0]图为在广元市朝天区发现的海南虎斑鹀。[1]广林局供。[2]中新网广元3月15日电。[3]记者15日从四川省广元市野生动物救治中心获悉：近日，该市朝天区东溪河乡的群众发现一只受伤的“怪鸟”引起多方关注，随后，上报到广元市林业部门。[4]后经当地野生动物保护专家鉴定“怪鸟”为世界最濒危鸟类海南虎斑鹀。.....[7]据广元市野生动物救治中心工作人员介绍，经四川省和广元市野生动物保护专家鉴定，该鸟是我国特有的珍稀鸟类、国家二级保护动物海南虎斑鹀，被列为世界最濒危的30种鸟类之一，目前全世界仅存1000余只。 .....[12]此后一直再没有关于该鸟踪迹的报道。 [0] The picture shows the Hainan tiger bitter found in Chaotian District, Guangyuan City. [1] Provided by Guanglin Bureau. [2] China News Service Guangyuan, March 15th. [3] The reporter was appraised by a local wildlife protection expert from the wild animals in Guangyuan City, Sichuan Province on the 15th [4]: .....[7] According to the staff of Yuan City Wildlife Rescue Center, it was approved by a wildlife protection expert in Guangyuan City, Sichuan Province It is identified that the bird is a unique rare bird. , The national protected animal Hainan tiger clam was sent home to the 30 most dangerous bird species in the world, and there are more than 1,000 wine spots in stock at the moment. ...[12] Since then, there has been no report about the bird's trace.
Summary	广元一市民发现受伤“怪鸟”，经鉴定系世界濒危鸟类海南虎斑鹀，全球仅存1000只。 A citizen of Guangyuan found an injured "strange bird", which was identified as the world's endangered bird, the Hainan tabby bitter, and only 1,000 left in the world.

High abstraction









CNewSum

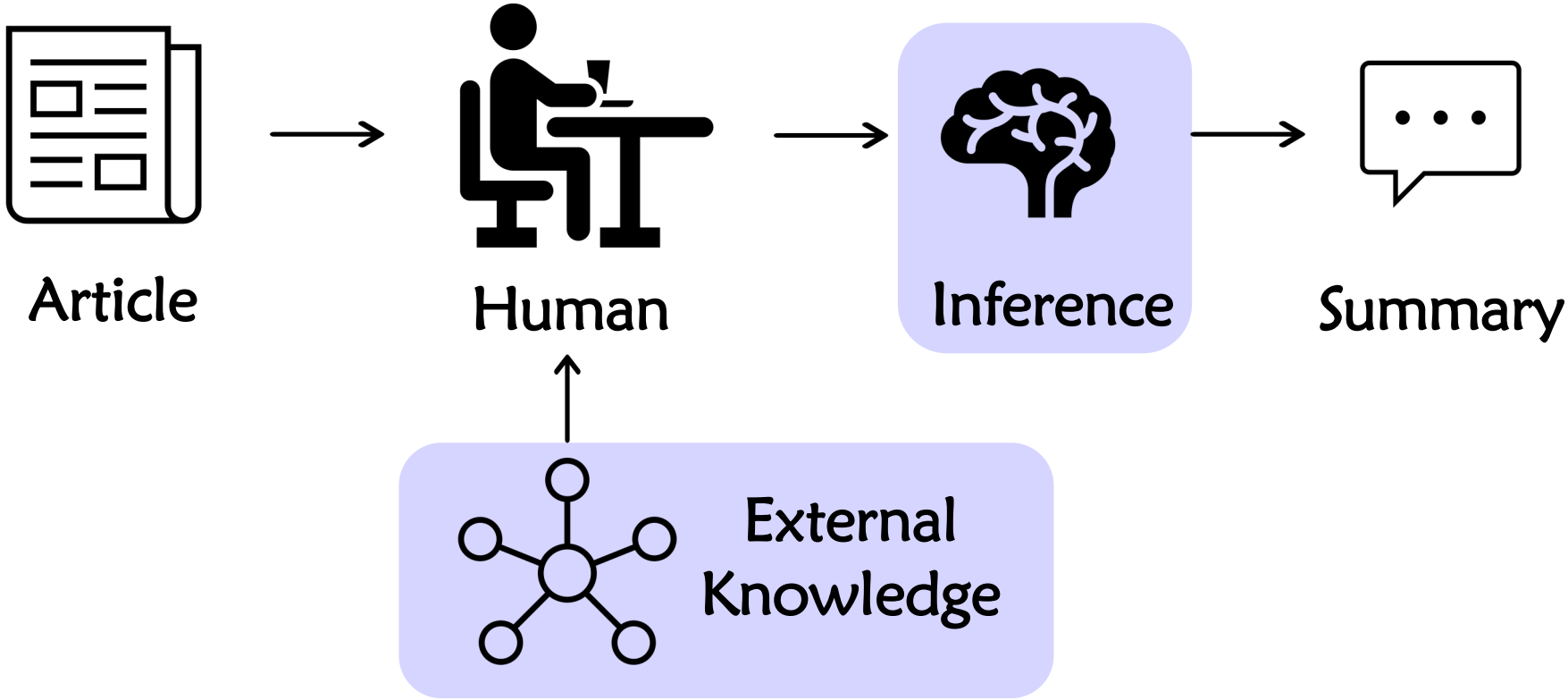
- Large-scale and plenty of publishers
- Long article and human-written summary
- High abstraction

Common characteristics

- Adequacy & Deducibility

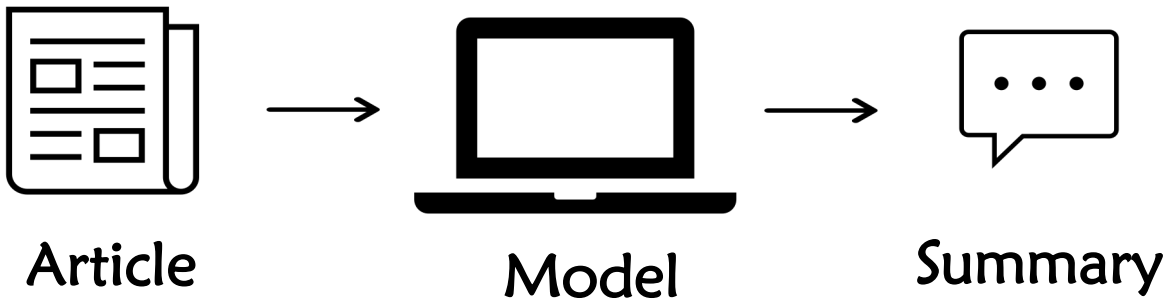
New features

How do human summarize an article?



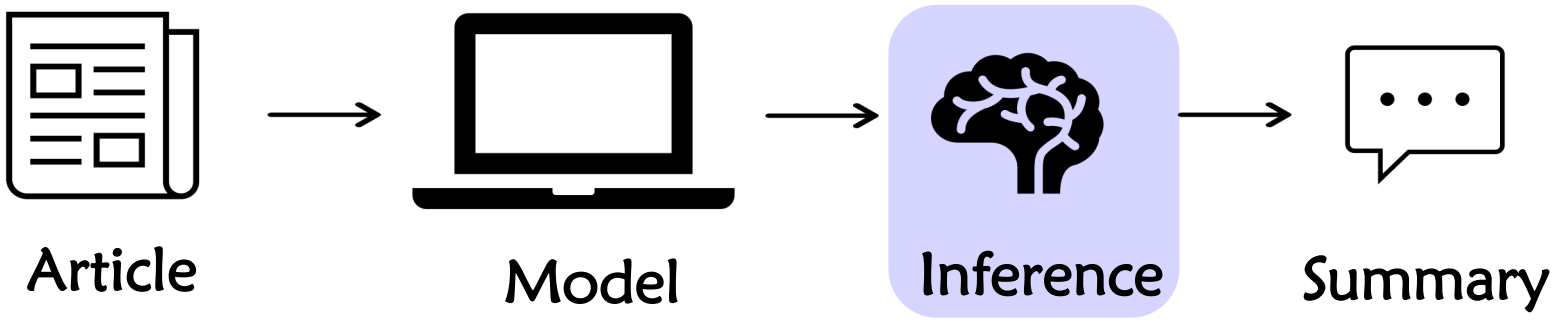
Adequacy

Has necessary information of the summary been included in the document?



Deducibility

Can the information of the summary be easily inferred from the document?



Adequacy & Deducibility

A& D	Article	Summary
A=1 & D=1 (91.08%)	北京时间1月6日,在小牛双加时117-116战胜国王的比赛中,德克-诺维茨基首节在无人防守的情况下扣篮不中,赛后,德克在推特上笑称篮筐有点高。德克转发了自己扣篮不中的视频,并写道:“Yooo。那个篮筐好像有12英尺高……”	诺天王比赛中上演一幕:最简单扣篮扣丢,自己调侃因篮筐太高。
A=0 & D=1 (4.11%)		
A=1 & D=0	-	-
A=0 & D=0 (4.81%)		

Adequacy & Deducibility

A& D	Article	Summary
A=1 & D=1 (91.08%)	北京时间1月6日,在小牛双加时117-116战胜国王的比赛中,德克-诺维茨基首节在无人防守的情况下扣篮不中,赛后,德克在推特上笑称篮筐有点高。德克转发了自己扣篮不中的视频,并写道:“Yooo。那个篮筐好像有12英尺高……”	诺天王比赛中上演一幕:最简单扣篮扣丢,自己调侃因篮筐太高。
A=0 & D=1 (4.11%)	王大陆在发布会现场王大陆、柴智屏,资料图1月6日消息,据台湾媒体报道,王大陆和柴智屏合约纠纷闹了1个多月未解,王大陆今下午2点召开记者会,说明与柴智屏间的官司,他透过律师指“柴智屏开的条件我无法负担。”……律师称王大陆在柴公司3年,赚不到100万台币,现在解约,柴却要王大陆付清3千万台币,未来4年照合约抽成,条件严苛。……因此公司全然交由律师处理。	王大陆开发布会:无法负担柴智屏的条件;律师称其3年赚不到25万元,解约要付600万。 [Exchange Rates] 100万台币 => 25万元 1 million NTD => 0.25 million RMB
A=1 & D=0	-	-
A=0 & D=0 (4.81%)		

Adequacy & Deducibility

A& D	Article	Summary
A=1 & D=1 (91.08%)	北京时间1月6日,在小牛双加时117-116战胜国王的比赛中,德克-诺维茨基首节在无人防守的情况下扣篮不中,赛后,德克在推特上笑称篮筐有点高。德克转发了自己扣篮不中的视频,并写道:“Yooo。那个篮筐好像有12英尺高……”	诺天王比赛中上演一幕:最简单扣篮扣丢,自己调侃因篮筐太高。
A=0 & D=1 (4.11%)	王大陆在发布会现场王大陆、柴智屏,资料图1月6日消息,据台湾媒体报道,王大陆和柴智屏合约纠纷闹了1个多月未解,王大陆今下午2点召开记者会,说明与柴智屏间的官司,他透过律师指“柴智屏开的条件我无法负担。”……律师称王大陆在柴公司3年,赚不到100万台币,现在解约,柴却要王大陆付清3千万台币,未来4年照合约抽成,条件严苛。……因此公司全然交由律师处理。	王大陆开发布会:无法负担柴智屏的条件;律师称其3年赚不到25万元,解约要付600万。
A=1 & D=0	-	-
A=0 & D=0 (4.81%)	杨云微博截图中新网2月1日电。1日早上,杨威妻子杨云在微博晒出爱子生病的照片,还留言称:“上吐下泻,快快好起来吧孩子,妈妈心疼。”照片中,杨威儿子戴着大大的眼罩,嘴巴微微向下撇,脸色惨白。对此,网友们纷纷送上关心与祝福,“好心疼我的小男神啊,快点好起来吧”、“懂事又可爱的杨阳洋,希望你能够快快好起来”、“快好快好我要你陪我玩。”	心疼!杨阳洋上吐下泻脸色惨白,戴大眼罩,嘴巴微微向下撇,身裹羽绒服卷缩角落。 [Unknown Entity] 杨云儿子——杨阳洋 Son of Yun Yang —— Yangyang Yang

More inference cases

- Unit conversion
  - ✓ 4 kg -> 4000 g
- Number Calculation
  - ✓ 300+1500 -> 1800
- Name Abbreviation
  - ✓ 武汉大学 -> 武大

CNewSum

- Large-scale and plenty of publishers (75.6k/14.4k/14.4k)
- Long article and human-written summary (730.4/35.1)
- High abstraction

Common characteristics

- Adequacy & Deducibility

New features

Evaluation Tool

- Chinese text is split by characters
- English words and numbers will be split by space
- All tokens are mapped to IDs

[Input] Surface Phone将装载Windows 10
[Output] Surface/phone/将/装/载/windows/10

Baseline Performance: Abstractive >> Extractive

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead	30.43	17.26	25.33
Oracle	46.84	30.54	40.08
TextRank	24.04	13.07	20.08
NeuSum	30.61	17.36	25.66
Transformer-ext	32.87	18.85	27.59
BERT-ext	34.78	20.33	29.34
Pointer Generator	25.70	11.05	19.62
Transformer-abs	37.36	18.62	30.62
BERT-abs	44.18	27.37	38.32

Wait for improvement: A=0

Model	Category	ROUGE-1	ROUGE-2	ROUGE-L
Transformer-ext	A=1&D=1	33.16	19.19	27.88
	A=0&D=1	30.89	15.60	25.38
	A=0&D=0	28.92	14.88	23.74
Transformer-abs	A=1&D=1	37.54	18.85	30.83
	A=0&D=1	36.36	16.70	29.63
	A=0&D=0	34.73	15.95	27.52
BERT-ext	A=1&D=1	35.05	20.67	29.62
	A=0&D=1	32.81	16.90	27.05
	A=0&D=0	31.07	16.57	25.72
Bert-abs	A=1&D=1	44.51	27.76	38.70
	A=0&D=1	41.75	23.64	35.34
	A=0&D=0	40.18	23.34	33.60

Thanks for your listening !

A Large-scale Chinese News Summarization Dataset with Human-annotated Adequacy and Deducibility Level

Danqing Wang, Jiaze Chen, Xianze Wu, Hao Zhou, Lei Li†
ByteDance AI Lab, †University of California Santa Barbara

Paper: PDF
Dataset: Google Drive
Evaluate tool: MLROUGE (Script for NLPCC2017 and NLPCC2018)

Overview

CNewSum is a large-scale Chinese news summarization dataset, which consists of 304,307 documents and human-written summaries from Toutiao. It is a extended version of **TTNews** for NLPCC2017 and NLPCC2018, which is much larger and has several features:

- The news article are collected from hundreds of thousands of news publishers. A team of expert editors are hired to provide human-written summaries for the daily news feed.
- Human-annotated labels are provided for each example in the test set to figure out how much knowledge the model needs to generate a human- like summary.
  - Adequacy Level: *Does necessary information of the summary has been included in the document?*
  - Deducibility Level: *Can the information of the summary be easily inferred from the document?*

Dataset Information

We list the statistics of common English and Chinese summarization datasets. The 'Article' and 'Summary' are the average length of articles and summaries in the dataset. For English, it is calculated by words and for Chinese, it is calculated by characters.

Welcome to take a close look at our dataset and challenge yourself!