

A Brief Introduction to Cross-lingual Summarization

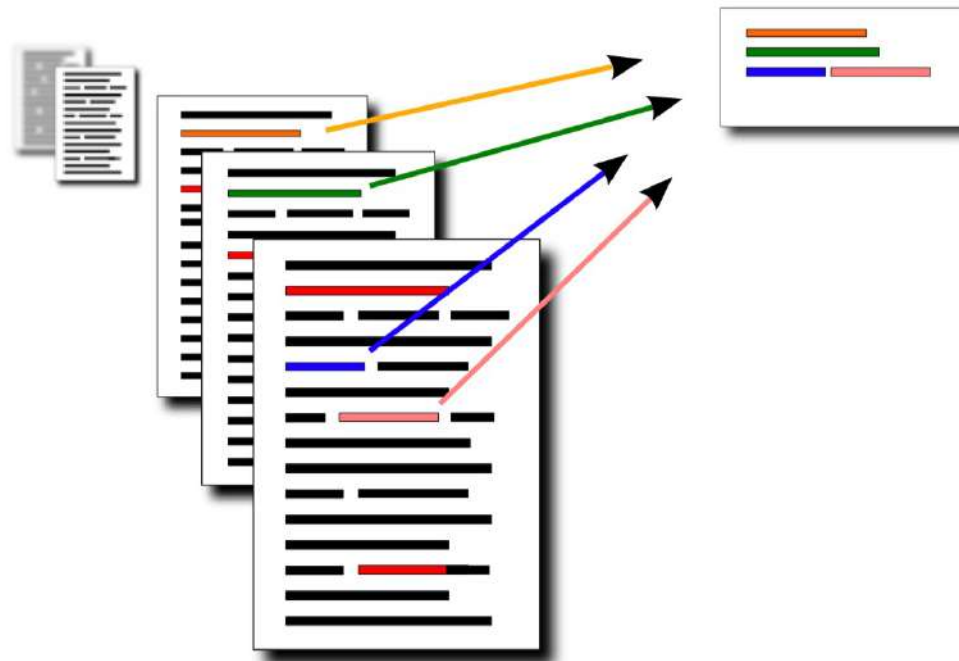
Xiachong Feng

2021-12-03

Introduction

Text Summarization

- Condense the input document into a concise version.



Task Classification



Single-Doc



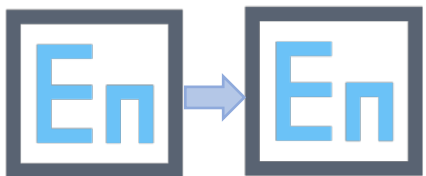
Multi-Doc



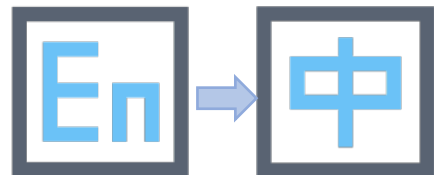
Single-modal



Multi-modal



Mono-lingual



Cross-lingual



News



Patent



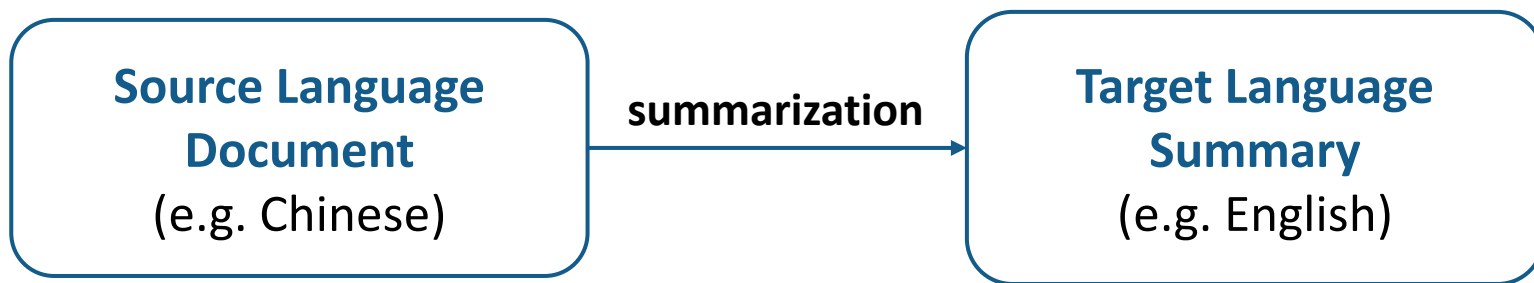
paper



Dialogue

Cross-lingual Summarization

- Cross-lingual summarization aims at summarizing a document in one language (e.g., Chinese) into another language (e.g., English).



TMZ体育记者称，詹姆斯今天接受了三次新冠检测，第一次检测呈阳性，随后他进行了第二次检测，但结果是阴性，在第三次决定性的检测中，詹姆斯的检测结果依然是阳性。消息源透露，詹姆斯无症状，球队已经为他定了私人飞机，飞回洛杉矶。

LeBron James took 3 COVID tests, Test 1 and 3 were positive.

Datasets

Zh2EnSum && En2ZhSum

- Original **English** Mono-lingual datasets: MSMO + CNNDM



- Original **Chinese** Mono-lingual datasets: LCSTS

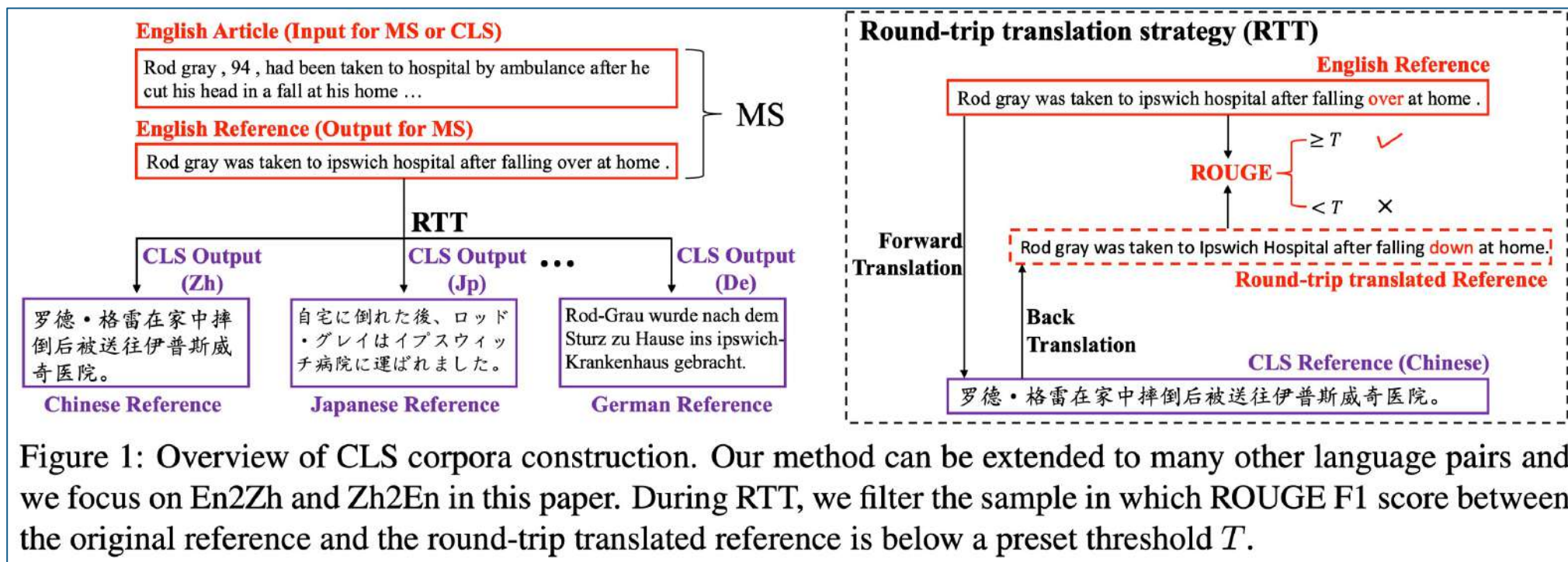
【江西高考被曝替考 有关考生已被警方控制】人民日报记者吴齐强消息，江西高考被曝光替考，7日中午江西省教育厅发布消息称，接到有人组织替考的举报后，江西省教育厅、江西省教育考试院立即部署南昌市教育考试院，联合南昌市警方开展调查核实，有关考生已被警方控制。调查进展情况将及时向社会公布。

LCSTS

MSMO: Multimodal Summarization with Multimodal Output
LCSTS: A Large Scale Chinese Short Text Summarization Dataset

Zh2EnSum && En2ZhSum

- Round-trip Translation



Zh2EnSum && En2ZhSum

- Original **English** : MSMO + CNNDM
- Original **Chinese**: LCSTS
- Quality Control:
 - Round-trip Translation

En2ZhSum	train	valid	test	Zh2EnSum	train	valid	test
#Documents	364,687	3,000	3,000	#Documents	1,693,713	3,000	3,000
#AvgWords (S)	755.09	759.55	744.84	#AvgChars (S)	103.59	103.56	140.06
#AvgEnWords (R)	55.21	55.28	54.76	#AvgZhChars (R)	17.94	18.00	18.08
#AvgZhChars (R)	95.96	96.05	95.33	#AvgEnWords (R)	13.70	13.74	13.84
#AvgSentsWords	19.62	19.63	19.61	#AvgSentsChars	52.73	52.41	53.38
#AvgSents	40.62	41.08	40.25	#AvgSents	2.32	2.33	2.30

Wikipedia

Chinese Lead Section

NBA [编辑]

维基百科，自由的百科全书

 本条目存在以下问题，请协助[改善本条目](#)或在[讨论页](#)针对议题发表看法。

提示：此条目的主题不是**MBA**。

国家篮球协会（英语：National Basketball Association，缩写：**NBA**）是北美的男子**职业篮球**联盟，由30支球队组成（29支在美国以及1支在加拿大），分属两个联盟（Conference）：**东部联盟**和**西部联盟**；而每个联盟各由三个赛区（Division）组成，每个赛区有五支球队。NBA是**四大北美职业体育联赛**之一，并被视为全世界水准最高的男子职业篮球赛事^[2]。NBA前身是1946年于**纽约**成立的**全美篮球协会**（BAA）^{[1][3]}，1949年和**国家篮球联盟**（NBL）合并后改为现名。联盟总部位于**曼哈顿中城**，现任总裁为**亚当·萧华**。

NBA正式赛季于每年10月中开始，分为**常规赛**、**季后赛**两大部分。常规赛为循环赛制，每支球队都要完成82场比赛；常规赛到次年的4月结束(因疫情期间延至5月)，每个联盟的前八名将有资格进入接下来进行的季后赛。季后赛采用七战四胜赛制，共分四轮；季后赛的最后一轮也称为**总决赛**，由两个联盟的冠军争夺NBA的最高荣誉——**总冠军**。整个NBA赛季当中，常规赛完结之后分区冠军不设奖杯，只给予得奖球队锦旗一个，但联盟冠军及总冠军均设有奖杯加锦旗。

English Lead Section

National Basketball Association

From Wikipedia, the free encyclopedia

"NBA" redirects here. For other uses, see NBA (disambiguation).

The **National Basketball Association** (NBA) is a **professional basketball league** in North America. The league is composed of 30 teams (29 in the United States and 1 in Canada) and is one of the four **major professional sports leagues in the United States and Canada**. It is the premier men's professional basketball league in the world.^[1]

The league was founded in **New York City** on June 6, 1946, as the **Basketball Association of America** (BAA).^[2] It changed its name to the National Basketball Association on August 3, 1949, after merging with the competing **National Basketball League** (NBL).^[3] The NBA's regular season runs from October to April, with each team playing 82 games. The league's **playoff** tournament extends into June. As of 2020, NBA players are the world's best paid athletes by average annual salary per player.^{[4][5][6]}

The NBA is an active member of **USA Basketball** (USAB),^[7] which is recognized by the **FIBA** (International Basketball Federation) as the **national governing body** for basketball in the United States. The league's several international as well as individual team offices are directed out of its head offices in **Midtown Manhattan**, while its **NBA Entertainment** and **NBA TV** studios are directed out of offices located in **Secaucus, New Jersey**.

In North America, the NBA is the third wealthiest **professional sport league** after the **National Football League** (NFL) and **Major League Baseball** (MLB) by **revenue**, and among the top four in the world.^[8]

The **Milwaukee Bucks** are the reigning champions, having beaten the **Phoenix Suns** 4–2 in the **2021 Finals**.

Chinese Content

...

English Content

....

XWikis

- Twelve language pairs and directions for four European languages, namely Czech, English, French and German
- Take 7,000 titles in the intersection across all language sets. (XWikis-comparable)

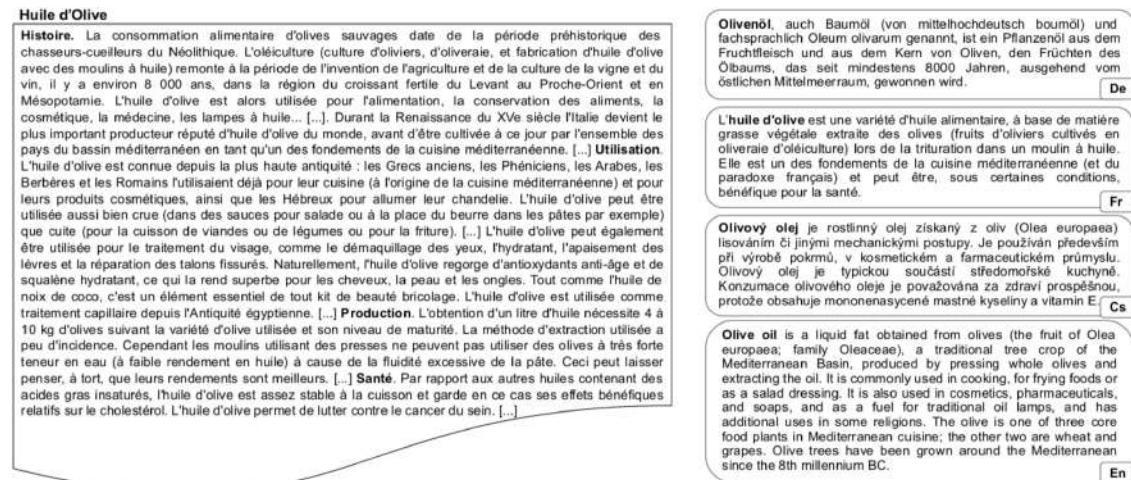


Figure 1: Example source document in French and target summaries in German, French, Czech and English.

	y	en	de	fr	cs
x					
en			425,279	468,670	148,519
de		376,803		252,026	109,467
fr		312,408	213,425		91,175
cs		64,310	53,275	51,578	

Table 1: Total number of document-summary pairs in the XWikis corpus considering all language pairs and directions. Each table cell corresponds to a cross-lingual dataset $\mathcal{D}_{x \rightarrow y}$.

MassiveSumm

- A large-scale multilingual summarization dataset containing articles in **92 languages**, spread across **28.8 million articles**
- Manual select news platform → automatic collection (*archive.org*) → quality control
- Use the summaries provided in the **HTML metadata**.

language	genus (family)	empty src	empty tgt	prefix	ellipsis	ellipsis/prefix	all-prefix	all-ellipsis	all	count	%invalid	valid count	
AFRICA													
2	Swahili	Bantoid (F4)	10,219	48,054	52,166	93,395	144,911	151,246	110,383	202,762	302,565	67.01%	99,803
3	Hausa	West Chadic (F13)	22,753	27,319	42,966	34,402	77,355	84,289	93,015	127,242	233,608	54.47%	106,366
4	Somali	Lowland East Cushitic (F13)	18,112	1,385	39,122	121,122	160,235	138,866	57,903	177,979	204,717	86.94%	26,738
5	Afrikaans	Germanic (F11)	374	8	121,056	5,549	126,173	5,927	121,434	126,551	198,792	63.66%	72,241
6	Kinyarwanda	Bantoid (F4)	17,791	6,878	40,893	21,241	62,062	45,307	65,477	86,128	92,674	92.94%	6,546
7	Amharic	Semitic (F13)	12,247	3,945	21,694	2,002	23,483	17,952	37,675	39,433	84,732	46.54%	45,299
8	North Ndebele	Bantoid (F4)	26,731	7	10,267	1,988	12,209	28,660	37,004	38,881	51,202	75.94%	12,321
9	Shona	Bantoid (F4)	25,130	5	12,505	715	13,205	25,840	37,638	38,330	46,681	82.11%	8,351
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
EURASIA													
20	Russian	Slavic (F11)	26,564	27,482	432,521	91,252	491,426	145,096	486,458	545,270	1,284,433	42.45%	739,163
21	Spanish	Romance (F11)	36,434	101,844	85,805	428,547	513,726	564,728	223,487	649,907	1,216,217	53.44%	566,310
22	Ukrainian	Slavic (F11)	29,968	37,652	358,697	243,248	598,286	302,697	424,432	657,735	1,252,150	52.53%	594,415
23	Persian	Iranian (F11)	16,277	147,711	428,787	44,699	470,156	195,272	579,432	620,729	1,150,653	53.95%	529,924
24	Arabic	Semitic (F13)	44,039	216,084	403,561	6,296	408,247	263,573	661,071	665,524	1,186,870	56.07%	521,346
25	Chinese	Chinese (F9)	838,069	62,003	36,335	388,542	424,829	1,016,062	890,620	1,052,349	1,171,189	89.85%	118,840
26	German	Germanic (F11)	23,358	246,308	323,190	15,901	333,184	284,787	592,185	602,070	1,080,213	55.74%	478,143
27	Urdu	Indic (F11)	19,236	2,291	469,175	4,213	472,516	25,514	490,602	493,817	1,115,555	44.27%	621,738
28	Hindi	Indic (F11)	6,388	1,059	469,614	34,754	502,814	41,977	477,057	510,037	1,073,514	47.51%	563,477
29	French	Romance (F11)	31,711	112,622	249,625	323,869	564,598	458,696	388,211	699,425	1,007,129	69.45%	307,704
30	Polish	Slavic (F11)	6,808	39,910	435,591	22,334	454,093	68,471	482,246	500,230	983,252	50.88%	483,022
31	Vietnamese	Viet-Muong (F3)	532,441	21,410	125,609	81,298	199,344	590,681	672,481	708,727	920,166	77.02%	211,439
32	Bulgarian	Slavic (F11)	22,272	6,606	273,851	9,206	281,857	37,558	302,351	310,209	977,769	31.73%	667,560
33	Tamil	Southern Dravidian (F14)	1,074	11,654	703,881	126,331	829,332	138,242	715,826	841,243	886,482	94.90%	45,239
34	Hungarian	Ugric (F5)	17,332	28,724	220,577	1,229	221,511	43,082	262,478	263,364	885,749	29.73%	622,385
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
INTERNATIONAL													
86	Esperanto	Constructed (F11)	0	0	27	103	130	103	27	130	565	23.01%	435
NORTH AMERICA													
87	Haitian	Creoles and Pidgins (F6)	5,890	12	8,346	3,240	11,582	9,118	14,246	17,460	26,009	67.13%	8,549
PAPUNESIA													
88	Indonesian	Malayo-Sumbawan (F12)	57,358	7,899	131,349	81,850	213,191	146,982	196,586	278,323	492,909	56.47%	214,586
89	Filipino	Greater Central Philippine (F12)	5	0	40	52	92	57	45	97	294	32.99%	197
90	Tetum	Central Malayo-Polynesian (F12)	0	0	2	0	2	0	2	2	15	13.33%	13
91	Bislama	Creoles and Pidgins (F6)	3	0	0	0	0	3	3	3	4	75.00%	1
SOUTH AMERICA													
92	Aymara	Aymaran (F0)	32	0	110	104	213	129	142	238	827	28.78%	589
totals			2,981,925	1,775,581	10,315,099	5,145,760	15,238,148	9,404,789	14,891,856	19,497,177	31,940,180	58.04%	12,443,000

LANGUAGE FAMILY LEGEND	
Aymaran (F0)	
Kartvelian (F1)	
Altaic (F2)	
Austro-Asiatic (F3)	
Niger-Congo (F4)	
Uralic (F5)	
other (F6)	
Japanese (F7)	
Tai-Kadai (F8)	
Sino-Tibetan (F9)	
Mande (F10)	
Indo-European (F11)	
Austronesian (F12)	
Afro-Asiatic (F13)	
Dravidian (F14)	

Methods

Pipeline Systems

① Summarize-then-Translate



② Translated-then-Summarize

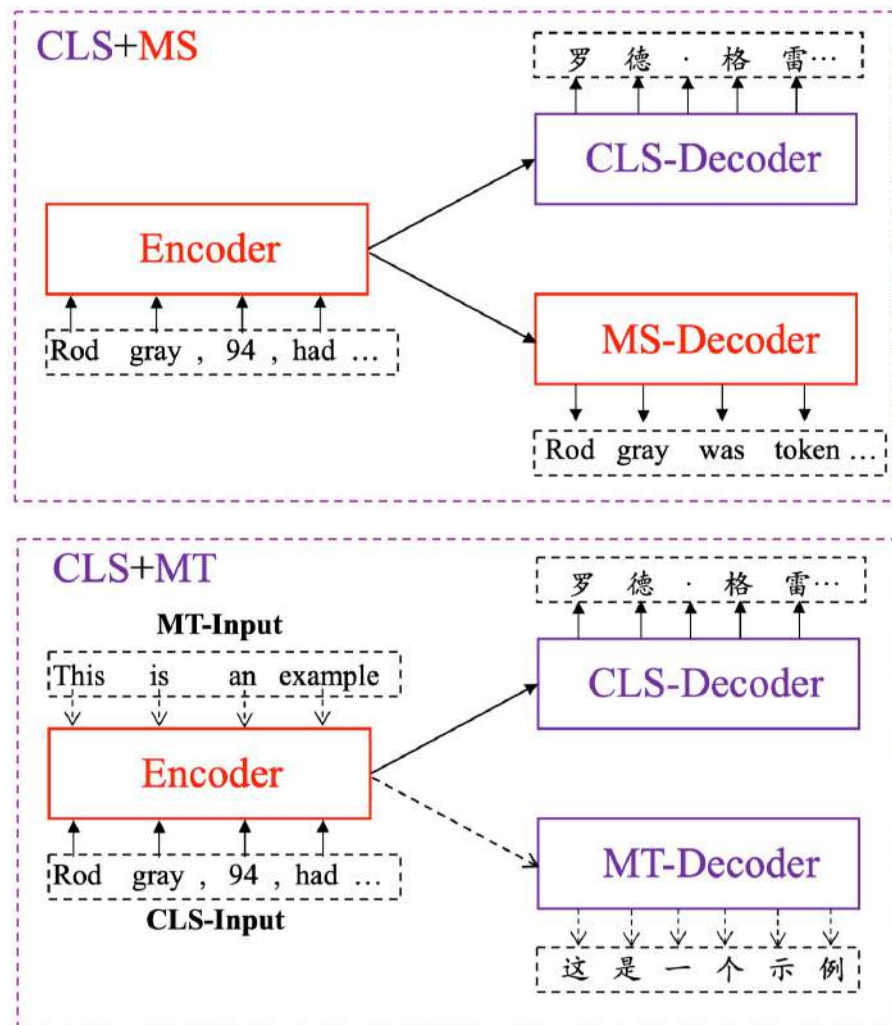


Mixed-Lingual Pre-training

Objective	Supervised	Multi-lingual	Inputs	Targets
Masked Language Model			France <X> Morocco in <Y> exhibition match.	<X> beats <Y> an
Denoising Auto-Encoder			France beats <M> in <M> exhibition .	France beats Morocco in an exhibition match.
Monolingual Summarization	✓		World champion France overcame a stuttering start to beat Morocco 1-0 in a scrappy exhibition match on Wednesday night.	France beats Morocco in an exhibition match.
Cross-lingual MLM	✓	✓	France <X> Morocco in <Y> exhibition match. 法国队在一场表演赛中击败摩洛哥队。	<X> beats <Y> an
Cross-lingual MLM	✓	✓	France beats Morocco in an exhibition match. <X>队在一场表演赛中<Y>摩洛哥队。	<X>法国<Y>击败
Machine Translation	✓	✓	France beats Morocco in an exhibition match.	法国队在一场表演赛中击败摩洛哥队。

Table 1: Examples of inputs and targets used by different objectives for the sentence “France beats Morocco in an exhibition match” with its Chinese translation. We use <X> and <Y> to denote sentinel tokens and <M> to denote shared mask tokens.

Multi-task Learning: MS+MT



- CLS+MS

- the reference in each of CLS datasets has a **bilingual version**.

- CLS+MT

- optimizes each task for a fixed number of mini-batches before **switching** to the next task

Multi-task Learning: + Monolingual Summarization

- Employing one unified decoder to generate the sequential **concatenation of monolingual and cross-lingual summaries**
- Making the **monolingual task** a prerequisite for the **cross-lingual task** through modeling interactions.

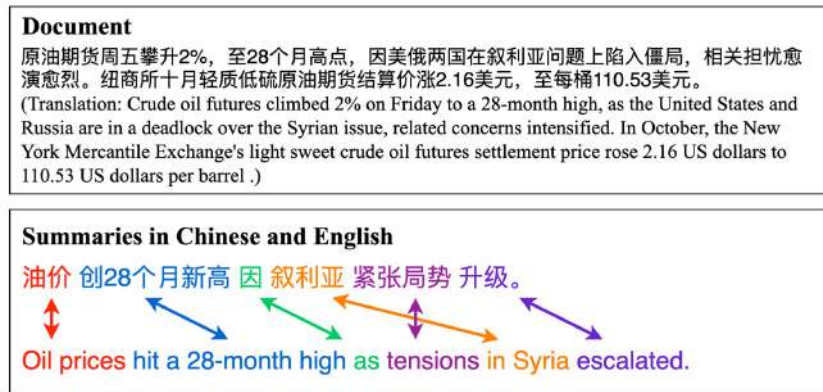


Figure 1: An example of the alignments across summaries in different languages. Each color represents phrases with one specific meaning.

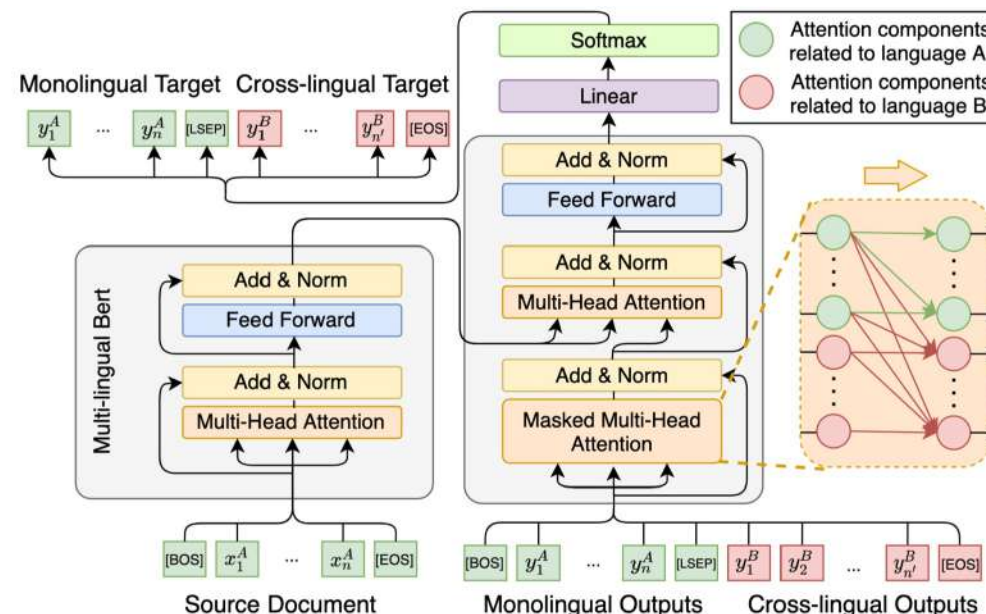


Figure 2: An overview of our proposed MCLAS. A unified decoder produces both monolingual (green) and cross-lingual (red) summaries. The green and red lines represent the monolingual and cross-lingual summaries' attention, respectively.

Multi-task Learning: + Machine Translation

- Task: Cross-Lingual Summarization with Compression Rate (CSC), regard MT task as a special CLS task with the compression rate of 100%.
- To bridge these two tasks smoothly, we propose a simple yet effective data augmentation method to produce document-summary pairs with different compression rates.

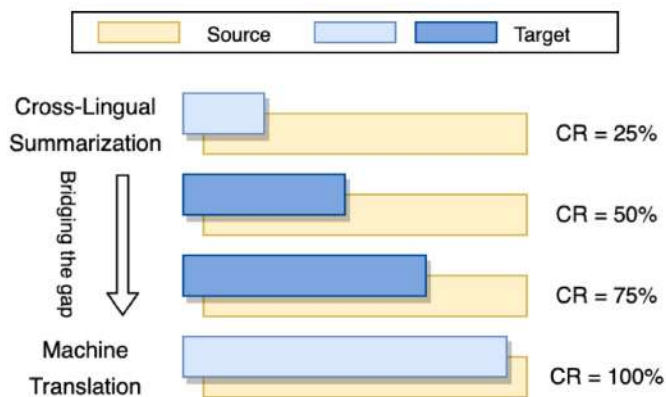


Figure 1: An illustration of the relationship between CLS and MT. The area of the text square represents its text length. CR means compression rate.

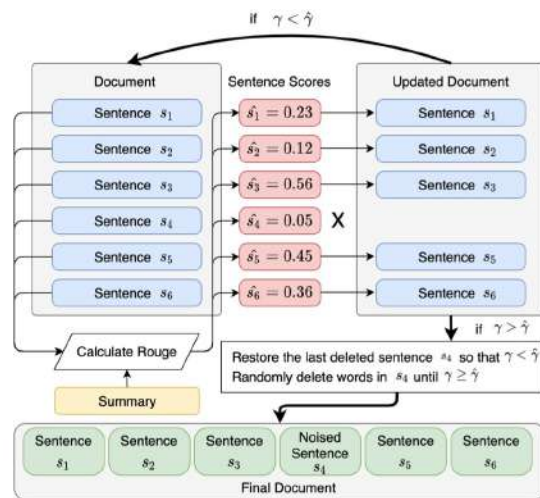


Figure 2: An example of our proposed compression rate based data augmentation method.

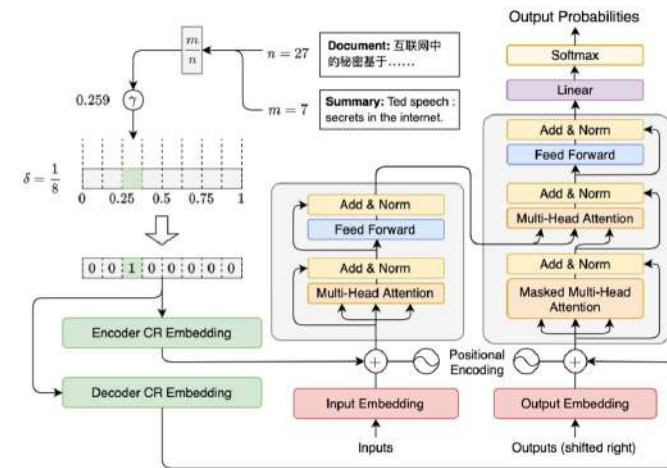
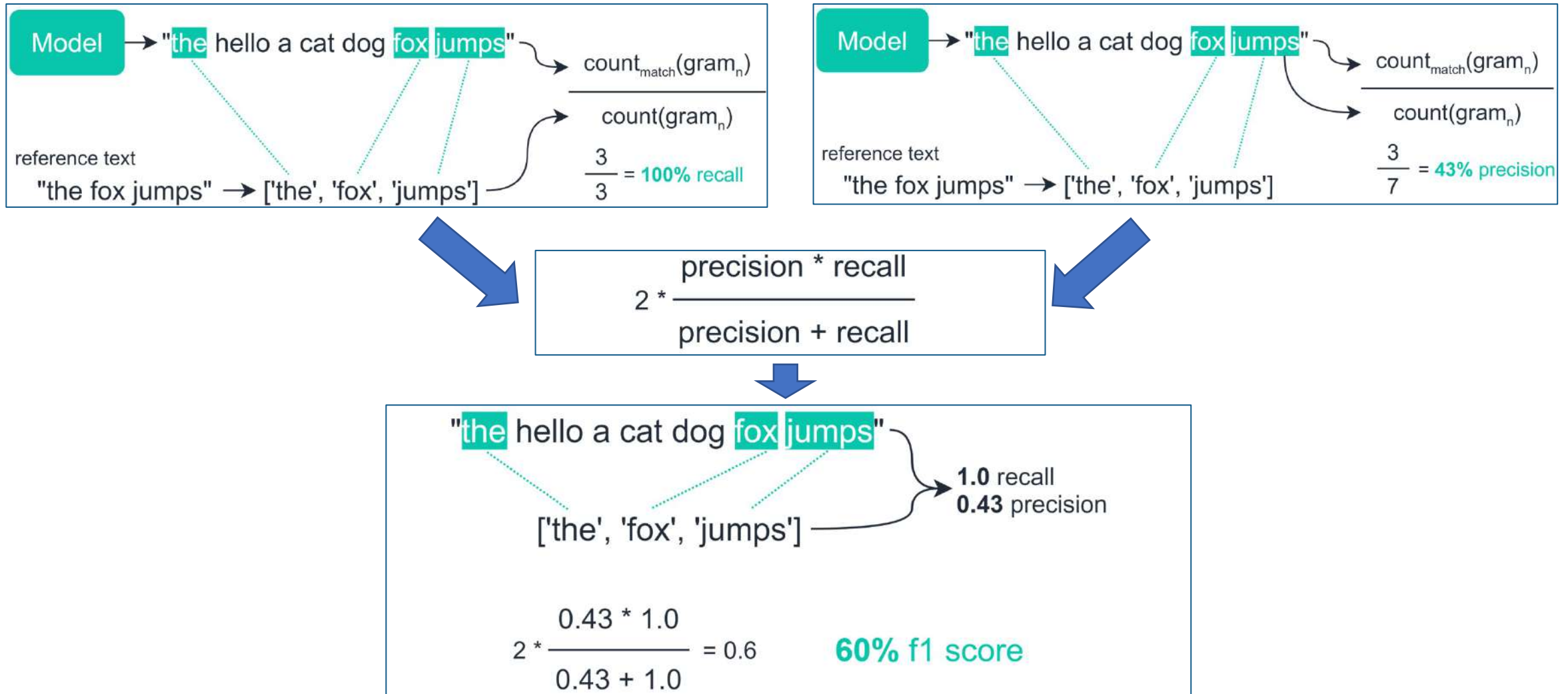


Figure 3: Model architecture of the modified Transformer incorporating compression rate.

Evaluation

Traditional ROUGE: ROUGE-1 For Example



① Multilingual ROUGE

Multilingual ROUGE Scoring

Overview

ROUGE is the de facto evaluation metric used for text summarization. However, it was designed specifically for evaluating English texts. Due to the nature of the metric, scores are heavily dependent on text tokenization / stemming / unnecessary character removal, etc. This repo tries to address these issues by adding the following main features using an adaptation of rouge-score: Google's rouge implementation.

- Enables multilingual ROUGE scoring by making use of popular word segmentation / stemming algorithms for various languages.
- Removes only punctuation characters according to unicode data tables as part of text normalization. This enables basic rouge scoring even with the absence of a segmenter / stemmer for any language.
- Provides an easy to use interface for using custom tokenization / stemming implementations.

Supported language names for stemming

bengali, hindi, turkish, arabic, danish, dutch, english, finnish, french, german, hungarian, italian, norwegian, portuguese, romanian, russian, spanish, swedish

Supported language names for word segmentation

chinese, thai, japanese, burmese

https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

② Multilingual ROUGE

ROUGE for Multilingual Summarization

Since the original summarization metric **ROUGE** is made only for English, we follow the method of [Hu et al.](#) and map words in other languages to numbers.

Languages without spaces (eg. Chinese, Japanese) will be segmented by characters and others will be split by spaces. For example, the Chinese text is split by characters, and the English words and numbers will be split by space.

```
[Input] Surface Phone将装载Windows 10 (The Surface Phone will be loaded with Windows 10)
[Segmentation] surface/phone/将/装/载/windows/10
```

Conclusion

Conclusion

- Cross-lingual summarization gains lots of research attentions recent days.
- Machine translation is an important related task, which needs to be explored more.
- Rather than single document cross-lingual summarization, multi-document cross-lingual summarization is also valuable.

Thanks~