The 6th Workshop on Asian Translation (WAT2019) 開催報告

中澤　敏明
東京大学

2019/11/19

@AAMT 2019, Tokyo

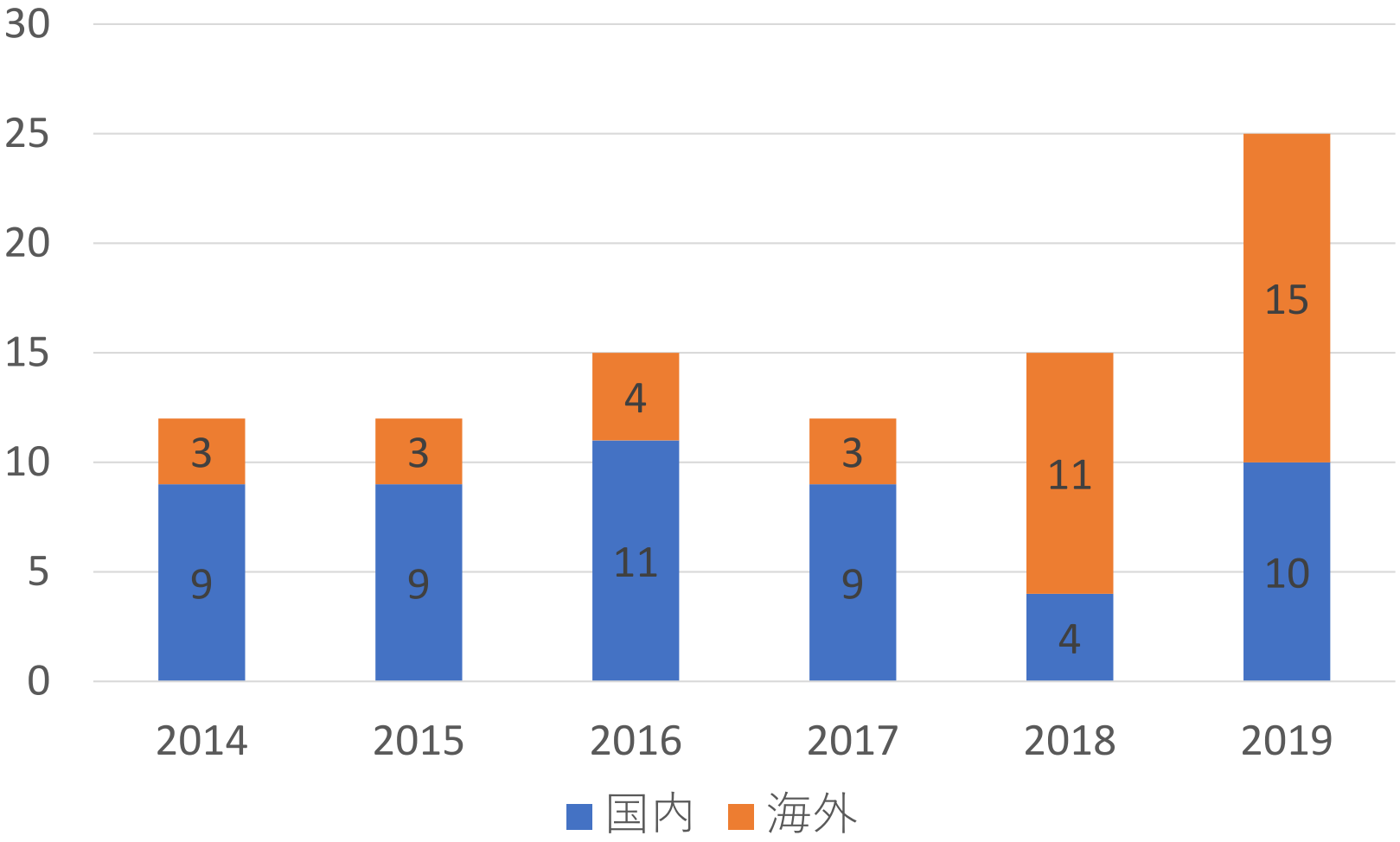
アジア翻訳ワークショップ (WAT)

- アジア言語を対象とした機械翻訳評価ワークショップ
  - 日本語、中国語、韓国語、インドネシア語、ヒンディー語、インド系諸語、ミャンマー語、ロシア語、クメール語、英語
- 2014年より毎年開催
  - WAT2014 @ 東京
  - WAT2015 @ 京都
  - WAT2016 @ 大阪 (Coling2016併設)
  - WAT2017 @ 台湾 (IJCNLP2017併設)
  - WAT2018 @ 香港 (PACLIC2018併設)
  - WAT2019 @ 香港 (EMNLP2019併設)
  - WAT2020 @ 中国・蘇州 (AACL-IJCNLP2020併設, 12月4-7日)

WATの翻訳タスク

ドメイン	言語対	2014	2015	2016	2017	2018	2019
科学技術論文	Ja ⇔ En	→	→	→	→	→	→
	Ja ⇔ Zh	→	→	→	→	→	→
特許	Ja ⇔ Zh		(Zh→Ja のみ) →	→	→	→	→
	Ja ⇔ Ko		(Ko→Ja のみ) →	→	→	→	→
	Ja ⇔ En			→	→	→	→
新聞	Id ⇔ En			→			
	Ja ⇔ En				→	→	→
	Ja ⇔ Ru						→
混合	Hi ⇔ En			→	→	→	→
	Hi ⇔ Ja			→	→	→	
	7 Indic ⇔ En					→	
	Ta ⇔ En						→
	My ⇔ En					→	→
	Km ⇔ En						→
レシピ	Ja ⇔ En				→	→	
適時開示文書	Ja → En						→
マルチモーダル	En → Hi						→

参加チーム数


30
25
20
15
10
5
0
3
9
3
9
4
11
3
9
11
4
15
10
2014
2015
2016
2017
2018
2019
国内
海外

WAT2019参加チーム一覧

Team ID	Organization	Country
TMU	Tokyo Metropolitan University	Japan
NICT	NICT	Japan
NTT	NTT Corporation	Japan
NICT-2	NICT	Japan
NICT-4	NICT	Japan
NICT-5	NICT	Japan
UCSYNLP	University of Computer Studies, Yangon	Myanmar
UCSMNLP	University of Computer Studies, Mandalay	Myanmar
cvit	IIIT Hyderabad	India
srcb	RICOH Software Research Center Beijing Co.,Ltd	China
sarah	Rakuten Institute of Technology	Japan
683	National Institute of Technology Silchar	India
KNU_Hyundai	Kangwon National University	Korea
NITSNLP	National Institute of Technology Silchar	India
ryan	Kakao Brain	Korea
PUP-IND	Punjabi University Patiala	India
FBAI	Facebook AI Research	USA
AISTAI	National Institute of Advanced Industrial Science and Technology	Japan
SYSTRAN	SYSTRAN	France
NHK-NES	NHK & NHK Engineering System	Japan
geoduck	Microsoft Research	USA
LTRC-MT	IIIT Hyderabad	India
ykkd	The University of Tokyo	Japan
IDIAP	Idiap Research Institute	Switzerland
NLPRL	Indian Institute of Technology (BHU) Varanasi	India

Team ID	ASPEC				JPC						TDDC	JIJI		NCPD	
	EJ	JE	CJ	JC	EJ	JE	CJ	JC	Ko-J	J-Ko	JE	EJ	JE	RJ	JR
TMU														✓	✓
NTT	✓	✓									✓				
NICT-2	✓	✓									✓				
NICT-5		✓												✓	✓
srcb	✓	✓	✓	✓											
sarah					✓	✓	✓	✓	✓	✓	✓	✓	✓		
KNU_Hyundai	✓	✓	✓	✓	✓	✓	✓	✓	✓						
ryan					✓	✓	✓	✓	✓	✓					
AISTAI	✓														
SYSTRAN														✓	✓
NHK-NES												✓	✓		
geoduck											✓				
ykkd		✓													

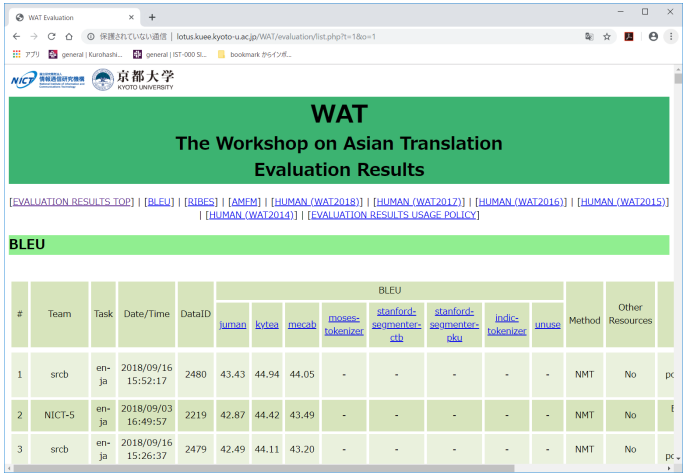
Team ID	Mixed-domain tasks								Mutimodal task
	ALT				IITB		UFAL (EnTam)		EV/CH
	EM	ME	E-Kh	Kh-E	EH	HE	ET	TE	EH
NICT	✓	✓							
NICT-4	✓	✓	✓	✓					
NICT-5						✓	✓	✓	
UCSYNLP	✓	✓							
UCSMNLP	✓	✓							
cvit					✓	✓	✓	✓	
sarah	✓								
683									✓
NITSNLP									✓
PUP-IND									✓
FBAI	✓	✓							
LTRC-MT						✓			
IDIAP									✓
NLPRL							✓	✓	

https://www.aclweb.org/anthology/events/emnlp-2019/#d19-52

研究論文

- 10件の投稿のうち7件を採択 (うち1件は後に取り下げ)
- 論文タイトル
  - Compact and Robust Models for Japanese-English Character-level Machine Translation
  - Controlling Japanese Honorifics in English-to-Japanese Neural Machine Translation
  - Designing the Business Conversation Corpus
  - Sentiment Aware Neural Machine Translation
  - Overcoming the Rare Word Problem for low-resource language pairs in Neural Machine Translation
  - Neural Arabic Text Diacritization: State of the Art Results and a Novel Approach for Machine Translation

翻訳結果評価



#	Team	Task	Date/Time	DataID	BLEU								Method	Other Resources
					juman	kytea	mecab	moses-tokenizer	stanford-segmenter-ctb	stanford-segmenter-pku	indic-tokenizer	unuse		
1	srcb	en-ja	2018/09/16 15:52:17	2480	43.43	44.94	44.05	-	-	-	-	-	NMT	No
2	NICT-5	en-ja	2018/09/03 16:49:57	2219	42.87	44.42	43.49	-	-	-	-	-	NMT	No
3	srcb	en-ja	2018/09/16 15:26:37	2479	42.49	44.11	43.20	-	-	-	-	-	NMT	No

- 自動評価
  - BLEU, RIBES, AM-FM
  - 自動評価サーバーに常時、翻訳結果を提出可
- 人手評価
  - 一対比較評価
  - 正確性評価
  - Direct Assessment (Multimodal task)
    - WMT(Conference on Machine Translation)の公式指標

- 評価費用はスポンサー費などによる（AAMT、NICT、サンフレア、川村インターナショナル、JPX）

一対比較評価 (Pairwise)

- 各システムの翻訳結果をベースライン（SMT or NMT）と文ごとに比較
- 評価対象文はテストセット中の400文
- 文ごとに
  - 5人の評価者がベースラインより良いか（+1）、悪いか（-1）、同程度か（0）を評価
  - 5人の評価を足し合わせて、+2以上ならば良い、-2以下ならば悪い、それ以外ならば同程度と判定
- 最終的に以下がスコア
  - 100 × (勝ち数 - 負け数) / 400

正確性評価 (Adequacy)

- 特許庁による特許文献機械翻訳の「内容の伝達レベルの評価」

評価値	評価基準
5	すべての重要情報が正確に伝達されている。 (100%)
4	ほとんどの重要情報は正確に伝達されている。 (80%～)
3	半分以上の重要情報は正確に伝達されている。 (50%～)
2	いくつかの重要情報は正確に伝達されている。 (20%～)
1	文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。 (～20%)

- 一対評価対象文中の200文を評価
- 文ごとに、2人が評価

Multi-Modal Translation と Direct Assessment

Sentence: 1

Is the English text (SRC) a good caption for the highlighted area of the image? : ○Yes ○No

SRC Text: Four baseball players on field.

Indicate to what extent each of these candidate translations expresses the meaning of the English source text (independently of the other candidate).

CAND1 Text: क्षेत्र में बेसबॉल खिलाड़ी

CAND1 Score: worst best

CAND2 Text: क्षेत्र में चार बेसबॉल खिलाड़ी।

CAND2 Score: worst best

評価結果（の一部）

詳細はoverview paperや各チームのsystem description paperを参照してください

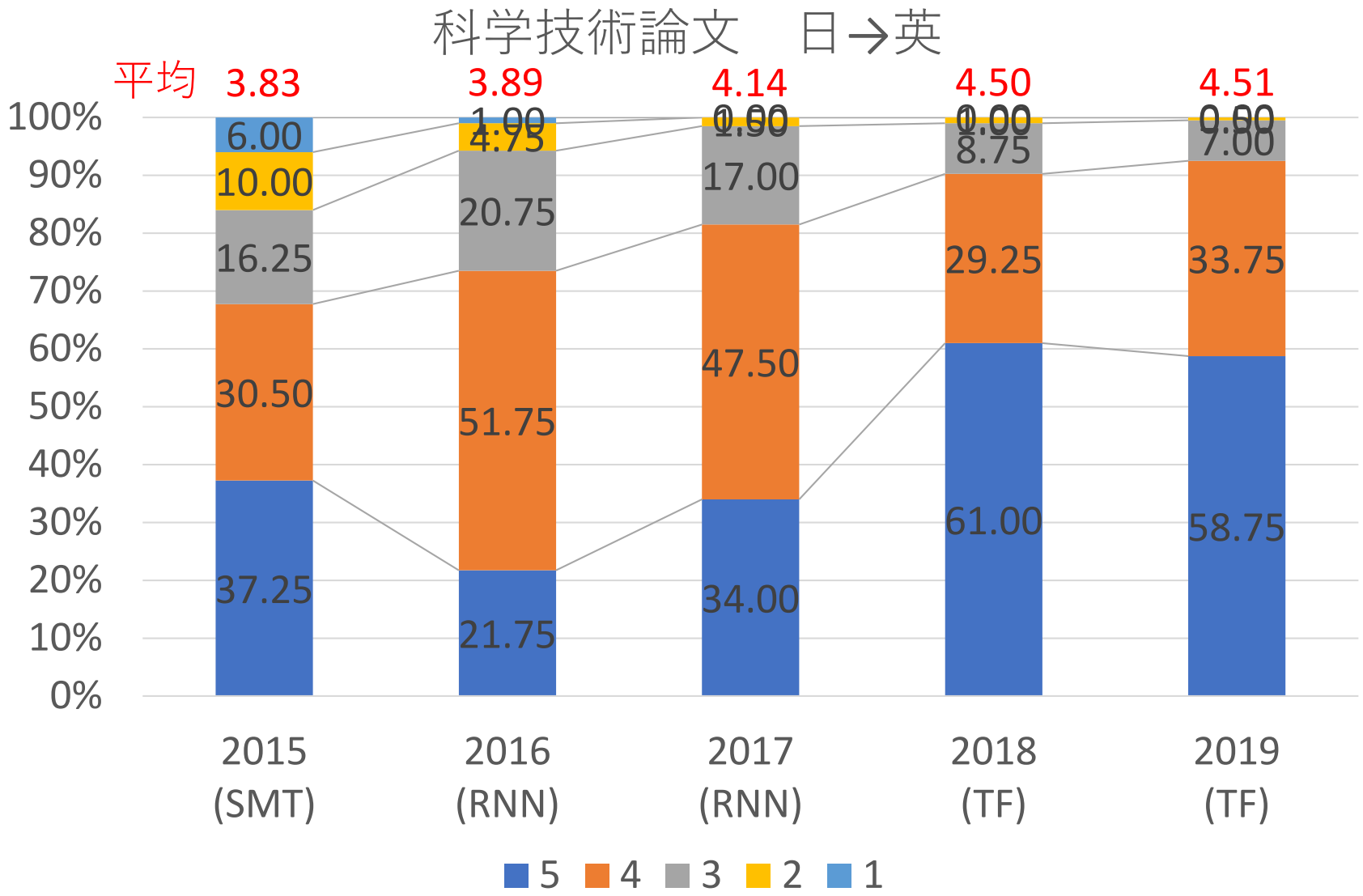
https://www.aclweb.org/anthology/events/emnlp-2019/#d19-52

用いられた技術

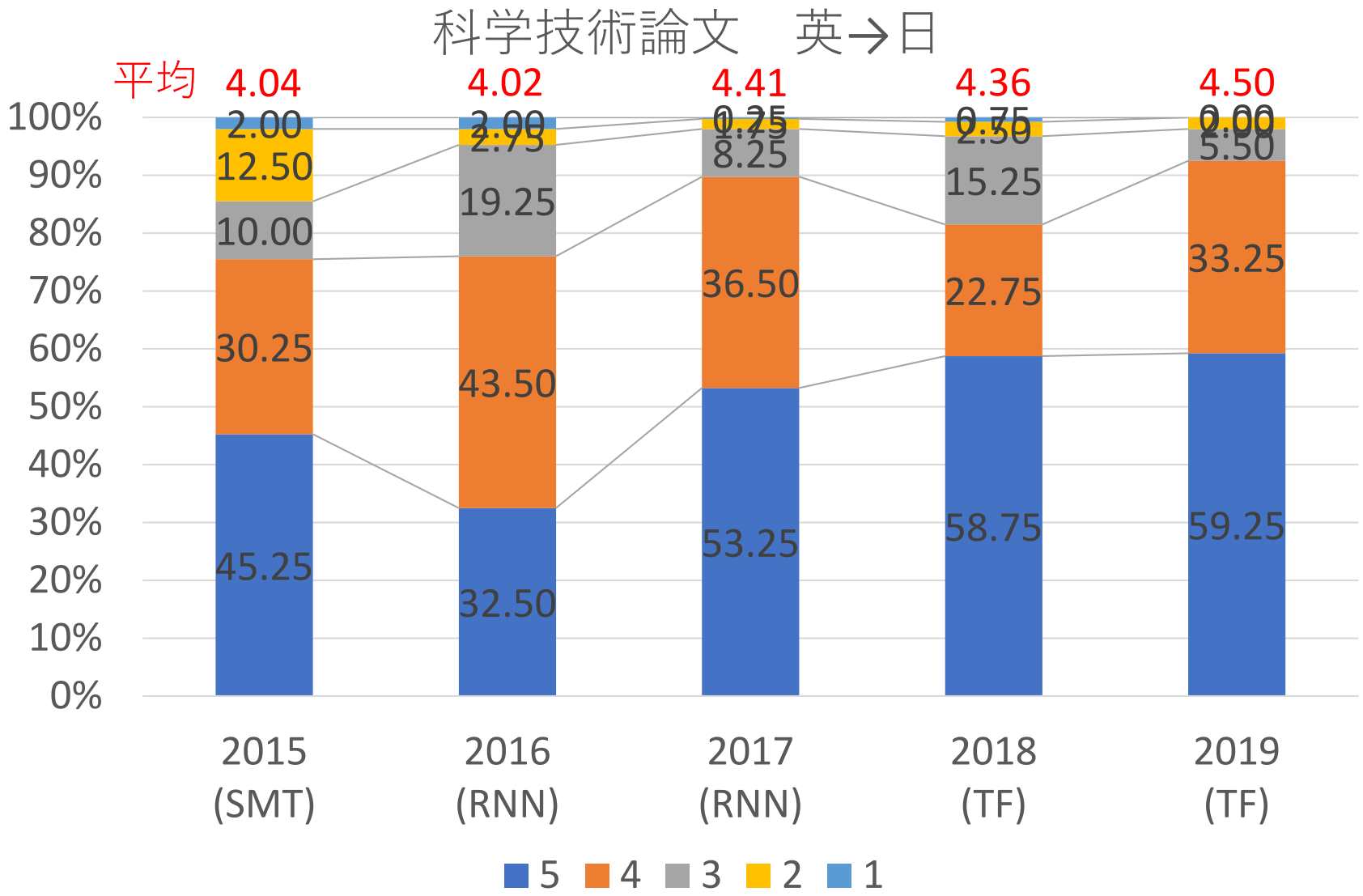
- Transformer
  - Relative positional encoding
- Data augmentation
  - Backtranslation
  - Multi-source translation
- Right-to-left reranking
- Additional resources
  - JParaCrawl
  - Manual translation and automatic extraction of Jiji Press news
- Task-specific fine-tuning
- Domain adaptation
- Long warm-up
- Weighted loss function
- Sentence-wise smooth Regularization
- Model Stabilization
  - Noise induction
  - Batch filtering
- Ensemble

科学技術論文と特許文

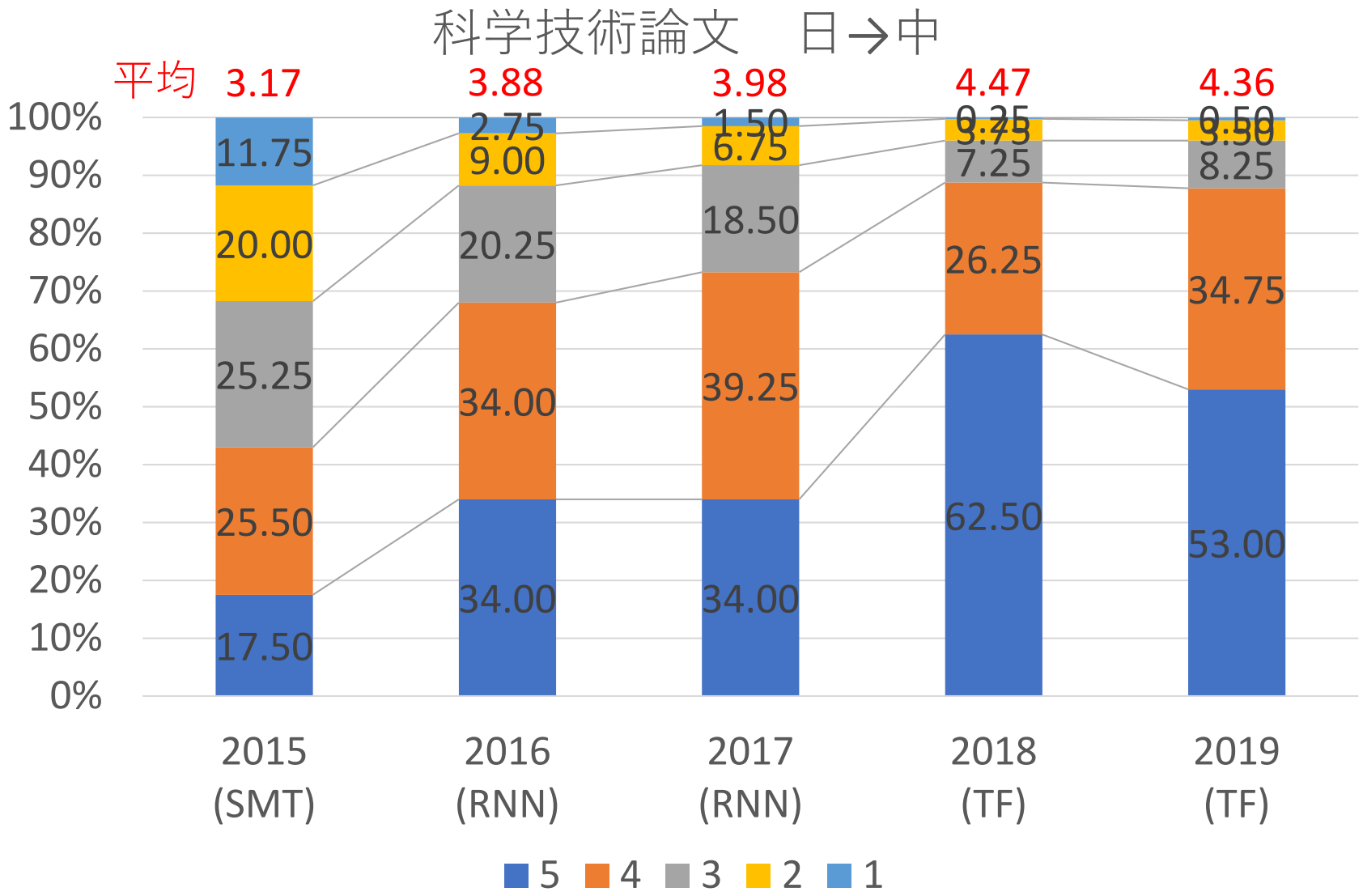
- どちらも文単位の翻訳品質はサチってきている
  - 5段階で平均4.5程度まで向上
- 文書単位での評価に移行した方が良いが・・・
  - 問題は文書単位のデータをどのように手に入れるか
- 次ページからのグラフは、これまでの各年度のトップチームの成績を並べたもの



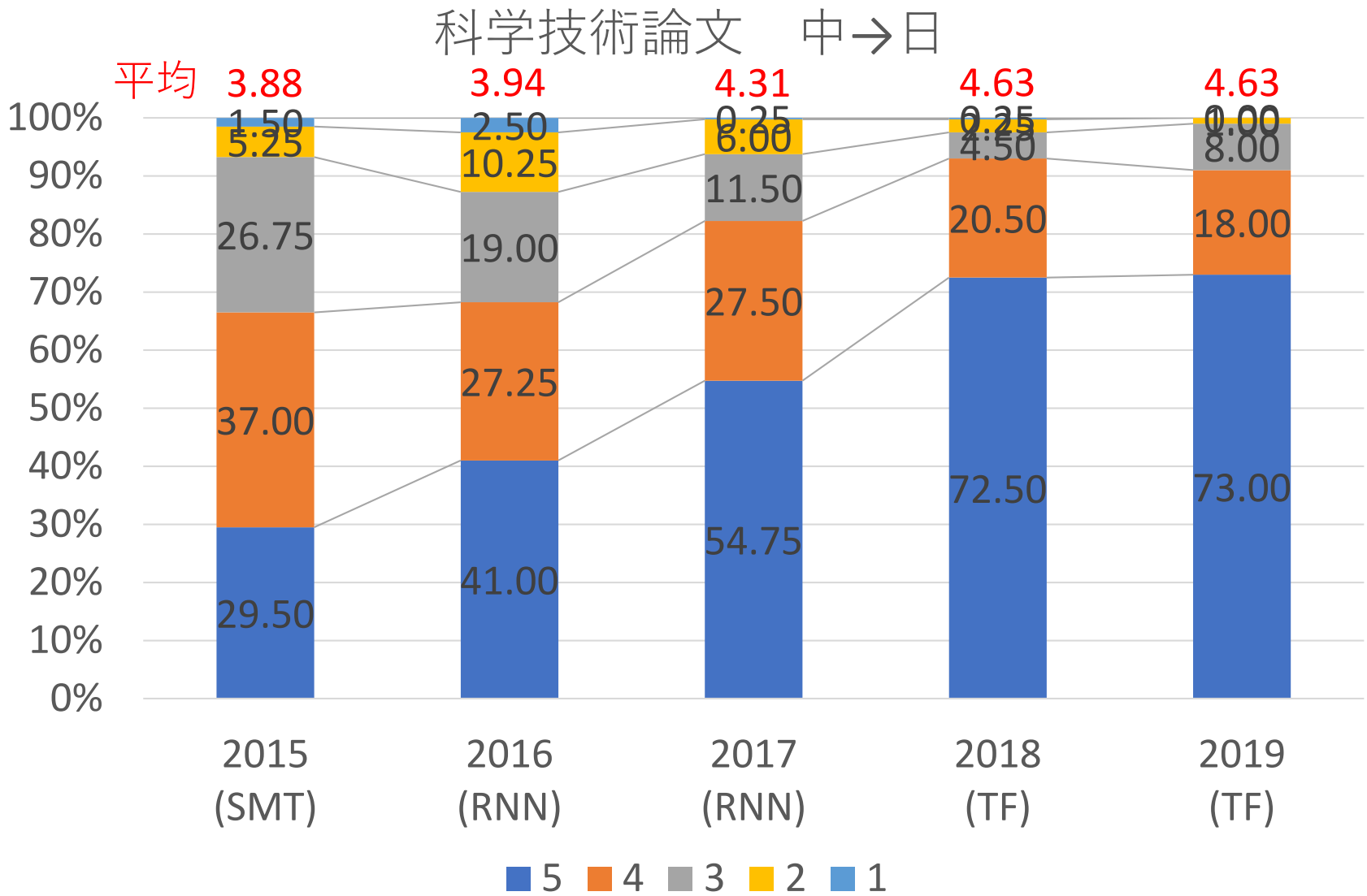
SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer



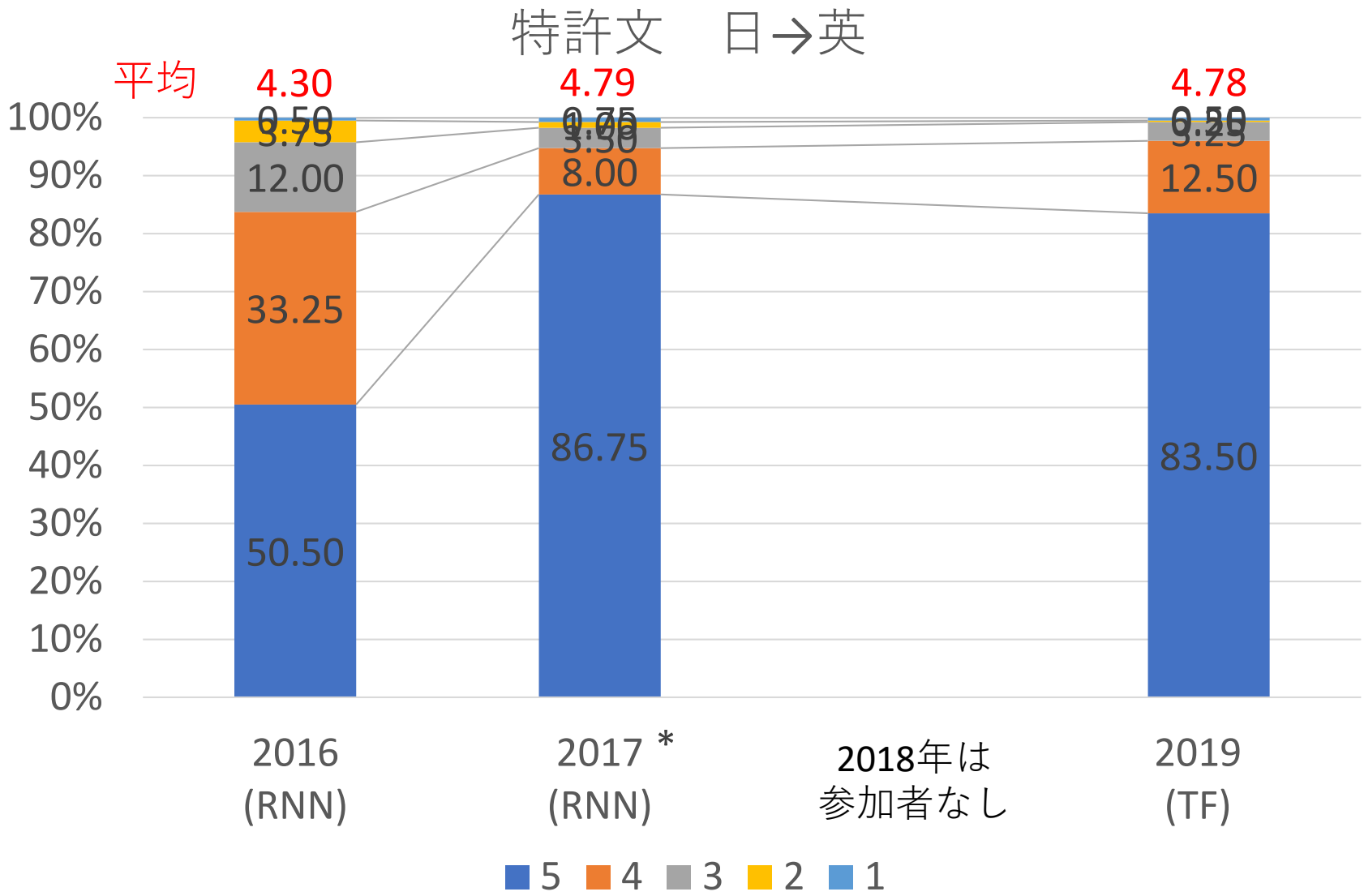
SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer



SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer

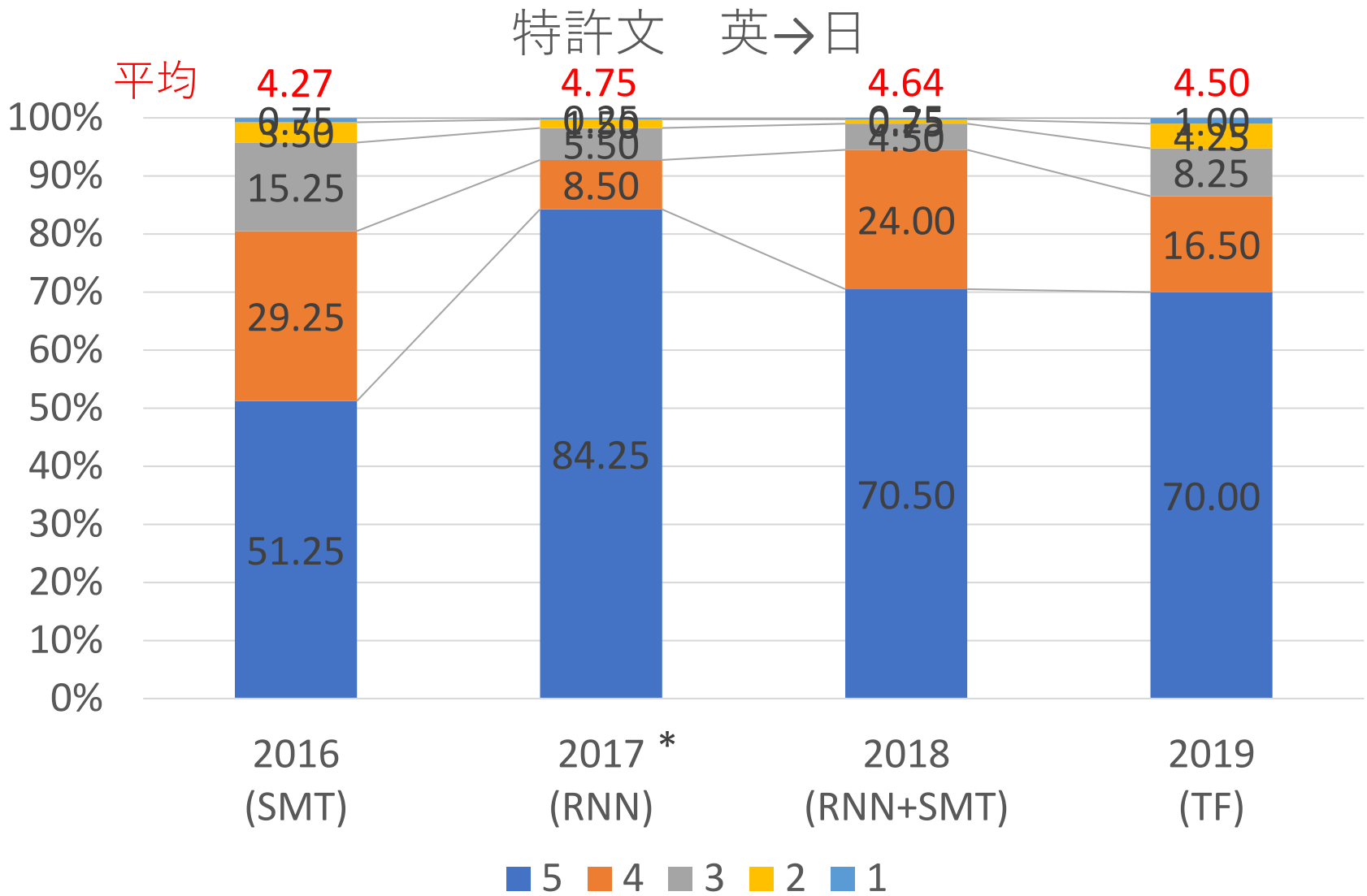


SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer



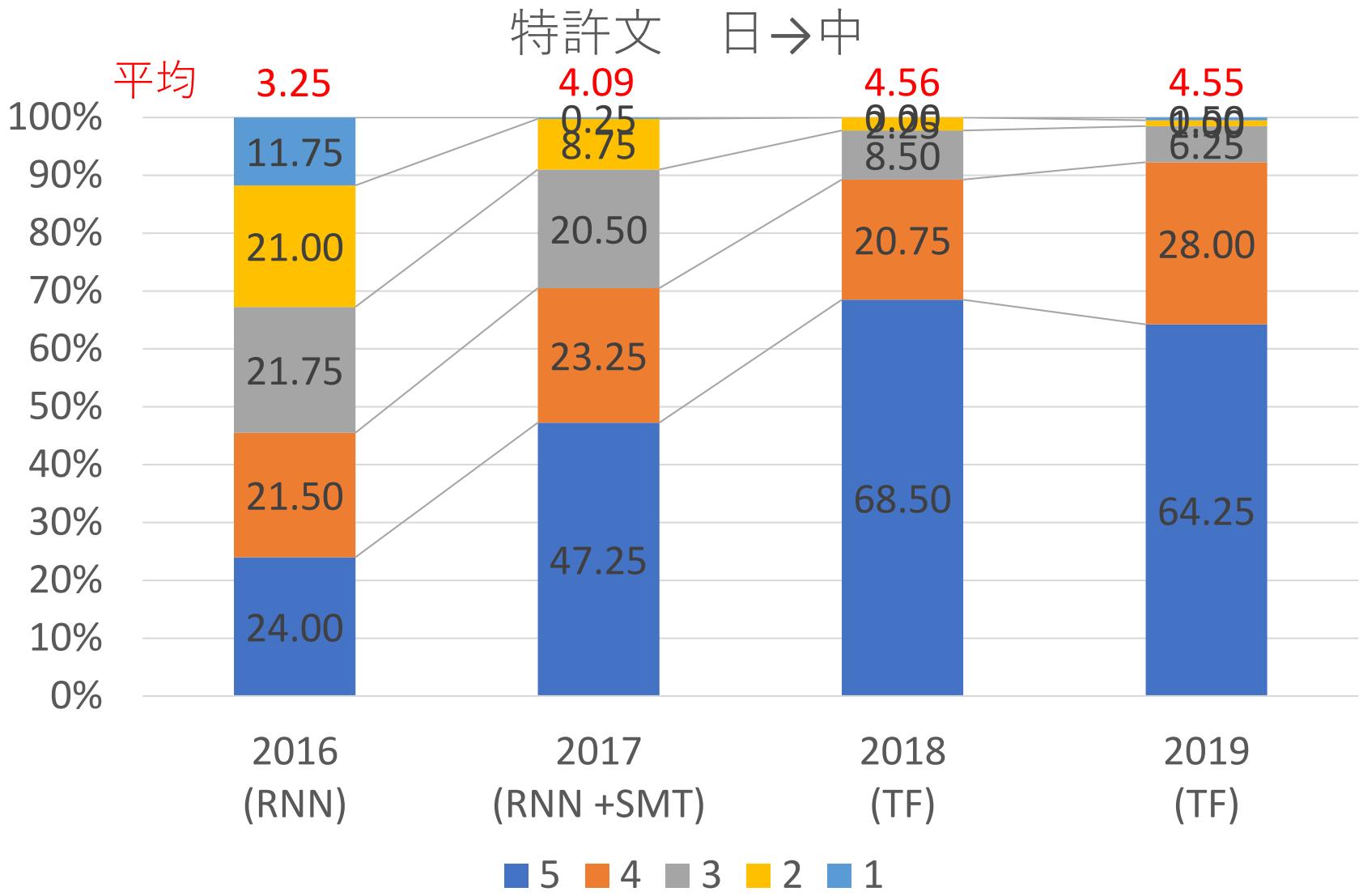
* 独自の対訳コーパスも利用

RNN: Recurrent Neural Network TF: Transformer

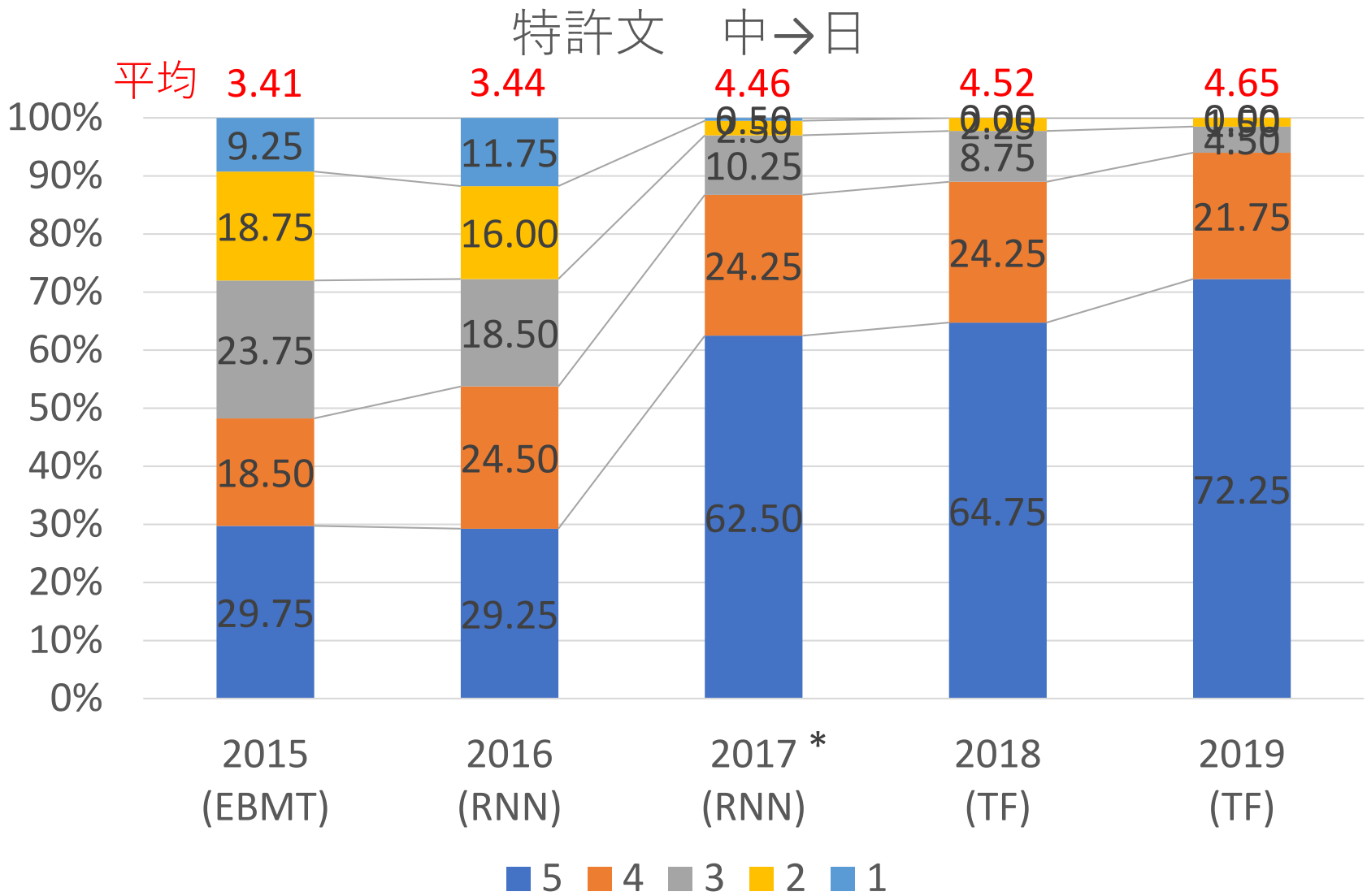


* 独自の対訳コーパスも利用

SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer

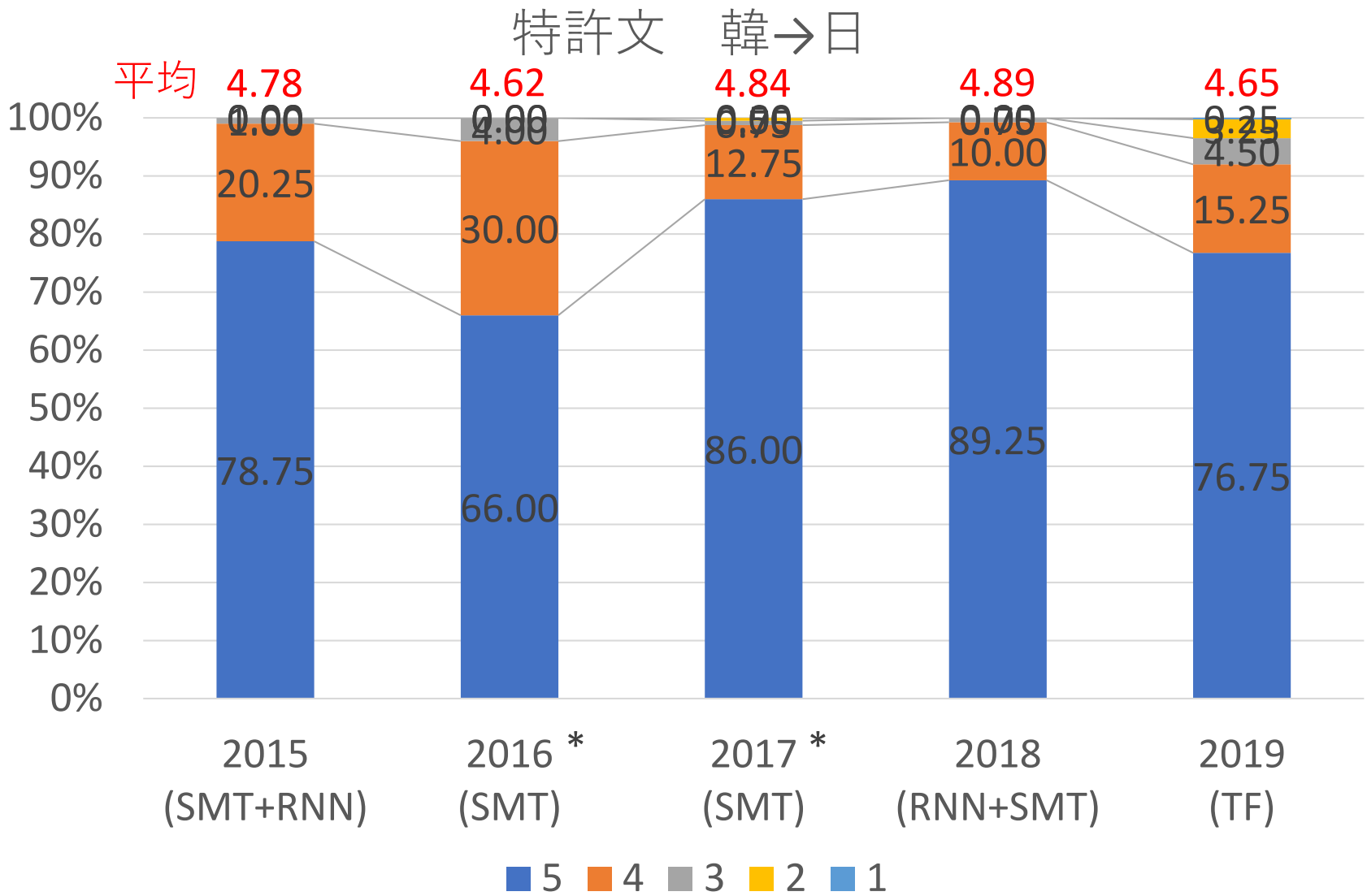


SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer



* 独自の対訳コーパスも利用

EBMT: Example-based Machine Translation RNN: Recurrent Neural Network TF: Transformer



* 独自の対訳コーパスも利用

SMT: Statistical Machine Translation RNN: Recurrent Neural Network TF: Transformer

適時開示情報

背景: 東京証券取引所 (TSE)

TSE listed companies

3,600

Market Capitalization

US$ 5.30 trillion

Foreigners: 30%

(End of 2018)

Annual Trading Value

US$ 7.21 trillion

Foreigners: 74% (As of 2018)

Timely Disclosure

77,000 docs/year

→ Over 500,000 pages/year

Investors

- Financial Institutions
- Individuals
- Foreigners

Examples of Timely Disclosure Documents

Update date	Heading
2019/1/29	Candidate for President of Japan Exchange Regulation
2019/1/29	Consolidated financial results for the nine months ended December 31, 2018
2019/1/29	[Explanatory Material] Overview of Earnings for Q3 FY2018
2019/2/22	Candidates for Directors and Executive Officers, etc.
2019/3/28	Basic Agreement Toward Realizing a Comprehensive Exchange
2019/3/28	3rd Medium-Term Management Plan

▶ TSE encourages the listed companies to **disclose information in English**.

背景: 東京証券取引所 (TSE)

TSE listed companies

3,600

Market Capitalization

US$ 5.30 trillion

Foreigners: 30%

(End of 2018)

Annual Trading Value

US$ 7.21 trillion

Foreigners: 74% (As of 2018)

Timely Disclosure

77,000 docs/year

→ Over 500,000 pages/year

Investors

- Financial Institutions
- Individuals
- **Foreigners**

Only 11% of timely disclosure documents are translated!
We need more **substantial** and **timely** information in English.

It is hard for us to translate more documents.
The main causes are **costs of translations, shortage of translators,** and **demand for timeliness.**

- It is not easy to meet demand for translations with **only human resources**.
- **Machine translation** can be considerable as one option.

Timely Disclosure Documents Corpus (TDDC)

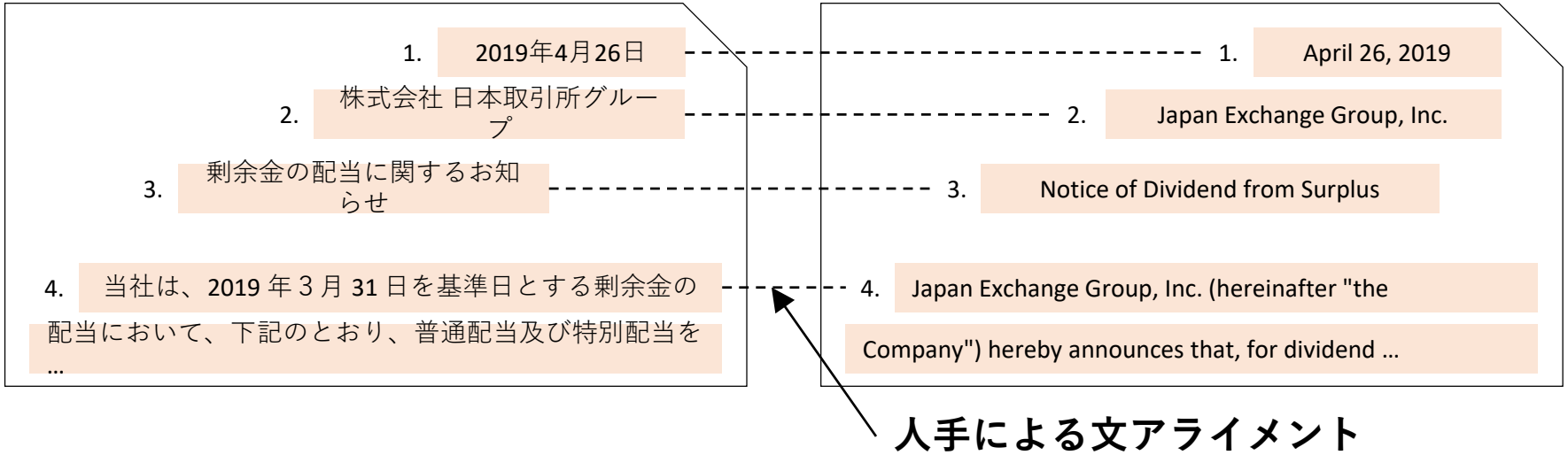
- 目的
  - 海外投資家向けに英語で正しい情報を素早く提供するために機械翻訳の精度向上を目指す
- キーポイント
  - 重要な数値、固有名詞

日本語	英語
2019年4月26日	April 26, 2019
株式会社 日本取引所グループ	Japan Exchange Group, Inc.
剰余金の配当に関するお知らせ	Notice of Dividend from Surplus
当社は、2018 年 3 月31 日を基準日とする剰余金の配当において、下記のとおり、普通配当に加えて記念配当を実施することとしましたので、お知らせいたします。	As a result, the year-end dividend per share for the fiscal year ended March 31, 2018 will be ¥43 (ordinary dividend of ¥33 plus commemorative dividend of ¥10).

'Items'

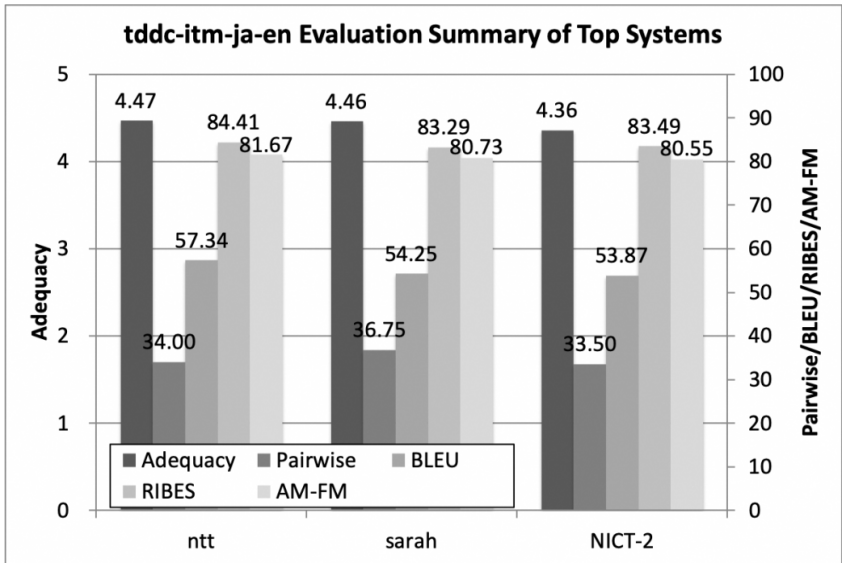
'Texts'

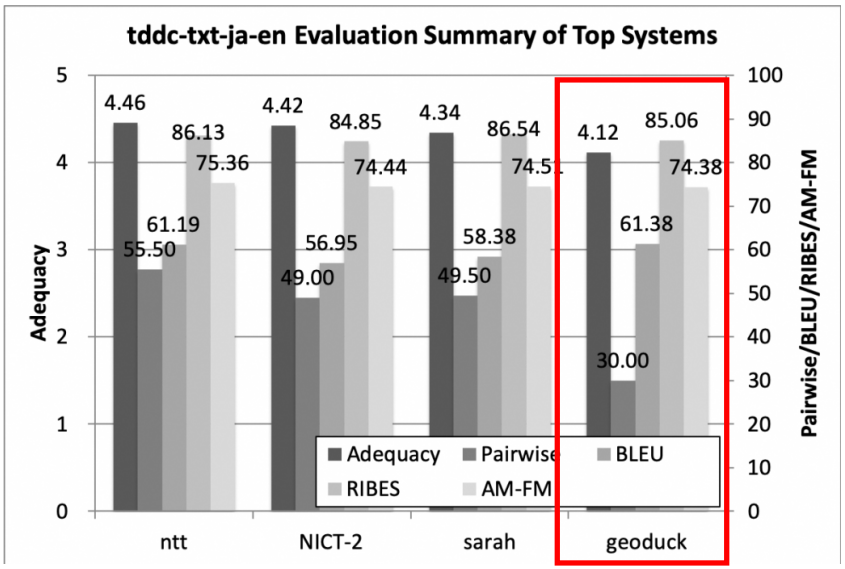
TDDCの構築



開示期間	文数 (重複を除く)		
2016 – 2017	**TRAIN_2016-2017** 1,089,346 (614,817)		
2018 (until 2018-06-30)	**TRAIN_2018** 314,649 (218,495)	**DEV_TEXTS** 1,153 (1,148)	**DEV_ITEMS** 2,845 (1,148)
		DEVTEST_TEXTS 1,114 (1,111)	**DEVTEST_ITEMS** 2,900 (2,671)
		TEST_TEXTS 1,153 (1,135)	**TEST_ITEMS** 2,129 (1,763)

TDDCの評価結果





人手評価で70%以上が評価4位上

- TDDCには必ずしも意味的に等価ではない文ペアが含まれており、翻訳エンジンに悪影響を与えている可能性がある

Japanese	Reference translation	Sample Output
当第3四半期連結会計期間末	the end of third quarter of **this fiscal year**	the end of the third quarter of **FY 2016**

TMとMTの組み合わせ

- 入力文との類似度の高い対訳文がある場合には、その訳文をそのまま出力 (TM)
  - Post-editはなし、つまり数字など異なる部分は直さずにそのまま出力
- 類似度の高い対訳文がない場合は機械翻訳 (MT)

- TDDCコーパスは重複や似た文が多いため、TMによりBLEUが稼げるが、数字などが異なるため人手評価は低い

重要な数値・固有名詞の評価

Japanese	English
株式会社日本取引所グループ	Japan Exchange Group, Inc.
業績予想及び配当予想の修正に関するお知らせ	Notice of Revision to Earnings Forecast and Dividend Forecast
当社は、2017年10月30日に開示しました2018年3月期(2017年4月1日～2018年3月31日)の通期連結業績予想及び期末の1株当たり配当予想について、下記のとおり修正することとしましたので、お知らせいたします。	We hereby announce that the consolidated earnings forecast and year-end dividend forecast for the fiscal year ending March 31, 2018 released on October 30, 2017 have been revised as follows.
剰余金の配当に関するお知らせ	Notice of Dividend from Surplus
これにより、2018年3月期の期末の1株当たり配当金は、普通配当33円に加え、記念配当10円を合わせた43円となります。	As a result, the year-end dividend per share for the fiscal year ended March 31, 2018 will be ¥43 (ordinary dividend of ¥33 plus commemorative dividend of ¥10).
投資活動によるキャッシュ・フローは、無形資産の取得による支出105億37百万円等により、261億64百万円の支出となりました。	There was cash outflow of ¥26,164 million from investment activities due mainly to ¥10,537 million in purchase of intangible assets.
発行済株式数に占める当社保有株式の比率	Shareholding ratio of JPX
SGXが保有する自己株式(515,063株)を含む。	Including the shares held by SGX as treasury stock (515,063 shares).

- 適時開示文書には重要な数値および固有名詞が多く含まれるが、NMTは一般にこれらの翻訳に弱い
- WATでの評価とは別に、これらも評価中

English -> Hindi Multi-Modal

概要

- 画像のキャプションを翻訳する際に、画像情報も利用

Street sign advising of penalty.

The penalty box is white lined.

An illustration of two meanings of the word “penalty” exemplified with two images (Hindi Visual Genome)

Tracks

- Text-Only Translation: translate short English captions (text) into Hindi. No visual information can be used.
- Hindi Captioning: generate captions in Hindi for the given rectangular region in an input image.
- Multi-Modal Translation: Given an image, a rectangular region in it and an English caption for the rectangular region, the task is to translate the English text into Hindi. Both textual and visual information can be used.

Results (Manual Evaluation)

Text-only

Multi-modal

Hindi Captioning

	Team ID	Data ID	Ave	Ave Z
EV TEXT	IDIAP	2956	72.85	0.70
	Reference		71.34	0.66
	683	3285	68.89	0.57
	683	3286	61.64	0.36
	NITSNLP	3299	52.53	0.00
CH TEXT	**Reference**		79.23	0.94
	IDIAP	3277	60.81	0.25
	IDIAP	3267	60.17	0.25
	683	3284	45.69	-0.28
	683	3287	45.52	-0.24
	NITSNLP	3300	28.48	-0.81
EV MM	**Reference**		70.04	0.60
	683	3271	69.17	0.61
	PUP-IND	3296	62.42	0.35
	PUP-IND	3295	60.22	0.28
	NITSNLP	3288	58.98	0.25
CH MM	**Reference**		75.96	0.76
	683	3270	54.51	0.08
	NITSNLP	3298	48.45	-0.20
	PUP-IND	3281	48.06	-0.13
	PUP-IND	3280	47.06	-0.17
EV HI	**Reference**		68.80	0.52
	NITSNLP	3289	51.78	-0.05
CH HI	**Reference**		72.60	0.61
	NITSNLP	3297	44.46	-0.35
	683	3304	26.54	-0.94

Multi-Modalタスクのまとめ

- 大規模なテキストデータのみを用いたシステムが画像情報も利用するシステムよりも良い結果
  - 参照訳よりも高精度
- 今後の課題
  - 画像情報が有効（必要）となるようなテストデータの構築

今後の展望

- アジアの翻訳研究の活性化、データ整備等を目的として2014年に始めたアジア翻訳ワークショップ（WAT）は一定程度の成果を得ている
  - 言語対、参加者数の増加
- 実用上問題となる訳抜け、専門用語や低頻度語の翻訳、訳語の統一、文脈レベルの翻訳などを評価できるタスク設計が必要
  - データの収集、評価指標の定義など
- WAT2020は2020年12月@中国・蘇州
  - スポンサー募集！