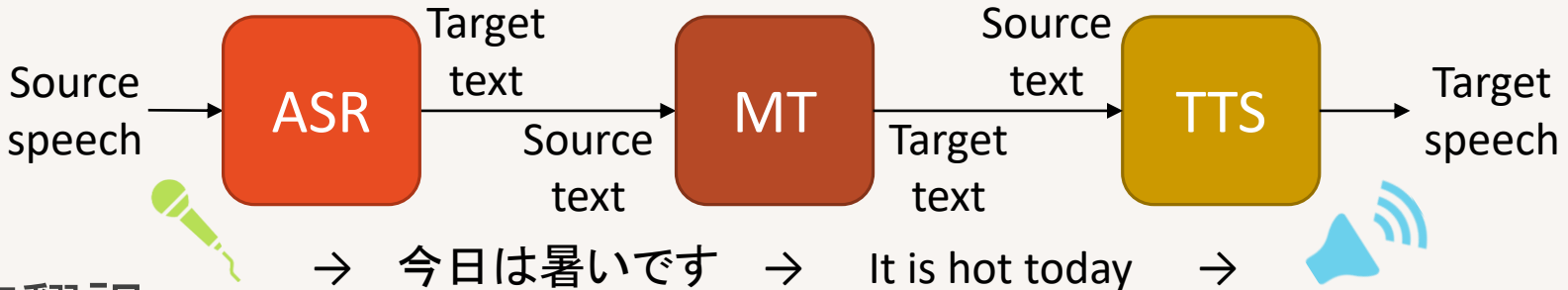
A4-1
言語情報とパラ言語情報を考慮したニューラル音声翻訳

徳山太顕[1] SAKRIANI SAKTI[1,2] 須藤克仁[1,2] 中村哲[1,2]

[1]奈良先端科学技術大学院大学

[2]理化学研究所革新知能統合研究センターAIP

音声翻訳(Speech-to-Speech Translation)



- 音声翻訳
  1. 音声認識 – ASR : Automatic Speech Recognition
  2. 機械翻訳 – MT : Machine Translation
  3. 音声合成 – TTS : Text to Speech
- 音声の持つ特徴
  - 言語情報
    - テキストで表される情報
  - 非言語情報 (**パラ言語情報**)
    - **強調**，抑揚，声の高さ etc…
- 問題点
  - パラ言語情報を反映できない

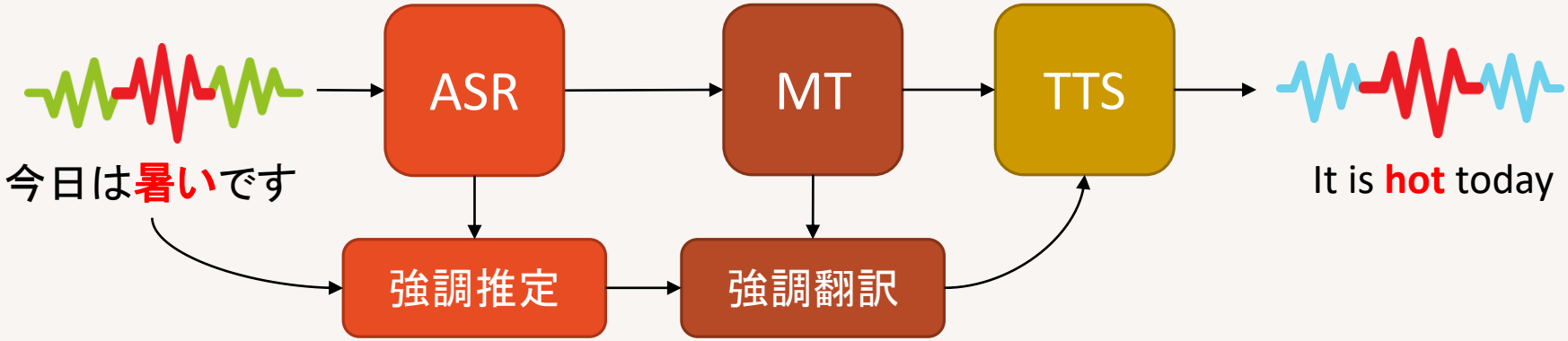
パラ言語情報の課題

- パラ言語情報を“どのように”他言語へ翻訳するか
  - 日本語から英語への翻訳の場合
  - 他言語同士の強調レベルは同じなのか？

- パラ言語情報を含めて翻訳する手法を検討する

先行研究

- Acoustic-to-Acoustic 強調翻訳 [1]
  - 入力音声と対応する強調箇所を翻訳後の出力音声にも反映させる

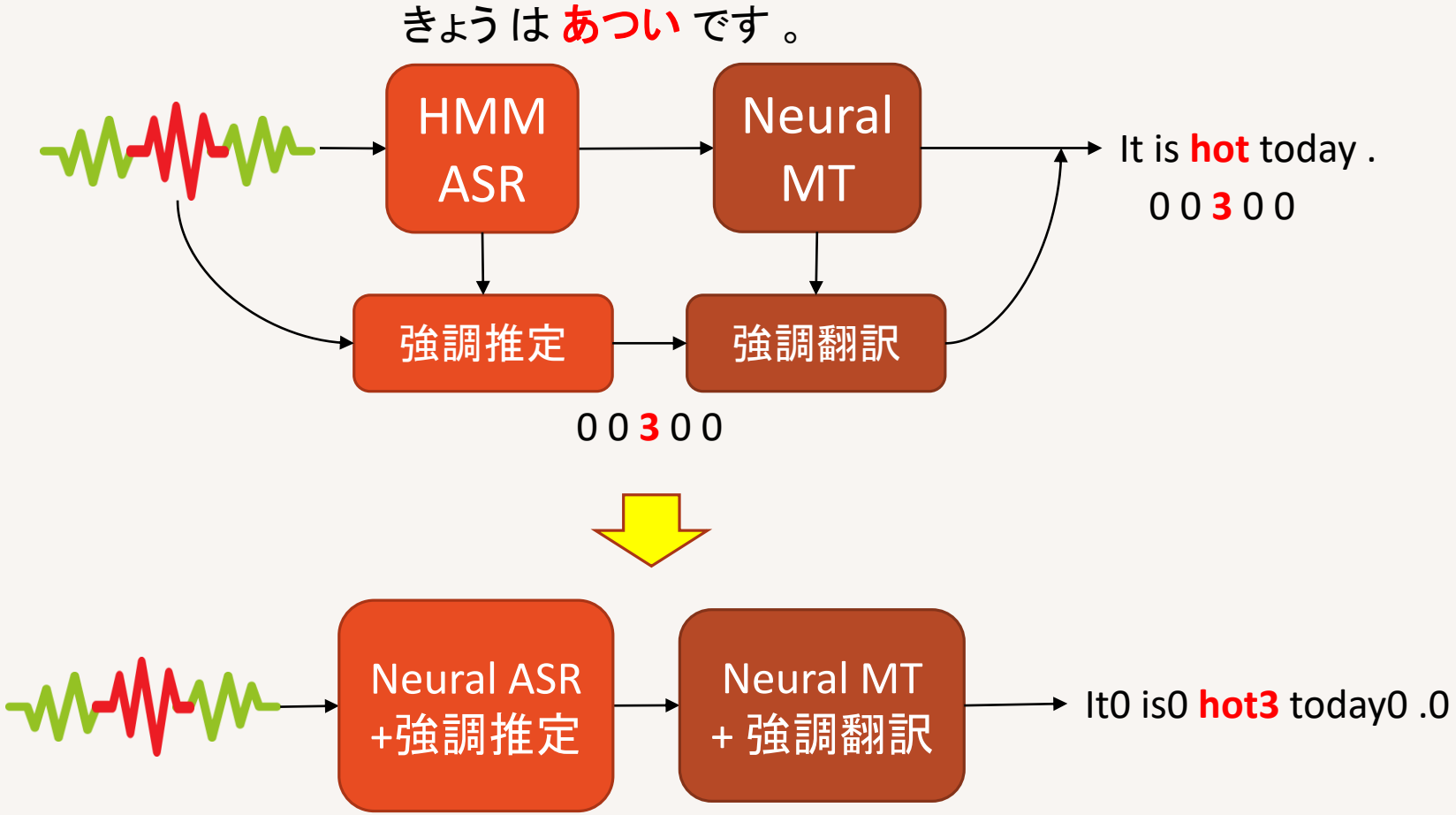


- 問題点
  - 強調の推定と翻訳は前の処理を待つ必要がある(カスケードモデル)
  - システム自体がかなり複雑
    - ASRとTTSはHMM(隠れマルコフモデル)ベース，MTはニューラルネットワーク
  - 音声の強調だけが他言語に対応するとは限らない
    - 言語によっては音声の強調以外で表現される可能性がある[2]

[1] Sequence-to-Sequence Models for Emphasis Speech Translation [Q. T. Do 2018]
[2] Toward Multi-Features Emphasis Speech Translation: Assessment of Human Emphasis Production and Perception with Speech and Text Clues [Q. T. Do 2018]

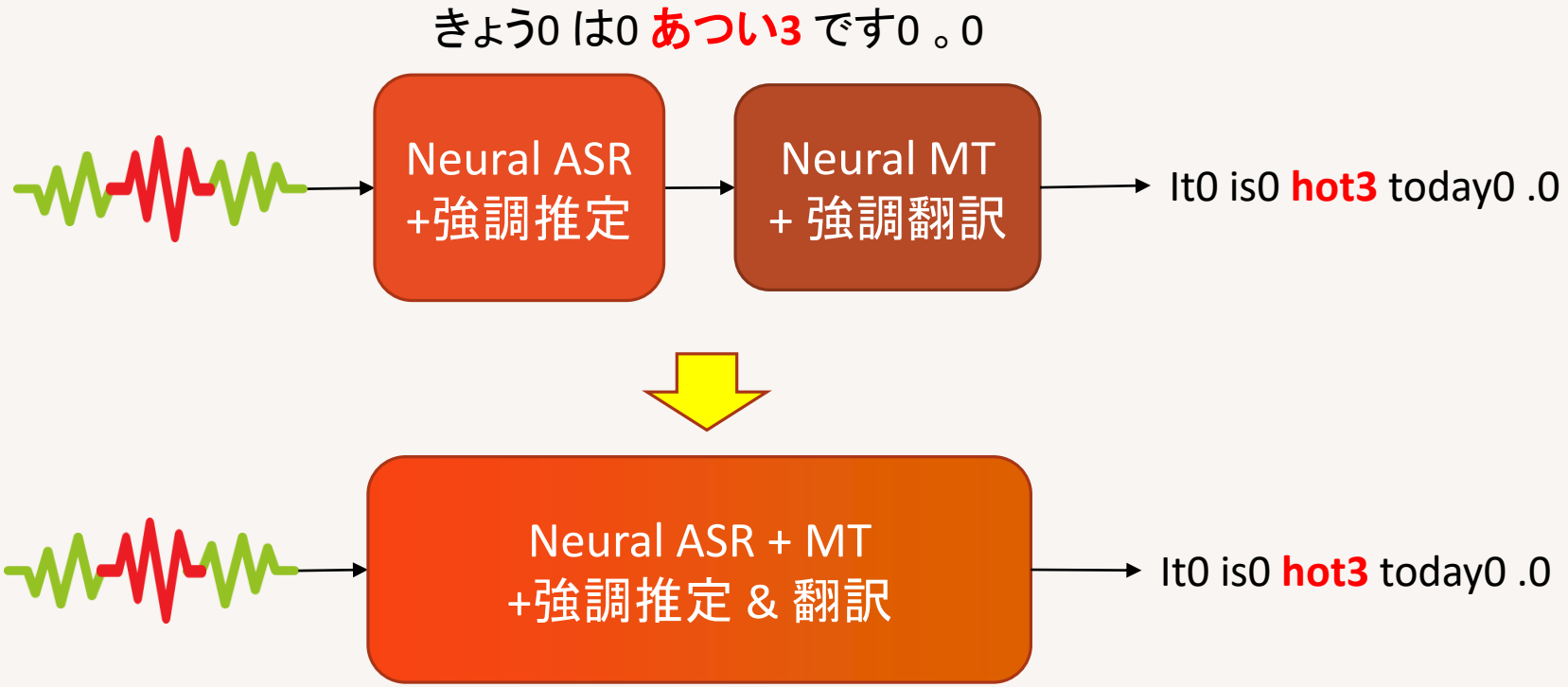
提案モデル1 – Cascade Neural S2T

- システムの単純化
  - 全てのシステムにおいてニューラルネットワークを利用
  - 強調推定，強調翻訳をASR, MT内で行う



提案モデル2 – Direct Neural S2T

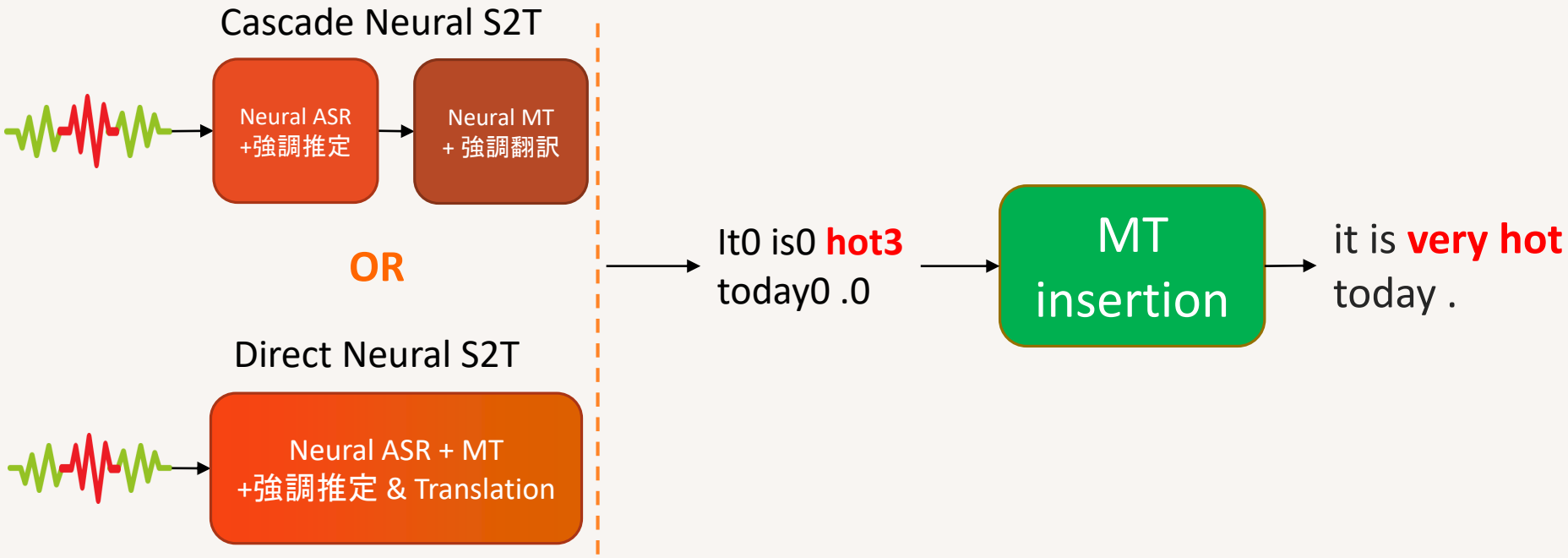
- ニューラルネットワークを用いた直接翻訳
  - 学習済み各システムを一つのモデルとして結合
  - 音声の強調を反映した翻訳
  - Yeら[3]の直接翻訳モデルを利用



[3] Direct speech-to-speech translation with a sequence-to-sequence model [Ye Jia 2019]

提案法 – パラ言語情報の翻訳 (音声→テキスト)

- 音声強調から言語強調への翻訳



- 目的に合わせてテキストに**強調表現**を埋め込んで学習
  - ASR ：言語，パラ言語の認識
  - MT ：言語，パラ言語の翻訳
  - MT-insertion: 言語的強調を含むテキストへと変換

提案手法 – テキストの強調表現

- データの編集
  - 強調が分かるようにテキストを変換

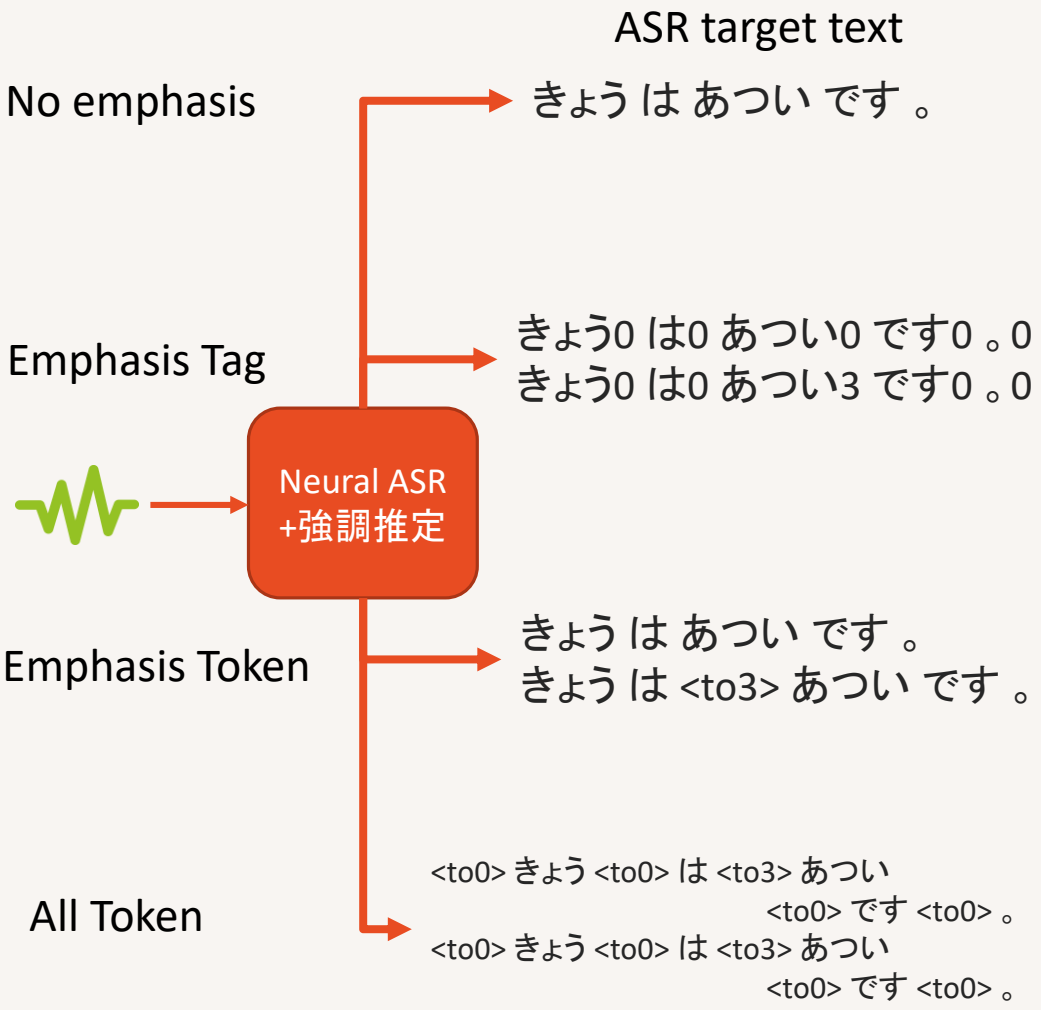
きょう0 は0 あつい0 です0 。0
it0 is0 hot0 today0 .0

きょう0 は0 あつい3 です0 。0
it0 is0 hot3 today0 .0

		例: hotの強調レベルが3の場合
No Emphasis	データに何も加えない	it is hot today .
Emphasis Tags	各単語に強調タグをつける	it0 is0 hot3 today0 .0
Emphasis 1-Token	強調単語の前にトークンを付ける	it is <to3> hot today .
Emphasis All Token	全単語の前にトークンを付ける	<to0> it <to0> is <to3> hot <to0> today <to0> .

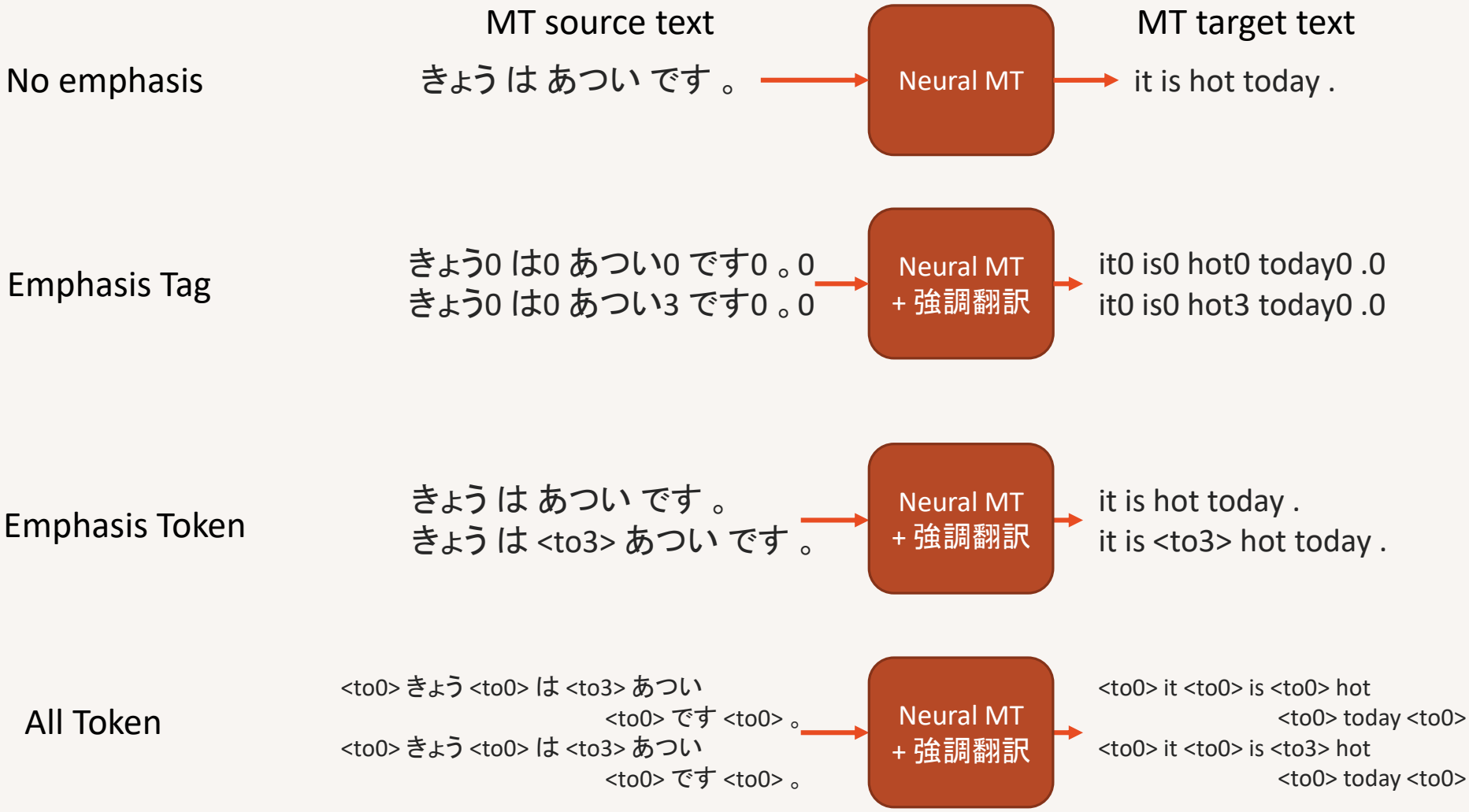
提案手法 ー 言語/パラ言語情報の認識

- ASR :同じ音声から強調を含んだテキストを認識できるよう学習



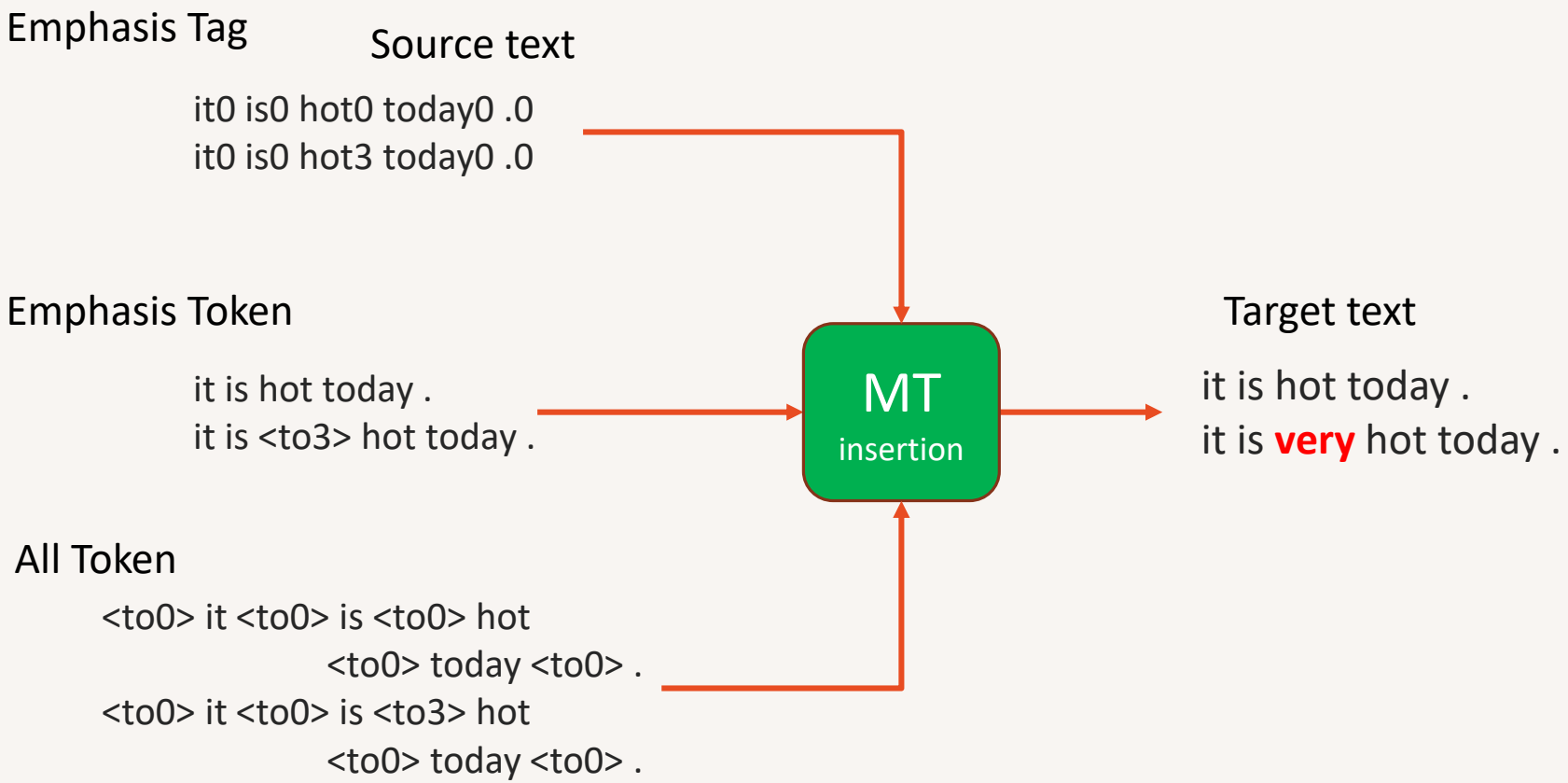
提案手法 ー 言語/パラ言語情報の翻訳

- MT : 強調された原文から強調された目的文を翻訳するよう学習



提案手法 – 挿入テキストの生成

- MT-insertion : 強調されたテキストから自然なテキストへと変換



モデルとデータセット

- 使用するモデル
  - 全システムにおいて Transformerを利用したOpenNMT-pyを使用

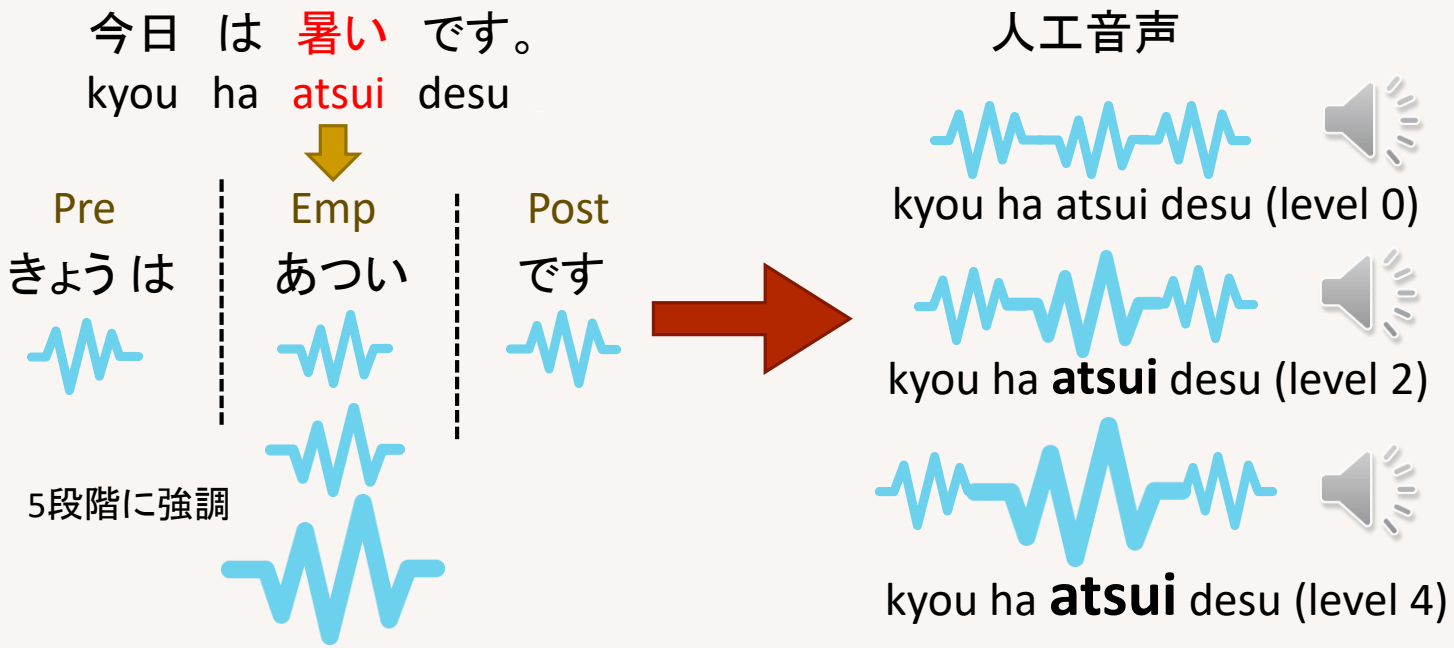
パラメータ	
Layer	3
RNN size	512
Word vec size	512
Transformer_ff	2048
Heads	8

- BTEC (Basic Travel Expressions Corpus)
- 自然音声[1]
  - BTECに基づいた“強調を含んだ”日英の音声とテキストのデータセット
    - 発話数1029かつ強調が5段階
  - 学習データ数としては非常に少ない
- データ拡張
  - BTECを利用してデータ拡張
  - 強調箇所を含んだ日英の人工データを作成する

[1] Sequence-to-Sequence Models for Emphasis Speech Translation [Q. T. Do 2018]

人工音声データの作成

1. 形容詞前後で分離
2. それぞれのテキストから音声を生成
3. 形容詞のみ5段階の音声を生成
   - Normal, Light, Medium, Strong and Very strong
4. 各生成音声を連結



人工テキストデータの作成

1. 形容詞前後で分離
2. それぞれの単語ごとに情報を追加する
   - 音声と同様に5段階

今日　は　暑い　です 。
It is hot today .

Pre	Emp	Post
きょう は	あつい	です 。
It is	hot	today .

Emphasis Tag

きょう0 は0 あつい3 です0 。0
It0 is0 hot3 today0 .0

Emphasis 1-Token

きょう は <to3> あつい です 。
It is <to3> hot today .

Emphasis All Token

<to0> きょう <to0> は <to3> あつい
<to0> です <to0> 。
<to0> it <to0> is <to3> hot <to0> today <to0> .

挿入テキストの作成

- 自然音声の単語が挿入されたテキストデータより抽出

Ex.1
level 0 : This is loose. (plain text)
level 1 : This is a tad loose.
level 2 : This is noticeably loose.
level 3 : This is so loose.
level 4 : This is completely loose.

Ex.2
My cholesterol is high.
My cholesterol is a bit high.
My cholesterol is quite high.
My cholesterol is certainly high.
My cholesterol is extremely high.

挿入された単語を抽出しリスト化

Light : **a tad, a bit …**
Medium : **noticeably, quite …**
Strong : **so, certainly …**
Very strong: **completely, extremely …**

- 形容詞の前に頻出の単語を挿入

I feel sick . 

Normal : I feel sick. (plain text)
Light : I feel **a bit** sick.
Medium : I feel **quite** sick.
Strong : I feel **so** sick.
Very strong: I feel **completely** sick.

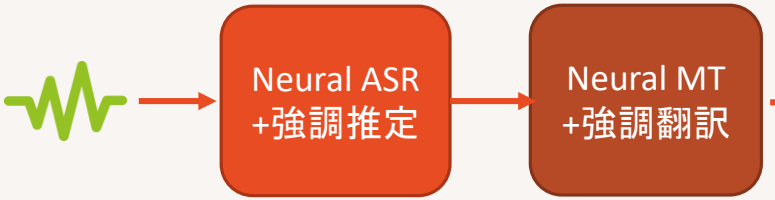
実験結果 – ASR & MT

Test data	Data type		Linguistic Evaluation			Emphasis Evaluation (F-score)	
			ASR	MT			
			WER⇓	Multi-bleu⇑	Sacrebleu⇑	ASR	MT
Synthetic (BTEC)	Previous	No Emphasis	0.72	44.84	47.80	---	---
		Emphasis Separated	0.38	43.99	46.71	95.52	5.41
	Proposed	Emphasis Tags	**0.64**	42.64	45.57	95.38	64.52
		Emphasis 1-Token	0.80	42.31	45.64	**98.51**	70.77
		Emphasis All Token	0.80	**44.84**	**47.71**	96.97	74.63
Natural	Previous	No Emphasis	24.72	31.81	34.68	---	---
		Emphasis Separated	19.50	28.74	31.76	59.66	34.31
	Proposed	Emphasis Tags	24.37	34.12	36.54	**69.23**	36.76
		Emphasis 1-Token	25.45	**36.05**	**38.71**	63.59	49.48
		Emphasis All Token	**22.40**	34.35	36.91	61.60	44.22

•提案法の構造がシンプルで翻訳性能が向上

実験結果 – S2T Task

- Cascade Neural S2T

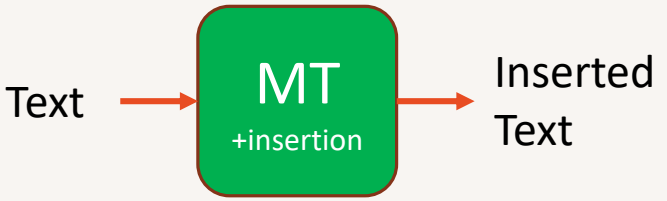


- Direct Neural S2T

Test data	Data type		Linguistic Evaluation				Emphasis Evaluation (F-score)	
			Cascade		Direct			
			Multi-bleu	Sacrebleu	Multi-bleu	Sacrebleu	Cascade	Direct
Synthetic (BTEC)	Previous	No Emphasis	**44.50**	**47.44**	33.88	36.88	---	---
	Proposed	Emphasis Tags	41.45	44.37	29.24	32.57	56.25	55.74
		Emphasis 1-Token	**44.39**	**46.94**	**37.92**	**41.29**	67.69	66.67
		Emphasis All Token	43.86	46.59	35.94	39.37	70.59	55.88
Natural	Previous	No Emphasis	12.32	14.44	2.72	4.16	---	---
	Proposed	Emphasis Tags	7.93	9.16	2.08	2.34	37.74	32.50
		Emphasis 1-Token	12.69	15.25	5.18	6.17	32.04	43.14
		Emphasis All Token	**14.20**	**16.75**	**5.29**	**6.75**	37.43	42.71

- DirectはCascadeよりも性能が低く，先行研究より3-4 bleuスコアが上昇
- Naturalではまだ改善の余地あるが，All Tokenが既存手法よりも高いスコアを出している

実験結果 MT insertion



Test data	System	Emphasis Level					Total
		0	1	2	3	4	
Synthetic (BTEC)	Emphasis Tags	98.18	87.88	93.94	94.94	90.91	96.81
	Emphasis 1-Token	99.20	90.91	100.00	100.00	100.00	98.90
	Emphasis All Token	94.33	93.94	96.97	90.90	93.94	94.25
Natural	Emphasis Tags	84.53	67.03	81.11	80.90	75.28	77.85
	Emphasis 1-Token	93.62	82.98	92.55	91.30	91.30	90.34
	Emphasis All Token	88.04	73.91	79.35	81.11	82.22	80.92

- 各強調レベルにおける挿入正解率を計算
- 1-Tokenの挿入正解率が高い
  - トークンと挿入する単語の対応が取れている

まとめ

- 目的：
  - パラ言語情報をニューラルネットを用いた直接翻訳
  - パラ言語情報から，言語情報への翻訳
- 結果:
  - 提案法の構造がシンプルで翻訳性能が向上
  - S2TにおけるAll Tokenのスコアが既存手法よりも高い
  - 単語挿入に関しては1-Tokenが最も効果的
- 今後の研究:
  - 挿入データ，音声データの主観評価
  - 自然音声のASRの認識向上
  - 他の直接翻訳モデルでの実験
  - TTSの導入

Appendix

抽出した副詞のリスト

- Light (Level 1)
  - 'a little', 'slightly', 'a tad', 'a bit', 'kind of', 'sort of', 'vaguely', 'almost', 'marginally', 'mildly', 'passably', 'faintly', 'a tiny', 'a touch', 'something of', 'just a', 'bit of', 'a slightly', 'more or', 'perceptibly', 'hardly', 'even a', 'imperceptibly', 'a mildly' …
- Medium (Level 2)
  - 'pretty', 'fairly', 'rather', 'quite', 'somewhat', 'relatively', 'moderately', 'reasonably', 'significantly', 'tolerably', 'noticeably', 'considerably', 'uncommonly', 'more than', 'unusually', 'admittedly', 'comparatively', 'visibly', 'partly', 'mostly', 'probably', …
- Strong (Level 3)
  - 'so', 'very', 'really', 'truly', 'super', 'largely', 'definitely', 'clearly', 'much', 'considerably', 'decidedly', 'certainly', 'undeniably', 'positively', 'deeply', 'greatly', 'undoubtedly', 'unquestionably', 'unmistakably', 'so very', 'such', 'vastly', 'prohibitively', 'assuredly', 'mostly' ….
- Very strong (Level 4)
  - 'seriously', 'terribly', 'extremely', 'completely', 'impossibly', 'perfectly', 'supremely', 'obviously', 'shockingly', 'totally', 'awfully', 'exceptionally', 'horribly', 'dreadfully', 'unbearably', 'wildly', 'powerfully', 'entirely', 'amazingly', 'wonderfully', 'hideously' …

自然音声データ

- Natural Japanese Speech
  - これ は ゆるい です 。
  - やっぱり とうきょう の らっしゅ あわー は ひどい です ね 。

Normal		
Light		
Medium		
Strong		
Very strong		

Analysis – ASR Results

- Input speech (Natural Speech)
  - この きかく は せいさん の めん で **まずい** 。(この企画は生産の面でまずい。)
  - “まずい” has emphasis level 3

Data type		Output
Previous	No Emphasis	この ちかく は がくせい が むえん で まずい 。
	Emphasis Separated	この きかく は せいさん の びん で **まずい** 。
		(Emphasis Estimation) 0 0 0 0 0 0 0 **3** 0
Proposed	Emphasis Tags	この0 きたく0 は0 ぜん0 の0 びん0 で0 **まずい3** 。0
	Emphasis 1-Token	この きかく は せーるすまん の びん で **<to3> まずい** 。
	Emphasis All Token	この <to0> ちかく <to0> は <to0> せいかつ <to0> の <to0> びん <to0> で **<to3> まずい** <to0> 。

- Emphasis recognition is mostly correct but language recognition is often different
  - More natural speech data is needed because the percentage of natural speech is small
- When testing natural speech data, sometimes recognize different emphasis level

Analysis – MT Results

- Input text -> Correct output
  - この きかく は せいさん の めん で **まずい** 。-> This scheme is **clumsy** production wise .
  - **Red word** is emphasis part (level 3)

Data type		Output
Previous	Emphasis Separated	This project **is going** to be tasteless
		(Emphasis Translation) 0 0 **3 3** 0 0 0
	No Emphasis	This project lacks control .
Proposed	Emphasis Tags	Something0 is0 **wrong3** with0 this0 plan0 .0
	Emphasis 1-Token	This project is **<to3> tasteless** .
	Emphasis All Token	<to0> This <to0> project <to0> is <to0> completely **<to3> wrong** <to0> .

- Translate depends on train dataset, but emphasis part is translated even if used different word
- Separated depends on Attention and sometimes translate emphasis to wrong emphasis part or other token (ex. <unk>)
- Emphasized “clumsy” word is not in train data
- Including Emphasis translation, proposed is better

Analysis – S2T Task with Synthetic Speech

- Input speech (Synthetic Speech) -> Correct output text
  - **やすい** れすとらん を しょうかい し て いただけ ます か 。
    -> could you recommend an **inexpensive** restaurant ?
  - **Red word** is emphasis part (level 3)

	Data type		Output
Cascade	Previous	No Emphasis	could you recommend a cheap restaurant ?
	Proposed	Emphasis Tags	could0 you0 recommend0 a0 **inexpensive3** restaurant0 ?0
		Emphasis 1-Token	could you recommend a **<to3> cheap** restaurant ?
		Emphasis All Token	<to0> could <to0> you <to0> recommend <to0> an **<to3> inexpensive** <to0> restaurant <to0> ?
Direct	Previous	No Emphasis	could you recommend an inexpensive restaurant ?
	Proposed	Emphasis Tags	could0 you0 recommend0 a0 **cheap3** restaurant0 ?0
		Emphasis 1-Token	could you recommend a **<to3> cheap** restaurant ?
		Emphasis All Token	<to0> could <to0> you <to0> recommend <to0> an **<to3> inexpensive** <to0> restaurant <to0> ?

- When ASR has good results, S2T Task works well

Analysis – S2T Task with Natural Speech

- Input speech (Natural Speech) -> Correct output text
  - この きかく は せいさん の めん で **まずい** 。-> This scheme is **clumsy** production wise .
  - **Red word** is emphasis part (level 3)

	Data type		Output
Cascade	Previous	No Emphasis	this project is better than his age .
	Proposed	Emphasis Tags	this0 machine0 is0 not0 **good3** enough0 .0
		Emphasis 1-Token	this is too **<to3> small** to be getting home .
		Emphasis All Token	<to0> the <to0> connection <to0> of <to0> this <to0> house <to0> is <to0> very <to0> poor <to0> .
Direct	Previous	No Emphasis	this kind of shoe is apt to slip on wet ground .
	Proposed	Emphasis Tags	this0 cloth0 is0 **resistant1** to0 heat0 .0
		Emphasis 1-Token	this north is not **<to3> clean** .
		Emphasis All Token	<to0> this <to0> tea <to0> is **<to3> tepid** <to0> .

- S2T task depends on ASR results
  - When ASR outputs wrong text, MT outputs worse texts …
- We need more natural speech dataset for ASR

Analysis - MT insertion

- Input text (English text) -> Correct output text (English inserted text)
  - this scheme is **clumsy** production wise . -> this scheme is **so** clumsy production wise .
  - Inserted words are 'so', 'very', 'really', 'undeniably', 'positively', 'unquestionably' and etc…

Data type	Output
Emphasis Tags	this scheme is detail production wise.
Emphasis 1-Token	this scheme is **so** clumsy production wise .
Emphasis All Token	this scheme is **so** clumsy production wise .

- Almost good results when using token
  - Guessing there are correspondence between the token and the word to be inserted
- Sometimes other words occurs when using tags
  - Emphasized "clumsy" is not in train data, so it may make to replace other word
    - Ex. "clumsy0" is in train data, but "clumsy1" is not in train data
  - But, some emphasized word works well
    - Ex. this0 is0 loose4 .0 -> this is stupendously loose. (呆気なく,凄まじく)

訓練データセット	
BTEC (強調有り)	184,620
BTEC (強調無し)	297,451
自然音声	92,900
合計	574,971