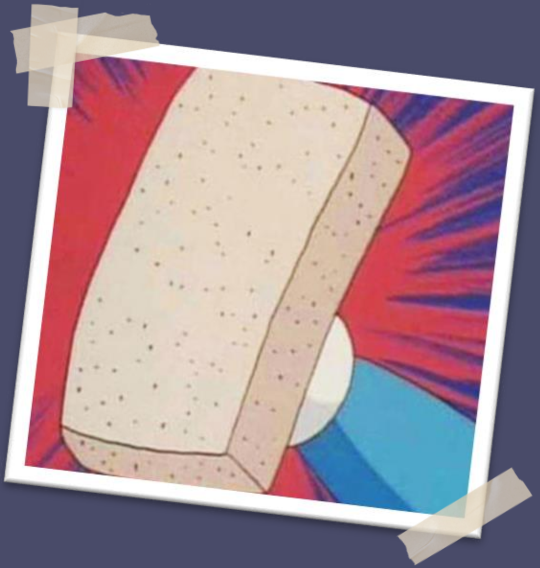
國立成功大學
人工智慧科技碩士學位學程

Improving Chinese-Japanese Neural Machine Translation with Joint Semantic-Phonetic Word Embedding



Student **Shih-Chieh Wang**
Advisor **Paul Horton**

哆啦A夢：大雄的平行西遊記 (52 分處)
ドラえもんのび太のパラレル西遊記 (1988. 03. 12)

Table of Contents

The structure of this presentation

1.1

- **Introduction**
- Method
- Experiment
- Result
- Discussion
- Conclusion

Neural Machine Translation (NMT)

NMT Model	
	RNNs
	Attention-based RNNs
	ConvS2S
	Transformer

Tokeni-zation	
	Word-level
	BytePair Encoding (BPE)
	WordPiece
	SentencePiece

Embedding	
	Word2Vec
	fastText
	ELMo
	BERT

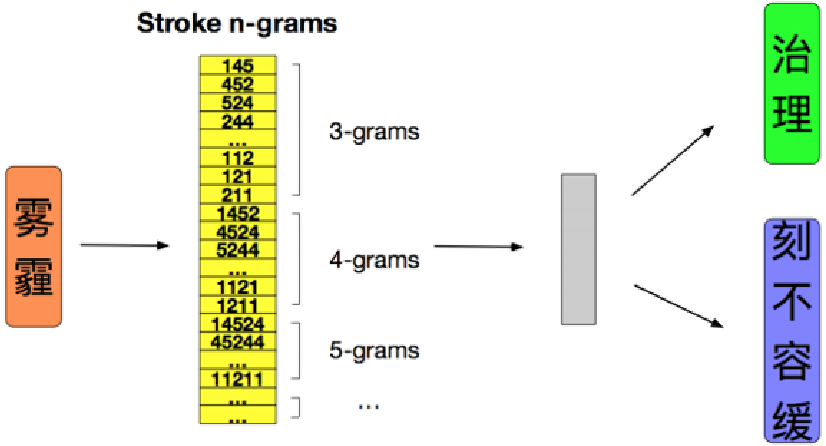
Corpus Problem	
	Back-translation
	Corpus Filtering
	Domain Adaptaion
	Unsupervised Learning

Chinese Hanzi Feature

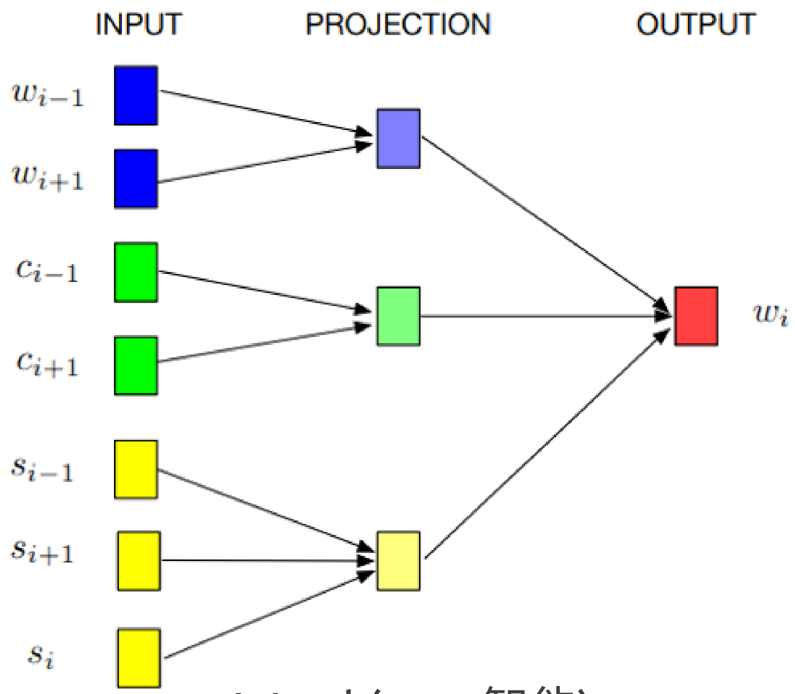
Radicals as an Additional Input Feature [1]

Source sentence	溝幅は１０mm以上が必要と推定した。 (estimated that the groove width should be 10 mm or more.)
Radical features	水巾水一雨艹艹人一力心襾止手宀丿大⺮

CW2VEC [2]



Joint Learning Word Embedding Model [3]



Word (e.g., 智能)

Character (e.g., 智)

Sub-character (e.g., 知, 日)

[1] [Zhang, J. and Matsumoto, T., 2017]
[2] [Bojanowski et al., 2017]
[3] [Yu et al., 2017]

1.3

Phonetic Feature

- **Introduction**
- Method
- Experiment
- Result
- Discussion
- Conclusion

Phonetic Encoding [4]

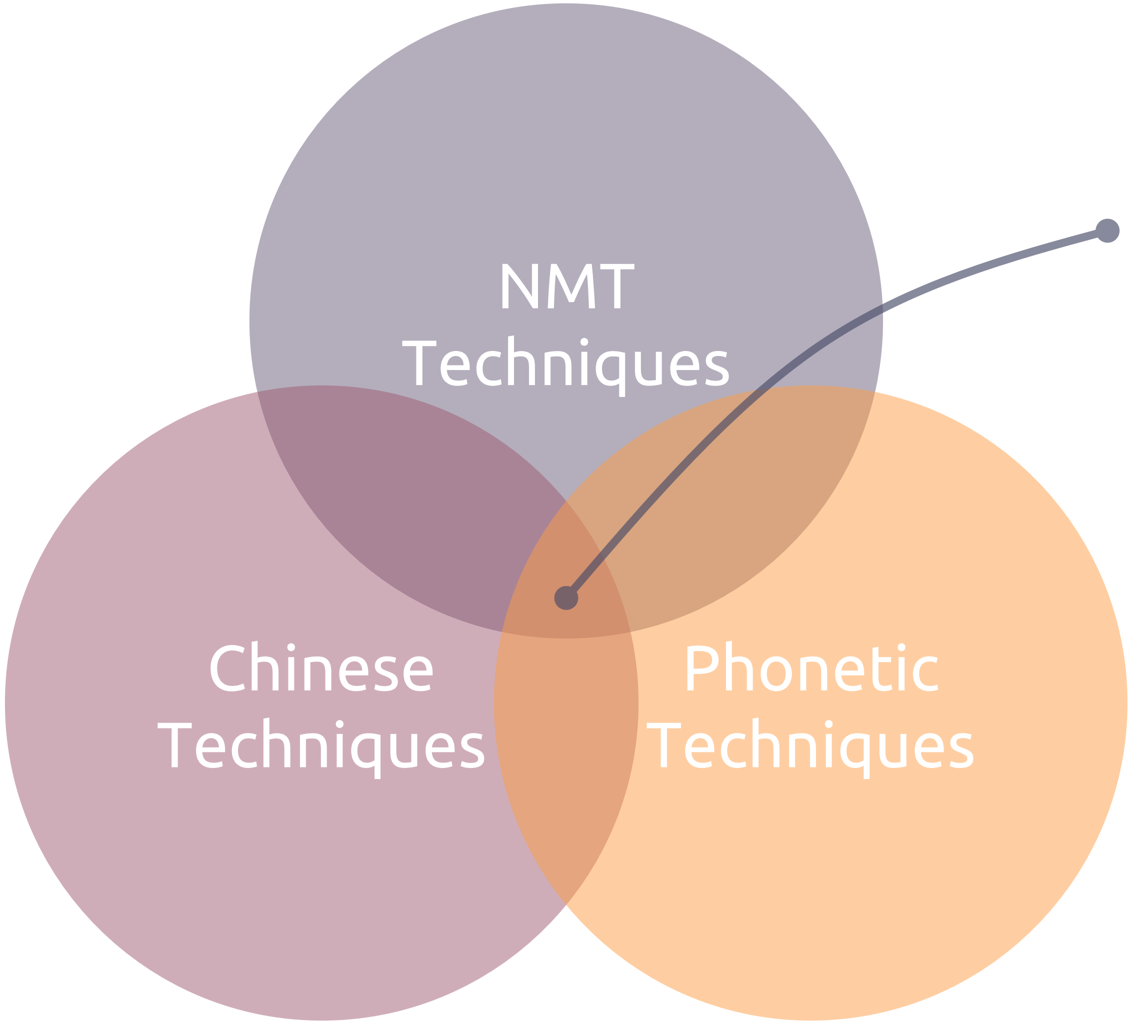
Idea	"Phonetics is a function that groups semantically distinct words. "
Extraction	*Soundex* *NYSIIS* *Metaphone* *Pinyin*
Method	BPE Concatenation

Robuestness to Homophones [5]

Idea	"Phonetic embedding can address the homophone noise problem in corpora."
Extraction	*Pinyin*
Method	$(1-\beta) \times \pi(a) +$ $\beta \times \pi(\psi(a))$

[4] [Khan and Xu, 2019]
[5] [Liu et al., 2019]

Objective



Languages

Chinese & Japanese

Phonetics

Bopomofo & Hiragana Encoding

Embedding

Joint Semantic-Phonetic Embedding

NMT Task

Deep Learning Framework

Analysis

Analyze translation and embedding

Introduction

1. NMT model
2. Chinese model
3. Phonetic model
4. Objective

Method

1. Corpus Filtering
2. Tokenization
3. Phonetics
4. Embedding
5. Model
6. Embedding Analysis

2.1

Corpus Filtering

Introduction

Method

Experiment

Result

Discussion

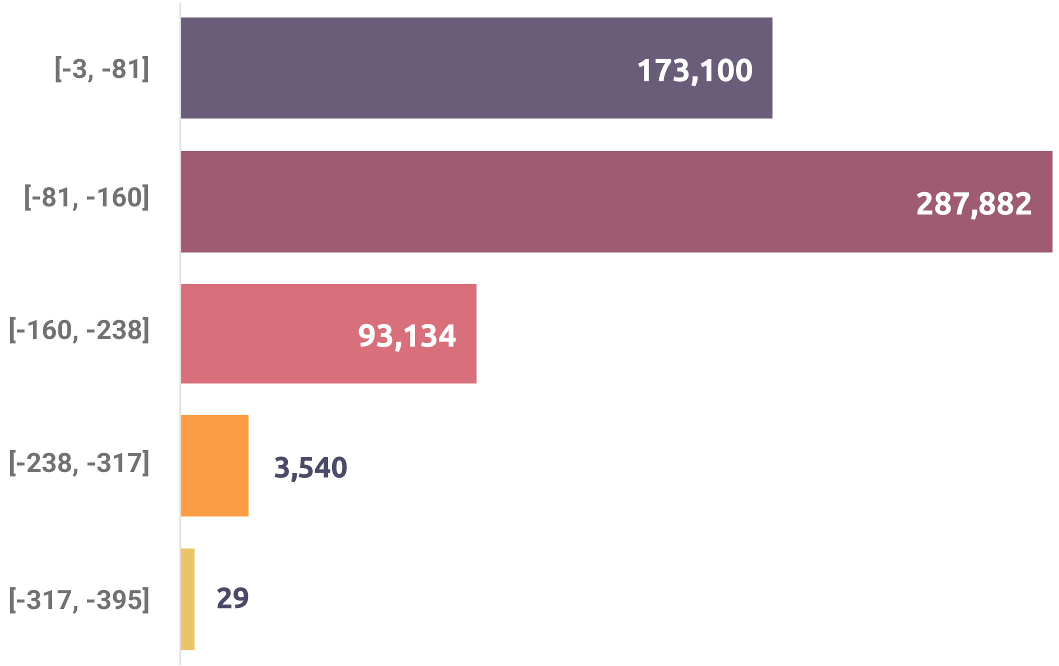
Conclusion

Pre-filtering Rules [6]

1. Too short / long
2. Excessive length ratio
3. Identical sentences
4. Language identifiable
5. Excessive English and numbers
6. One-to-many relationship

Scoring Functions

Alignment scores of ***fast_align*** [7]



[6] [Koehn et al., 2018]
[7] [Dyer et al., 2013]

2.2

Introduction

Method

Experiment

Result

Discussion

Conclusion

Tokenization

HuggingFace Tokenizers [8]		株式会社　ＫＡＤＯＫＡＷＡ
Normalizer	Normalize the texts	株式会社 KADOKAWA
Pre-tokenizer	Split input string with rules	[株式, 会社, KA, DO, KA, WA]
Model	Convert tokens into ids with WordPiece, BPE, etc	[株式会社, KADO, KAWA] [1123, 54, 78]
Post-Processor	Insert special tokens	[0, 1123, 54, 78, 1]
Decoder	Reverse the ids to tokens	[<BOS>, 株式会社, KADO, KAWA, <EOS>]

[8] github.com/huggingface/tokenizers

2.2

Tokenization

HuggingFace Tokenizers

Normalizer | Pre-tokenizer | Model | Post-Processor | Decoder

SentencePiece [9]

Same Unicode format

meta symbol " _ "

[_平成, _15, 年, 进行的研究, 内容, 如下, _。]

[_平成, _15, _年度, に行, なった, 研究, 内容, は次の通りである, _。]

Jieba [10]

Prefix dictionary

Longest path in DAG

Hidden Markov Model

[平成, 15, 年, 进行, 的, 研究, 内容, 如下, 。]

Janome [11]

Minimal Acyclic Subsequential Transducer (MAST)

Viterbi Algorithm

[平成, 15, 年度, に, 行なっ, た, 研究, 内容, は, 次, の, 通り, で, ある, 。]

[9] github.com/google/sentencepiece
[10] github.com/fxsjy/jieba
[11] mocobeta.github.io/janome

2.3

- Introduction
- **Method**
- Experiment
- Result
- Discussion
- Conclusion

Phonetic Extraction

DragonMapper [12]

Chinese to Bopomofo

CC-CEDICT [13] and *Unihan* [14] database

Applied to tokenized sentences

PyKakasi [15]

Japanese to Hiragana

Based on *kakasi* [16] library

SKK dictionary [17] *and UniDic* [18]

Applied to tokenized sentences

长大很快乐，音乐很长久

[长大, 很, 快乐, ,, 音乐, 很, 长久]

[ㄓㄤˇㄉㄚˋ, ㄏㄣˇ, ㄎㄨㄞˋㄌㄜˋ, ,, ㄧㄣㄩㄝˋ, ㄏㄣˇ, ㄔㄤˊㄐㄧㄡˇ]

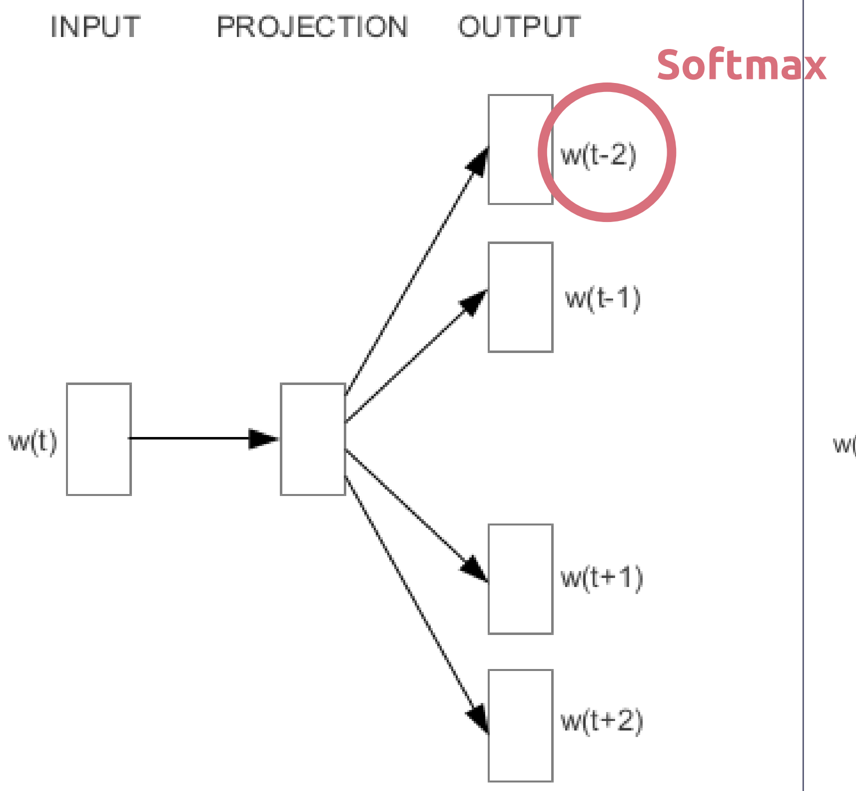
一生，芝生で生ビールを飲む

[一生, ,, 芝生, で, 生, ビール, を, 飲む]

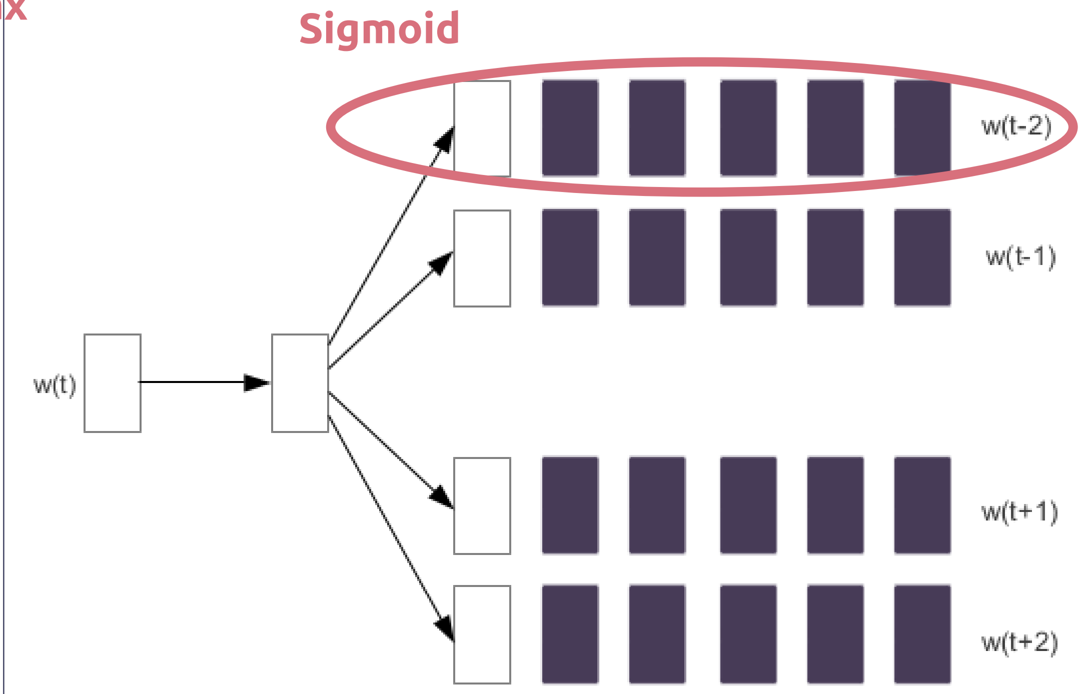
[いっしょう, ,, しばふ, で, なま, びーる, を, のむ]

[12] github.com/tsroten/dragonmapper
[13] cc-cedict.org/
[14] unicode.org/charts/unihan.html
[15] github.com/miurahr/pykakasi
[16] kakasi.namazu.org/index.html.en
[17] github.com/skk-dev/dict
[18] unidic.ninjal.ac.jp/

Embedding



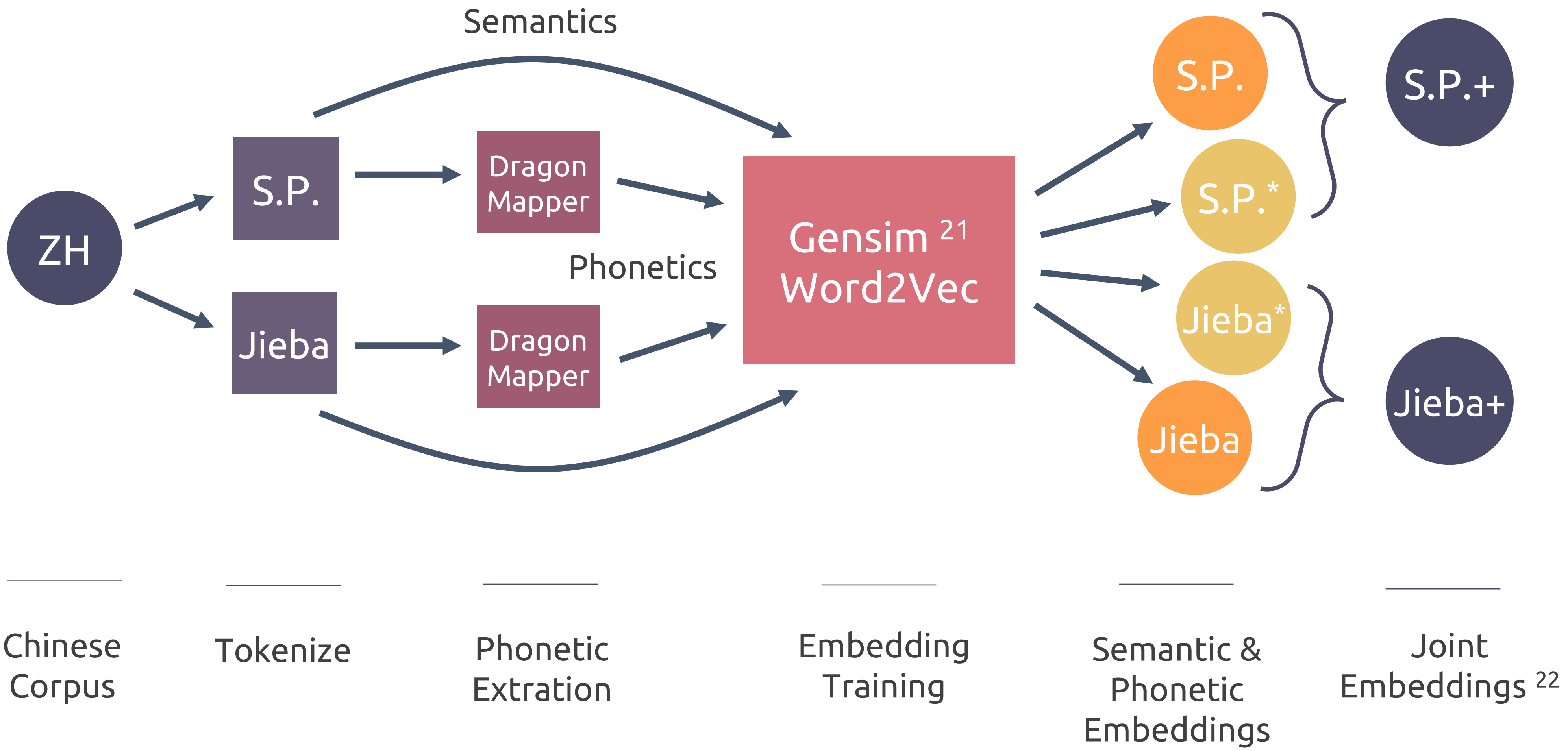
Word2Vec Skip-gram [19]



Negative Sampling [20]

[19] [Mikolov et al., 2013a]
[20] [Mikolov et al., 2013b]

Embedding



[21] radimrehurek.com/gensim/index.html
[22] [Coates and Bollegala, 2018]

Attention-based Bi-GRU Seq2Seq [23]

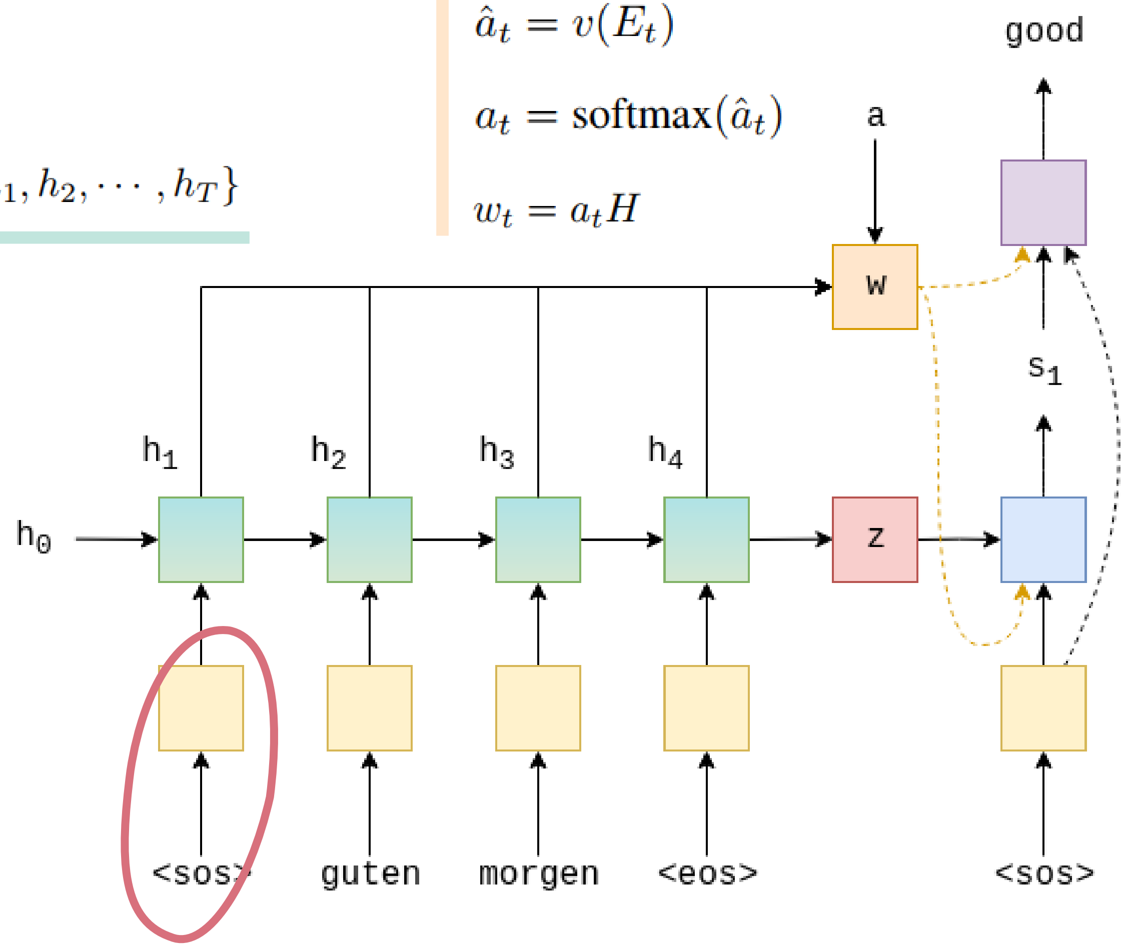
$$E_t = \tanh(\mathrm{attn}(s_{t-1}, H))$$

$$\hat{a}_t = v(E_t)$$

$$a_t = \mathrm{softmax}(\hat{a}_t)$$

$$w_t = a_t H$$

$$H = \{h_1, h_2, \cdots, h_T\}$$



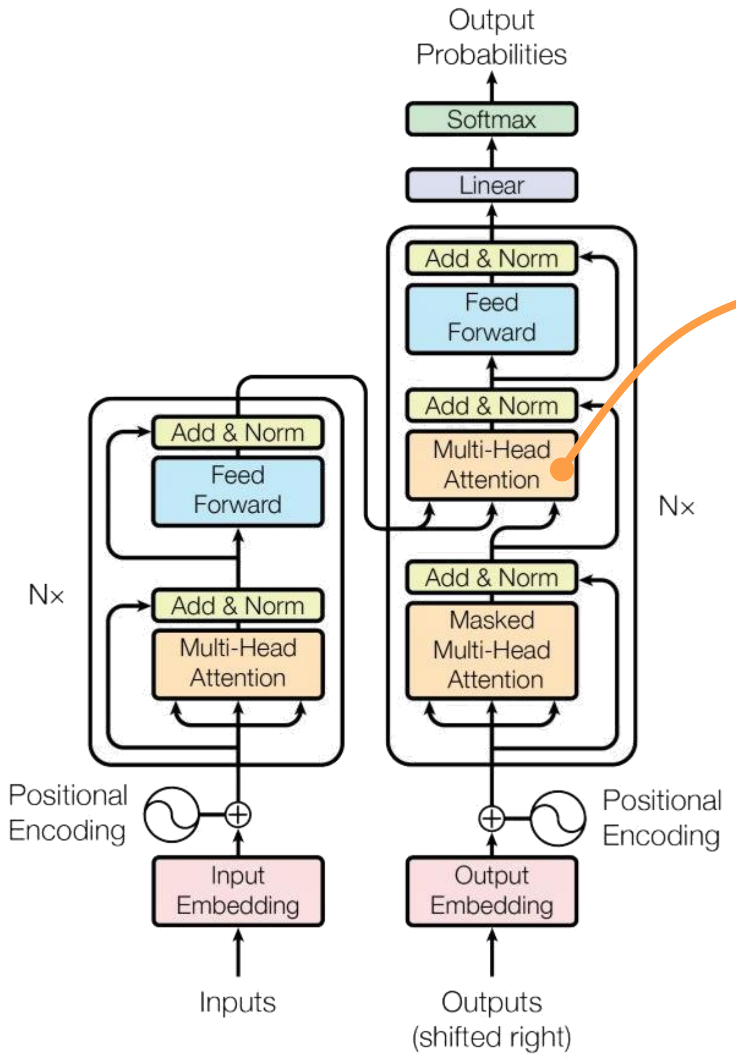
$$\hat{y}_{t+1} = f(d(y_t), w_t, s_t)$$

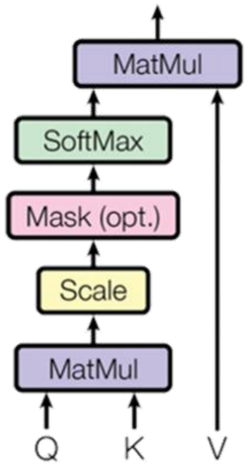
$$s_t = \mathrm{DecoderGRU}(d(y_t), w_t, s_{t-1})$$

$$z = \tanh(g(\mathrm{cat}(\overrightarrow{z}, \overleftarrow{z}))) = s_0$$

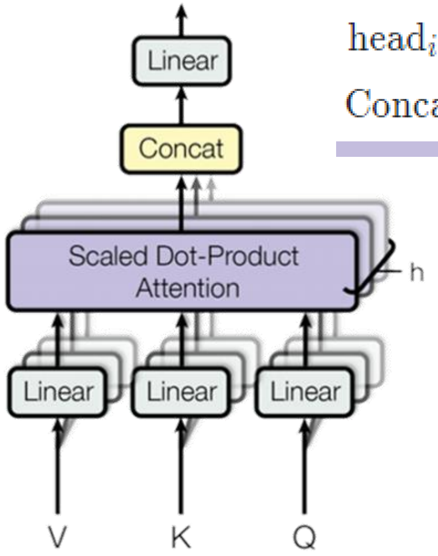
[23] [Bahdanau et al., 2014]

Transformer [24]





$$\mathrm{Attention}(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



$$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V)$$
$$\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)W^O$$

[24] [Vaswani et al., 2017]

2.6

Embedding Analysis

Analogy Reasoning

Apply arithmetic operations to reasoning
$a{:}a^* = b{:}b^*$ $(a - a^* + b = b^*)$

a= 東京，a*= 日本，b= 台北，b*= ?

Outlier Detection

Identify the semantically anomalous word

S = {人，貓，狗，花}，outlier= ?

Word Similarity

Evaluate the closeness of synonyms by calculating the cosine similarity

$\text{distance}^{\text{joint}}(X, X^*) \leqq \text{distance}^{\text{semantic}}(X, X^*)$?

Homonym & Heteronym

Reduces the impact of possible noise from homonyms and increase the semantics of heteronyms

$\text{corr}^{\text{joint}}(Y, Y^*) \geqq \text{corr}^{\text{semantic}}(Y, Y^*)$?

$\text{distance}^{\text{joint}}(Z, Z^*) \geqq \text{distance}^{\text{semantic}}(Z, Z^*)$?

Method

1. Corpus Filtering
2. Tokenization
3. Phonetics
4. Embedding
5. Model
6. Embedding Analysis

Experiment

1. Dataset
2. Parameter
3. Metric

3.1 Dataset

Asian Scientific Paper Excerpt Corpus Japanese-Chinese (ASPEC-JC) [25]

Translating the **excerpts** of Japanese **scientific papers** into Chinese.

Derived from the *Japan Science and Technology Agency* (JST) or the *Japan Science and Technology Information Aggregator, Electronic* (J-STAGE).

Train	Validation	Test
672,315	2,090	2,107

本模型适用于报告有关化学物质的堆积分布和堆积物的物理性状的详细数据的宍道湖及中海水系。

本モデルを,化学物質の堆積分布や堆積物の物理的性状に関する詳細なデータが報告されている宍道湖・中海水系に適用した。

阴离子HCO_3¯—、NO_3¯—、Cl¯—、SO_4¯(2—)对EE2光降解反应有抑制作用.

陰イオンのHCO_3—、NO_3—、Cl—とSO_4(2—)は17α—エチニルエストラジオールの光分解を抑制させた。

[25] [Nakazawa et al., 2016]

Parameter

Parameter	RNN	Transformer
sample_size	50,000	50,000
dictionary_size	32,000	32,000
embedding_dim	300	300
dropout	0.1	0.1
hidden_dim	512	512
learning rate	7e-4	5e-4
precision	16	32
layers	1	3
attention heads	-	6
freeze_epochs [26]	3	1

[26] [Kirkpatrick et al., 2017]

3.3

BLEU (Bilingual Evaluation Understudy) [27]

$$\text{BLEU} = \underbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}_{\text{brevity penalty}} \underbrace{\left(\prod_{i=1}^{4} precision_i\right)^{1/4}}_{\text{n-gram overlap}}$$

0 100

Reference 実践場面における質的研究法

Output 実践場面における質研究法

Unigram	11/12 = 0.917
Bigram	10/11 = 0.910
Trigram	8/10 = 0.8
4-gram	6/9 = 0.667
overlap	$(0.917 \times 0.91 \times 0.8 \times 0.667)^{1/4}$

[27] [Papineni et al., 2002]

Experiment

1. Dataset
2. Parameter
3. Metric

Result

1. Performance
2. Best Model

4.1

Introduction

Method

Experiment

Result

Discussion

Conclusion

BLEU Scores based on Sampled Data

Attention-based Bi-GRU Model

Tokenization	Baseline	Semantic	Phonetic	Joint
SentencePiece	21.63	21.66	21.32	**22.33**
Jieba + Janome	25.16	26.71	26.18	**27.05**

Transformer

Tokenization	Baseline	Semantic	Phonetic	Joint
SentencePiece	24.32	25.72	23.48	**26.44**
Jieba + Janome	29.31	31.23	30.90	**32.48**

Best Model

Transformer + Jieba + Janome

Dataset	Baseline	Semantic	Phonetic	Joint
Sampled	29.31	31.23	30.90	**32.48**
Complete & Filtered	52.78	52.83	53.04	**53.13**

Workshop on Asian Translation (WAT) 2020

The baseline system [28] of ASPEC-JC zh-ja task of WAT 2020 is designed using *OpenNMT* [29], BPE, and attention mechanisms.

zh-ja NMT task	Juman 7.0	KyTea 0.4.6	Mecab 0.996
Baseline System	46.87	47.30	47.00

[28] [Nakazawa et al., 2020]
[29] github.com/OpenNMT/OpenNMT

Result

1. Performance

2. Best Model

Discussion

1. Case Study

2. Embedding Analysis
    1. Analogy Reasoning
    2. Outlier Detection
    3. Word Similarity
    4. Homonym & Heteronym

Case Study of Translation Results

Retent correct words

Source	从背景知识B和观测结果O中获得行动规则的集合γ的集合H
Target	背景知識Bと観測結果Oより行動規則の集合γの集合を獲得する
Semantic	背景背景知識Bと観測結果Oから行動ルールの集合γの集合H獲得する
Joint	背景知識Bと観測結果Oから行動規則の集合γの集合H獲得する

Preserve English acronyms

Source	另一方面,GUIServer和GUIClient是用Java语言进行安装
Semantic	一方,GUISververやGUIIClientはJava言語で実装した
Joint	一方,GUIServerとGUIClientはJava言語で実装した

Utilize English loanwords

Source	在可能范围里保证变化丰富
Semantic	豊かな範囲で変化 (change; alternation) が豊かになる
Joint	可能な範囲でバリエーション (variation) が豊かになる

Select proper words

Source	微小粒子测量装置的比较试验
Semantic	微小粒子計測 (instrumentation) 装置の比較実験 (experiment)
Joint	微小粒子測定装置 (measuring instrument) の比較試験 (test)

5.2.1

Embedding Analysis - Analogy Reasoning

Chinese

Input			Ouput (B*)	
A	A*	B	Semantic	Joint
东京	日本	北京	中国 (p=0.492)	中国 (**p=0.531**)
长期	三年	短期	一年 (p=0.379)	两周 (**p=0.387**)
进口	买入	出口	卖出 (p=0.364)	卖出 (**p=0.442**)
Japanese				
男性	女性	父親	母親 (p=0.487)	母親 (**p=0.508**)
長期	年	短期	月 (p=0.550)	月 (**p=0.570**)
左右	前後	水平	垂直 (**p=0.432**)	垂直 (p=0.402)

5.2.2

Embedding Analysis - Outlier Detection

Chinese

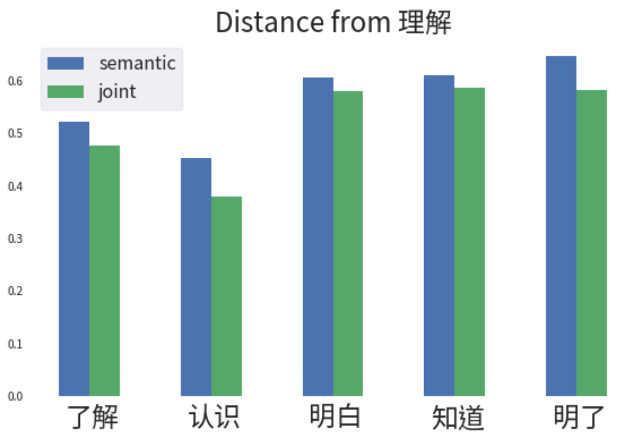
Word Cluster	Ouput (B*)	
	Semantic	Joint
[鸟, 狗 , 猫 , 花]	花	花
[可行, 不行, 行, 可以]	可以	不行
[广岛, 名古屋, 爱知, 上海]	上海	上海

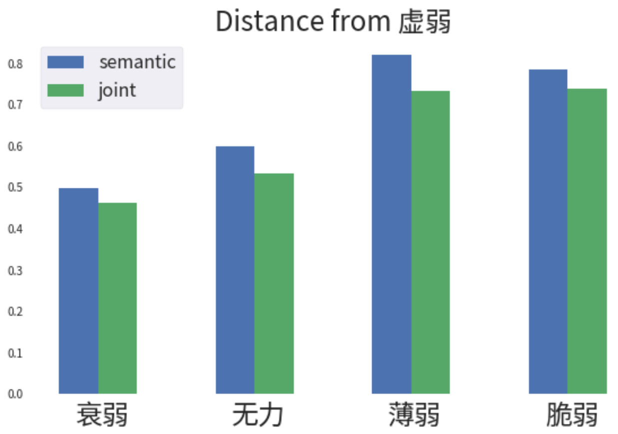
Japanese

[生み, 創造, 作る, 破壊]	破壊	破壊
[普通, 一般, 通常, 異常]	異常	異常
[平成, 昭和, 大正, 明治, 京都]	京都	京都

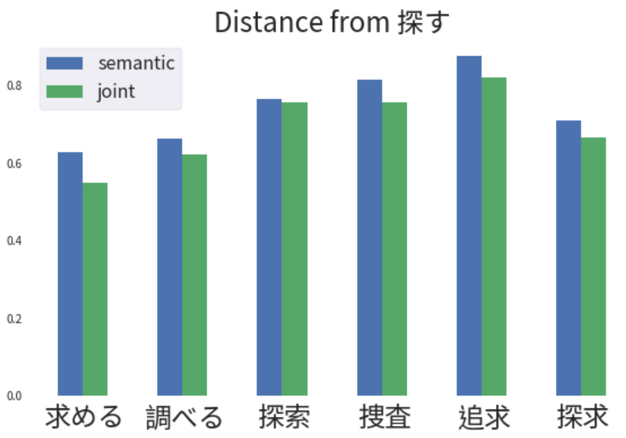
Embedding Analysis - Word Similarity

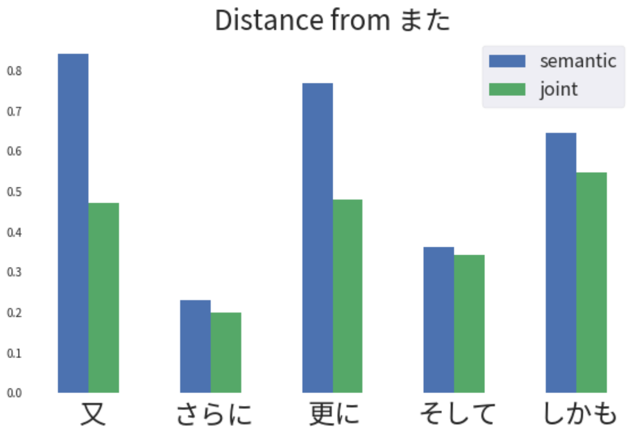
Chinese





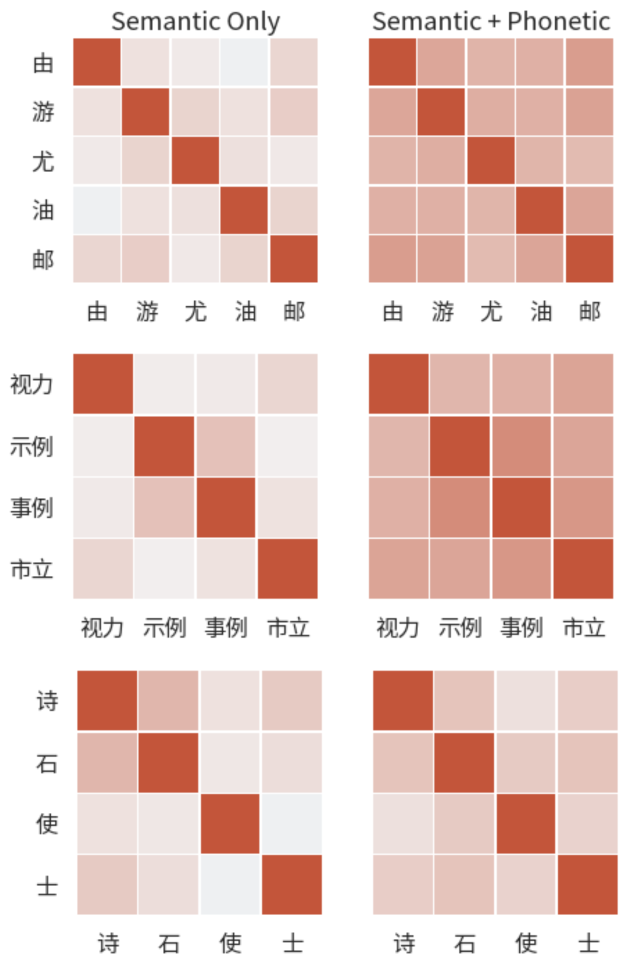
Japanese



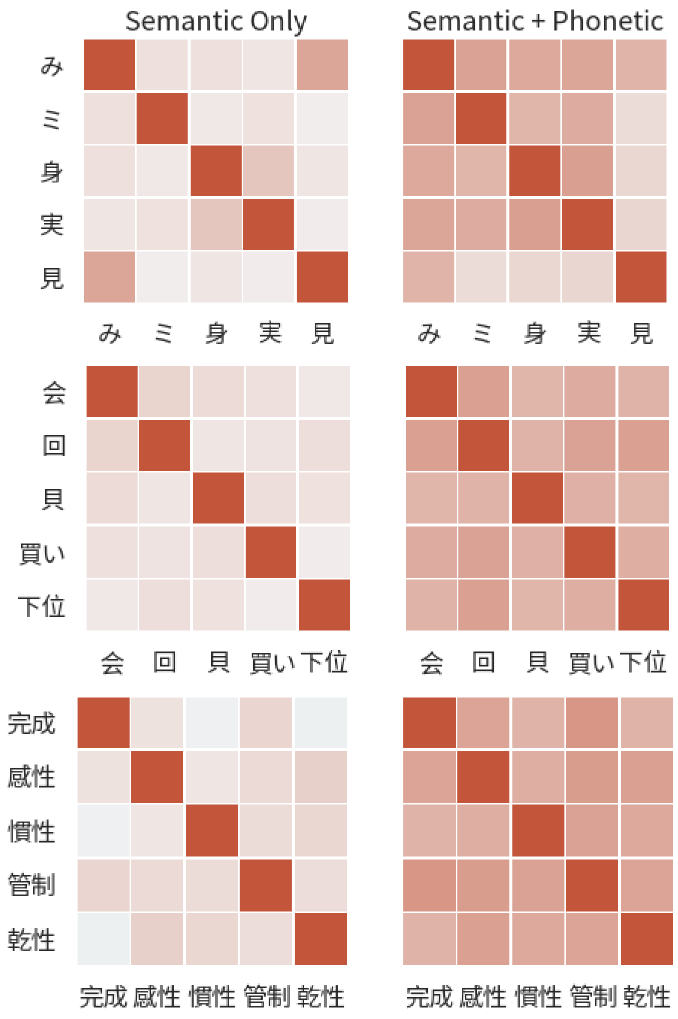


Embedding Analysis - Homonyms

Chinese



Japanese



Embedding Analysis - Heteronyms

Chinese

A	B	Similarity Distance	
		Semantic	Joint
長 (ㄔㄤˊ) 度	長 (ㄓㄤˇ) 大	0.826	**0.853**
樂 (ㄌㄜˋ) 趣	音樂 (ㄩㄝˋ)	0.636	**0.682**
中 (ㄓㄨㄥ) 午	中 (ㄓㄨㄥˋ) 毒	0.842	**0.866**

Japanese

生 (なま)	一生 (しょう)	0.879	**0.889**
生 (なま)	生 (う) む	0.830	**0.839**
生 (なま)	生 (き) 地	0.769	**0.867**

Discussion

1. Case Study
2. Embedding Analysis

Conclusion

1. Contribution
2. Future Work

6.1

Contribution

New Phonetic Encodings

Utilizing Bopomofo and Hiragana as phonetic information in Chinese and Japanese respectively

Joint Semantic-Phonetic Word Embedding

Training and combining word embeddings effectively with a small corpus

Better Translation Results

Obtaining higher BLEU scores based on joint embedding, corpus filtering, and other techniques

Analysis of Translations and Embeddings

Analyzing the advantages of applying phonetic information to embeddings and NMT systems

Future Work

Suggestion #1

Using any basic attempts to refine the research. For example, using other tokenization methods, NMT models, and phonetic extraction methods.

Suggestion #2

Training an embedding by combining phonetic information with Chinese characters and various sub-character features.

Suggestion #3

Using ELMo or BERT as the base embedding framework. Consider training a semantic model and a phonetic model and combining the two.

References

References

1 [Zhang, J. and Matsumoto, T., 2017] Zhang, J., & Matsumoto, T. (2017, December). Improving character-level japanese-chinese neural machine translation with radicals as an additional input feature. In 2017 International Conference on Asian Language Processing (IALP) (pp. 172-175). IEEE.

2 [Bojanowski et al., 2017] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

3 [Yu et al., 2017] Yu, J., Jian, X., Xin, H., and Song, Y. (2017). Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 286–291.

4 [Khan and Xu, 2019] Khan, A. R. and Xu, J. (2019). Diversity by phonetics and its application in neural machine translation. arXiv preprint arXiv:1911.04292

5 [Liu et al., 2019] Liu, H., Ma, M., Huang, L., Xiong, H., and He, Z. (2019). Robust neural machine translation with joint textual and phonetic embedding. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3044–3049, Florence, Italy. Association for Computational Linguistics

6 [Koehn et al., 2018] Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pages 726–739.

7 [Dyer et al., 2013] Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648.

19 [Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

20 [Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. arXiv preprint arXiv:1310.4546.

22 [Coates and Bollegala, 2018] Coates, J. and Bollegala, D. (2018). Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings. In Proceedings of the 2018 Conferenceof the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 194–198, New Orleans, Louisiana. Association for Computational Linguistics.

23 [Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

24 [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

25 [Nakazawa et al., 2016] Nakazawa, T., Yaguchi, M., Uchimoto, K., Utiyama, M., Sumita, E., Kurohashi, S., and Isahara, H. (2016). ASPEC: Asian scientific paper excerpt corpus. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2204–2208, Portoroz, Slovenia. European Language Resources Association (ELRA).

26 [Kirkpatrick et al., 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526.

27 [Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

28 [Nakazawa et al., 2020] Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., Goto, I., Pa, W. P., Kunchukuttan, A., Parida, S., et al. (2020). Overview of the 7th workshop on asian translation. In Proceedings of the 7th Workshop on Asian Translation, pages 1–44.

THANK YOU !

Improving Chinese-Japanese Neural Machine Translation with Joint Semantic-Phonetic Word Embedding

Shih-Chieh Wang, Paul Horton
National Cheng Kung University
Master Degree Program on Artificial Intelligence

Backup: Case Study - Retention of correct words

Source	使用有机溶剂的溶剂提取法作为物质的分离手段被广泛利用。
Target	有機溶媒を用いた溶媒抽出法は物質の分離手段として広く利用されている。
Semantic	有機溶媒を用いた溶媒抽出法は物質の分離手段として広く用いされている。
Joint	有機溶媒を用いた溶媒抽出法は物質の分離手段として広く利用されている。

用いられ

Source	LearnAR从背景知识B和观测结果O中获得行动规则的集合γ的集合H。
Target	LearnARは,背景知識Bと観測結果Oより行動規則の集合γの集合を獲得する。
Semantic	LearnARは背景背景知識Bと観測結果Oから行動ルールの集合γの集合H獲得する。
Joint	LearnARは,背景知識Bと観測結果Oから行動規則の集合γの集合H獲得する。

Backup: Case Study - Preservation of English acronyms

Source	RMCPRGIGCenter和RMCPRGIGTransceiver使用C语言安装,在二者间的通信中和RemoteGIG同样使用TCP/IP上的RMCP。
Semantic	RMCPRGIGCenterとRMCPRGIGTransceiはC言語を実装し,両者間通信通信でTCPTCPと同様にTCP/IPででRMCPを用いるて。
Joint	RMCPRGIGCenterとRMCPRGIGTransceiverをC言語を実装し,両者間の通信においてはRemoteGIGと同様にTCP/IP上でRMCPを用いた。

Source	另一方面,GUIServer和GUIClient是用Java语言进行安装。
Semantic	一方,GUISververやGUIIClientはJava言語で実装した。
Joint	一方,GUIServerとGUIClientはJava言語で実装した

Backup: Case Study - Utilization of English loanwords

Source	构成这些乐曲时在可能范围里保证变化丰富。
Target	これらの楽曲は,可能な範囲でバリエーションが豊かになるように構成されている。
Semantic	これらの楽曲を,豊かな範囲で変化が豊かになる.にしされている。
Joint	これらの楽曲を,可能な範囲でバリエーションが豊かになるように構成されている。

Variation

Source	铝焊接烟气的特性
Target	アルミニウム溶接ヒュームの物性
Semantic	アルミニウム溶接煙の特性
Joint	アルミニウム溶接ヒュームの特性

Fume

Backup: Case Study - More appropriate word selection

Source	微小粒子测量装置的比较试验。
Target	微小粒子測定装置の比較試験
Semantic	微小粒子計測装置の比較実験
Joint	微小粒子測定装置の比較試験

experiment vs. test

Source	胃运动机能和NERD的病状的关连性的相关讨论
Target	胃運動機能とNERDの病態との関連性に関する検討
Semantic	胃運動機能とNERDの病態のの関連性に関する議論
Joint	胃運動機能とNERDの病態との関連性に関する検討

debate vs. consideration

Backup: Model Train Loss & Valid Loss

Baseline

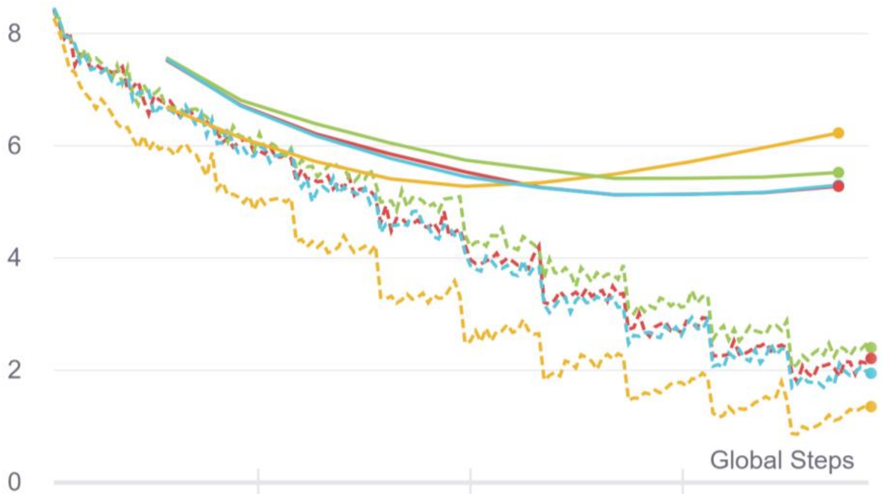
Semantic

Phonetic

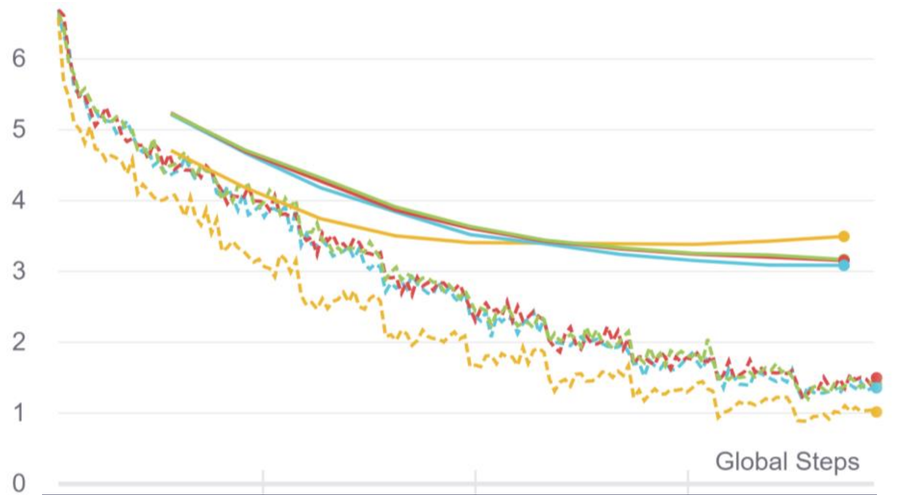
Joint

RNN & SentencePiece

RNN & Jieba + Janome



Transformer & SentencePiece



Transformer & Jieba + Janome