言語処理分野の最前線２
日本語を対象とした統計的機械翻訳の進展

Graham Neubig

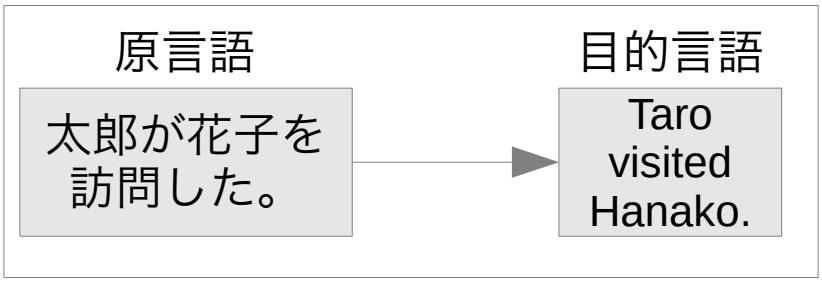
奈良先端科学技術大学院大学　助教

日本語を対象とした統計的機械翻訳の進展

Graham Neubig
奈良先端科学技術大学院大学 (NAIST)
2015-2-24

機械翻訳

- 原言語から目的言語へと自動的に翻訳



- 近年に著しい発展と実用化

Google translate

excite

統計的機械翻訳 [Brown+ 93]

- 大量の学習データからシステムを自動的に学習

対訳文

太郎が花子を訪問した。
Taro visited Hanako.

花子にプレゼントを渡した。
He gave Hanako a present.

...

モデル

翻訳モデル

並べ替えモデル

言語モデル

2010 年からの進展

1)並べ替えの問題を克服
事前並べ替え、統語ベース統計翻訳

2)原言語に含まれない単語の推定
助詞の挿入、ゼロ照応解析

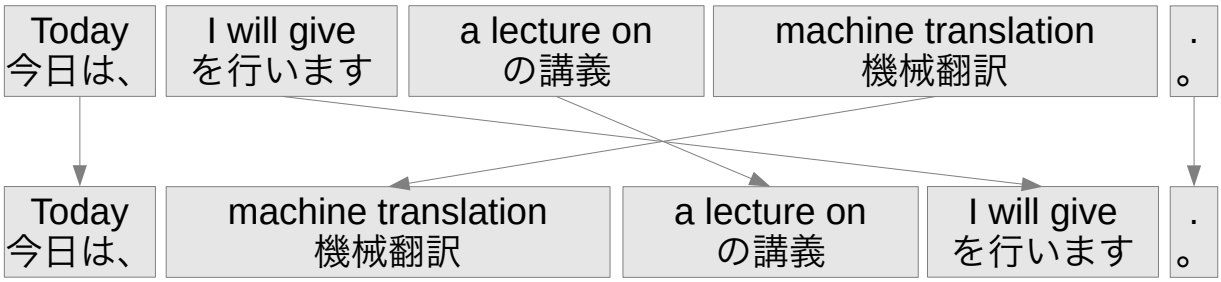
3)確率推定の高度化
ニューラルネットモデル

1) 並べ替えの問題の克服

フレーズベース機械翻訳 [Koehn+ 03]

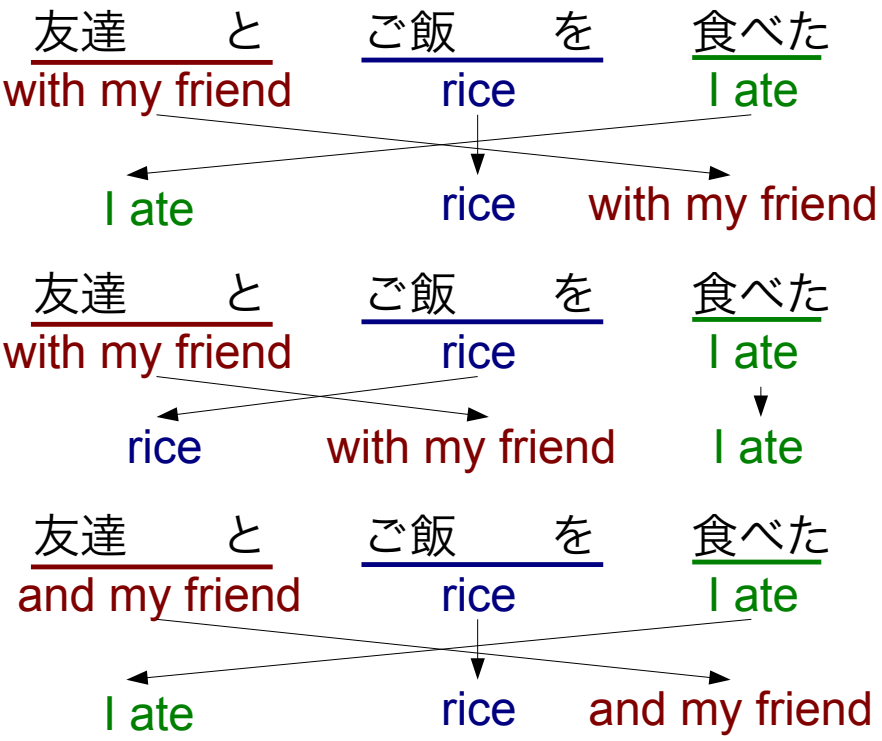
- 文をフレーズ（単語列）ごとに翻訳して、並べ替え

Today I will give a lecture on machine translation .

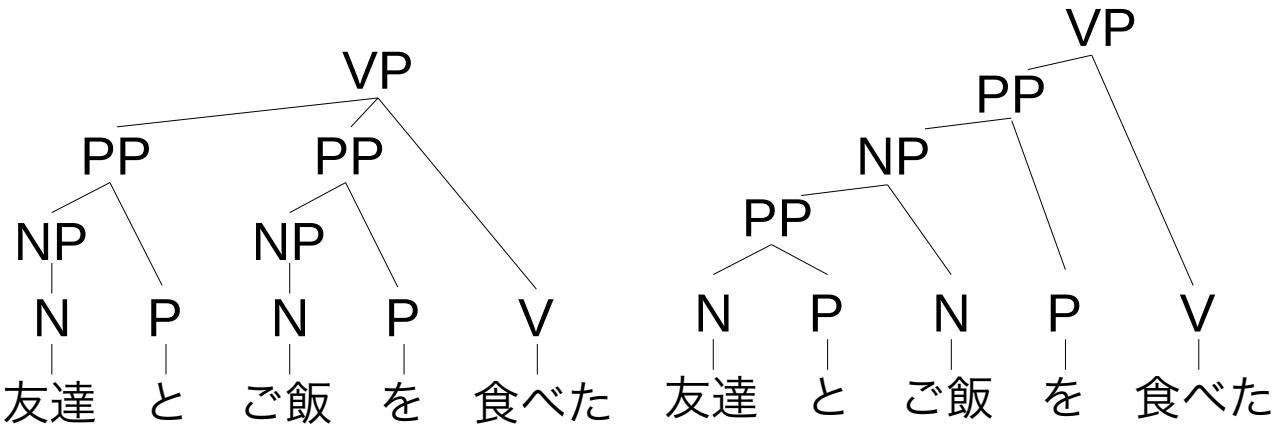


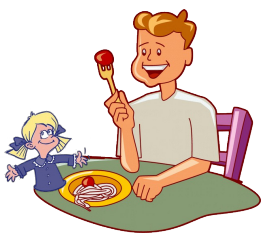
今日は、機械翻訳の講義を行います。

フレーズベース翻訳と並べ替え



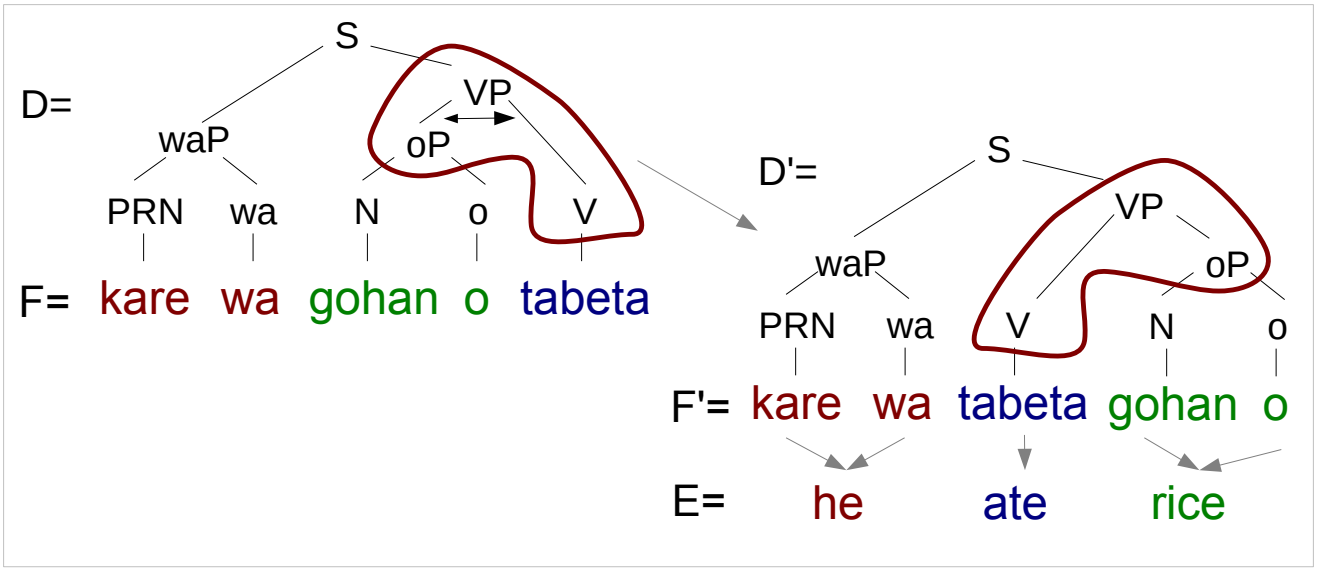
構文解析



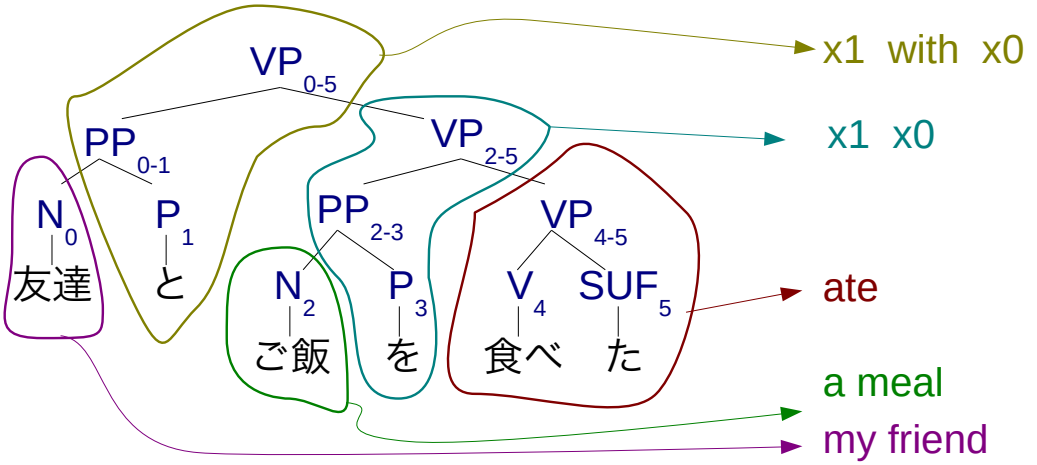


事前並べ替え [Xia+ 04, Isozaki+ 10]

- 原言語の構文木に対して並べ替えルールを定義



Tree-to-String 翻訳 [Liu+ 06, Neubig+ 14]



日本語における Tree-to-String 翻訳実験

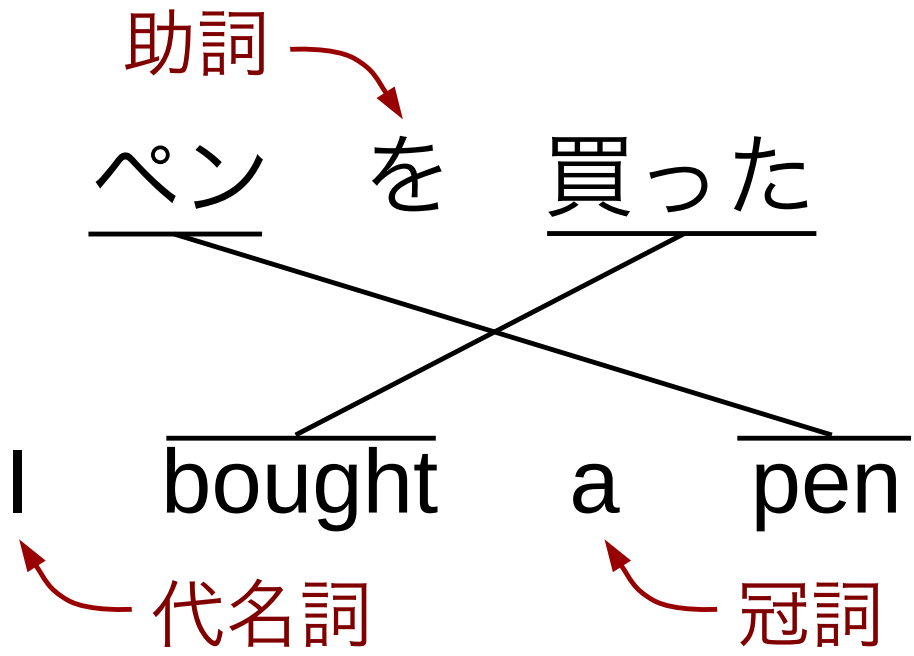
- 少しの工夫で既存手法を大幅に上回る [Neubig+14]

入力	In equipment which generates CT by pressure waves, it was confirmed that development hours of CT bubbles increase in proportion to the 0.5th power of input power of pressure waves.
PB	気泡を生成することを確認した圧力波によりｃｔ装置において，０．５乗に比例して増加し，圧力波の入力電力のｃｔの開発時間。
T2S	圧力波によるｃｔを発生する装置において，ｃｔ気泡の開発時間が圧力波の入力パワーの０．５乗に比例することを確認した。
正解	圧力波によりｃｔを発生させる装置では圧力波の入力パワーの０．５乗に比例してｃｔ気泡の発達時間が増大することを確認した。

- オープンソース公開
  http://www.phontron.com/travatar

2) 原言語に含まれない単語の推定

日英のヌル対応



13

助詞・冠詞の前処理 [Isozaki+ 10]

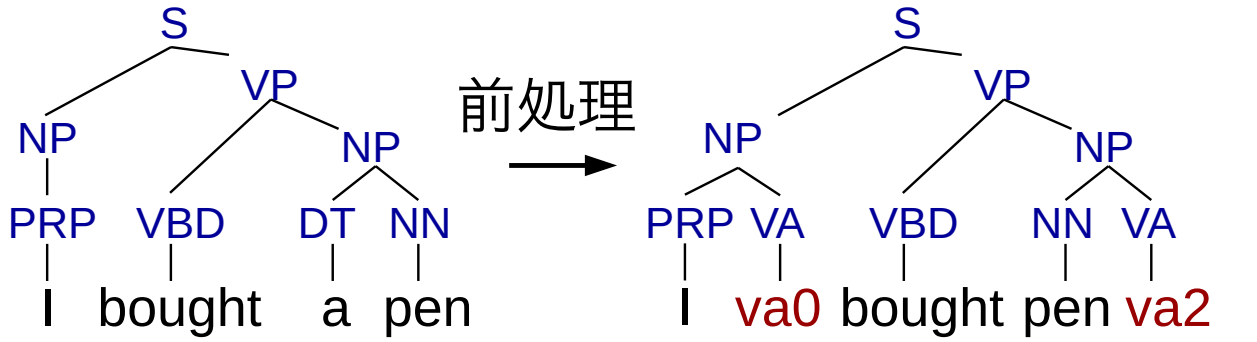
I bought a pen

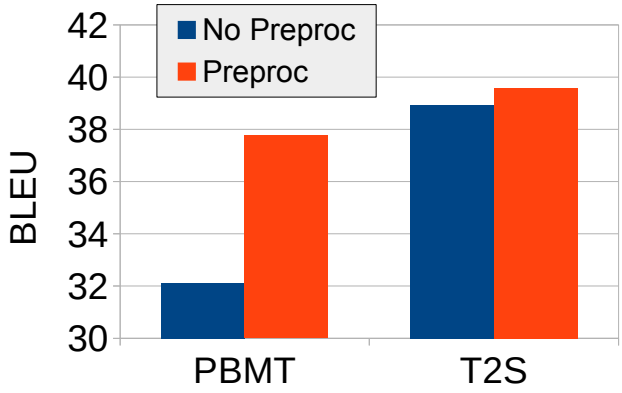
I va0_ pen va2_ bought

私 は ペン を 買った

14

統語ベース翻訳における前処理 [Hatakoshi+ 14]





統語ベース翻訳

私 は 本 を 買った

主語の補完・深層格解析 [Kudo+ 14]

- 同時に主語と深層格を解析し、事前並べ替えに利用

今日は {d=other} 酒が {d=obj} 飲める {v=potential, z=I}
I can drink alcohol today

ニュースが {d=subj} 伝えられた {v=passive, z=already_exist}
The news was conveyed

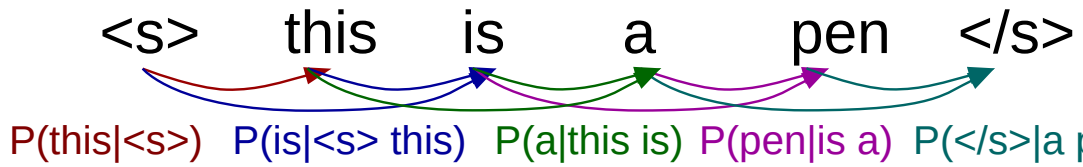
パスタは {d=obj} 食べましたか {v=active, z=you}
Did you eat pasta?

あなたは {d=subj} 食べましたか {v=active, z=already_exist}
Did you eat?

3) 確率モデルの高度化

n-gram 言語モデルとその弱点

- 次の単語の確率を前の n-1 単語で推定



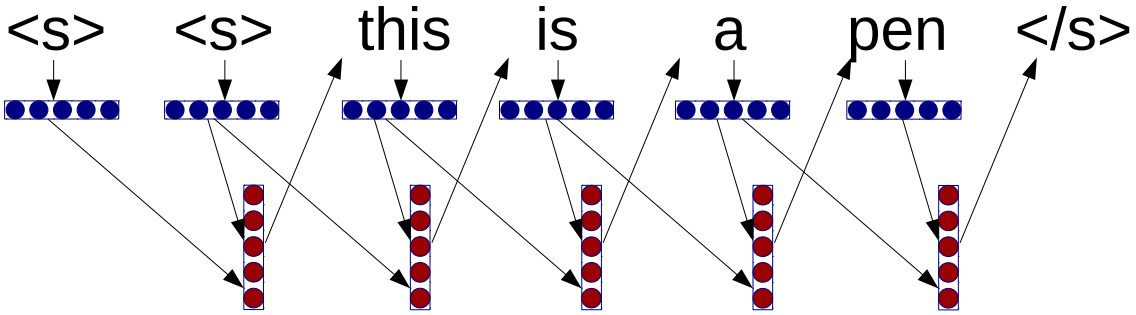
- 単語の類似性は未考慮

my pet cat
my pet dog
my pet wallaby

- 珍しい文脈で壊れる

my name is john smith
my name is jacob smith
my name is sylvan smith

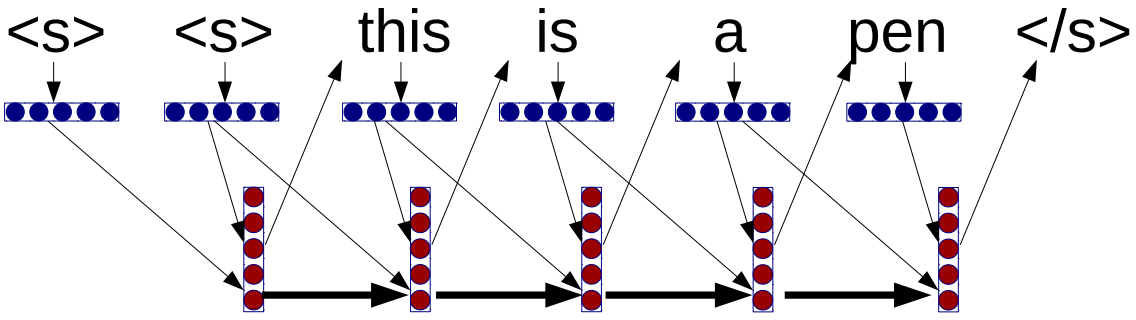
ニューラルネット言語モデル [Nakamura+ 90, Bengio+ 06]



- 低次元隠れ層で出力の類似性を考慮
- 単語表現で文脈の類似性を考慮
- 文脈のすべての単語を直接考慮するため、未知語を含めた文脈で壊れない

19

リカレントニューラルネット言語モデル [Mikolov+ 10]

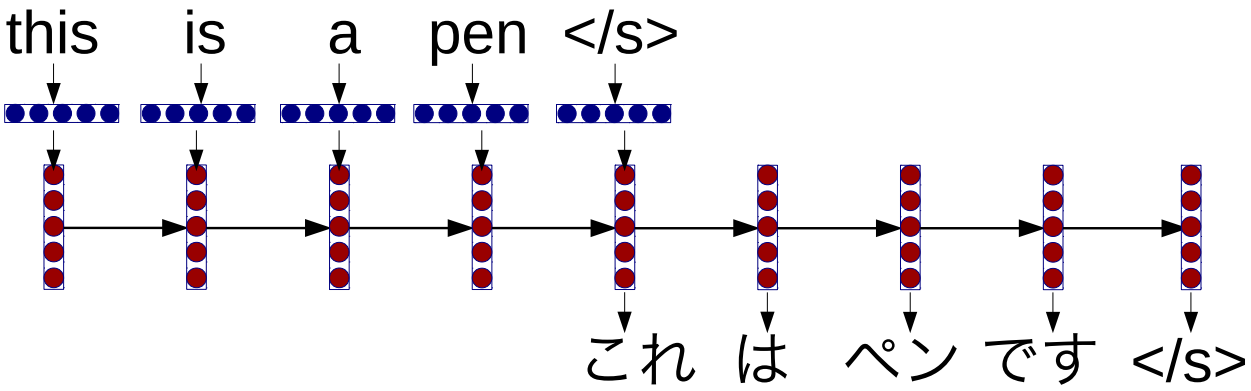


- 以前の単語を「記憶」することが可能
- ４言語対の翻訳における BLEU 改善 [Neubig 14]

  英日 : +0.71 日英 : +0.96 中日 : +0.79 日中 : +0.51

20

リカレントニューラルネット翻訳モデル [Sutskever+ 14]



- 独立、もしくは既存手法の一部として利用可能
- 簡単な割に高精度
- これから大きな波紋？

英日・日英翻訳の現状？

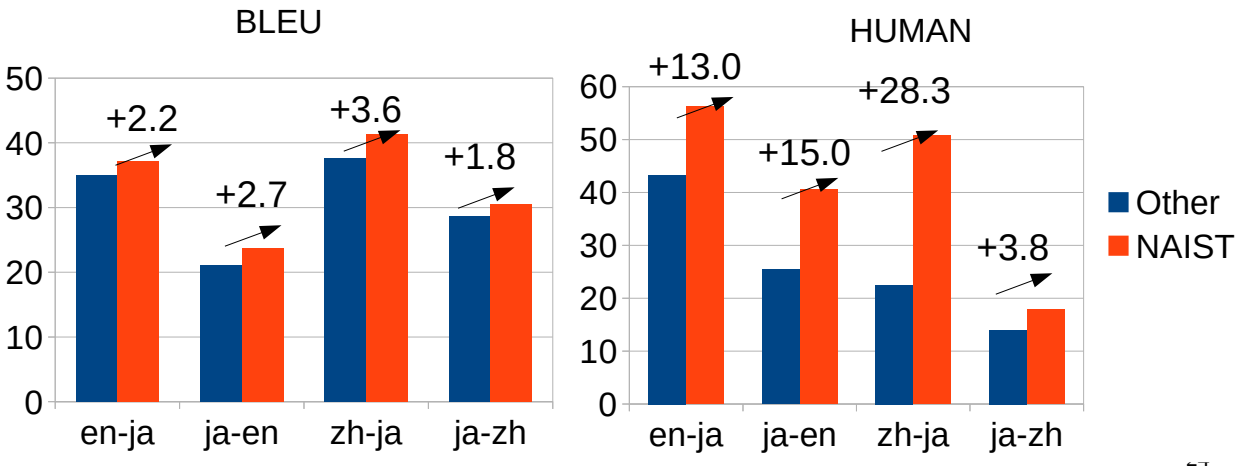
現状の印象

- ルールベース？統計ベース？
  - 英日一般分野： 統計ベース ＞ ルールベース
  - 日英一般分野： ルールベース ＞ 統計ベース
  - 専門分野： 統計ベース ＞ ルールベース
- 問題の進捗度合い：
  - 高：並べ替え（長文でも）
  - 高：メジャーな慣用句・固有表現
  - 中：主語の補間
  - 中：口語、崩れた文
  - 低：曖昧な語彙の選択

23

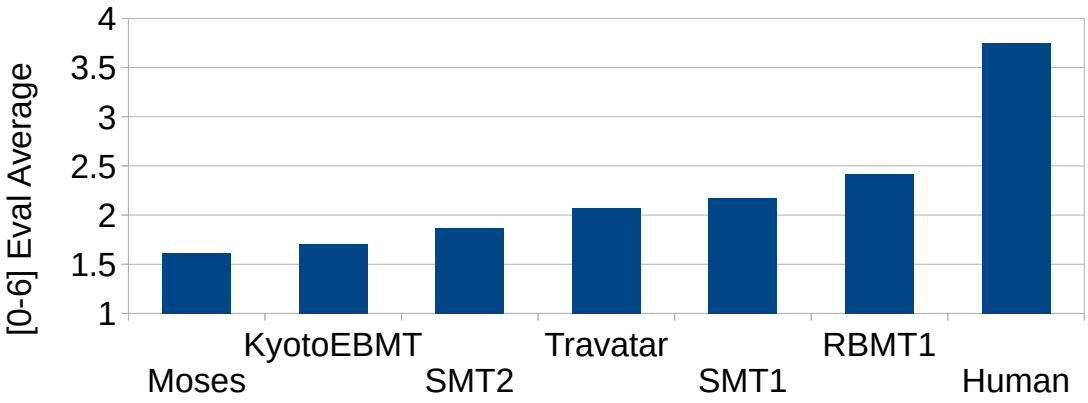
事例： Workshop on Asian Tranlsation

- 科学論文の英日・日英・中日・日中翻訳
  - 専門分野、 50~200 万文の豊富なデータ
- T2S+NNLM でルールベース・統計ベースを凌いで首位



事例： Project Next

- 現代日本語書き言葉均衡コーパス
- 日英の様々な分野、ぴったりのデータがない

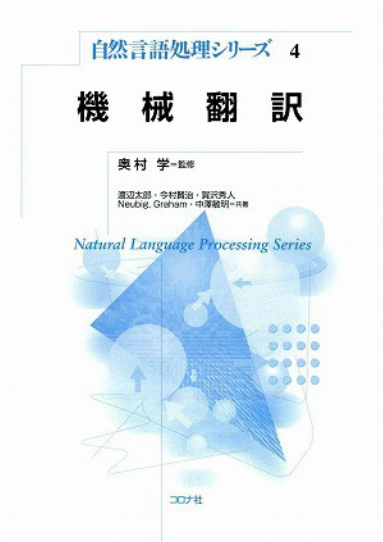


- すべてのシステムに共通の課題：
  文脈依存の語彙選択、解釈が曖昧な文

ご清聴ありがとうございました！

更に勉強するには：

コロナ社「機械翻訳」



参考文献

- Y. Bengio, H. Schwenk, J.-S. Sencal, F. Morin, and J.-L. Gauvain. Neural probabilistic language models. In Innovations in Machine Learning, volume 194, pages 137–186. 2006.
- P. F. Brown, V. J. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19:263–312, 1993.
- Y. Hatakoshi, G. Neubig, S. Sakti, T. Toda, and S. Nakamura. Rule-based syntactic preprocessing for syntax-based machine translation. In Proc. SSST, 2014.
- H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh. Head finalization: A simple reordering rule for SOV languages. In Proc. WMT and MetricsMATR, 2010.
- P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In Proc. HLT, pages 48–54, 2003.
- T. Kudo, H. Ichikawa, and H. Kazawa. A joint inference of deep case analysis and zero subject generation for Japanese-to-English statistical machine translation. In Proc. ACL, pages 557–562, 2014.
- Y. Liu, Q. Liu, and S. Lin. Tree-to-string alignment template for statistical machine translation. In Proc. ACL, 2006.
- T. Mikolov, M. Karafi´at, L. Burget, J. Cernocky`, and S. Khudanpur. Recurrent neural network based language model. In Proc. 11th InterSpeech, pages 1045–1048, 2010.
- M. Nakamura, K. Maruyama, T. Kawabata, and K. Shikano. Neural network approach to word category prediction for English texts. In Proc. COLING, 1990.
- G. Neubig and K. Duh. On the elements of an accurate tree-to-string machine translation system. In Proc. ACL, 2014.
- I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems, pages 3104–3112, 2014.
- F. Xia and M. McCord. Improving a statistical MT system with automatically learned rewrite patterns. In Proc. COLING, 2004.