複数の目的言語の同時生成による統計的機械翻訳

Graham Neubig, Philip Arthur, Kevin Duh
奈良先端科学技術大学院大学 (NAIST)

言語処理学会第 21 回年次大会 (NLP2015)

人間は２つ目の目的言語を活用できる！

機械は同じようなことができるか？

統計的機械翻訳の定式化

京都议定书 → ?? Kyoto Treaty / Kyoto Sanction / Kyoto Protocol

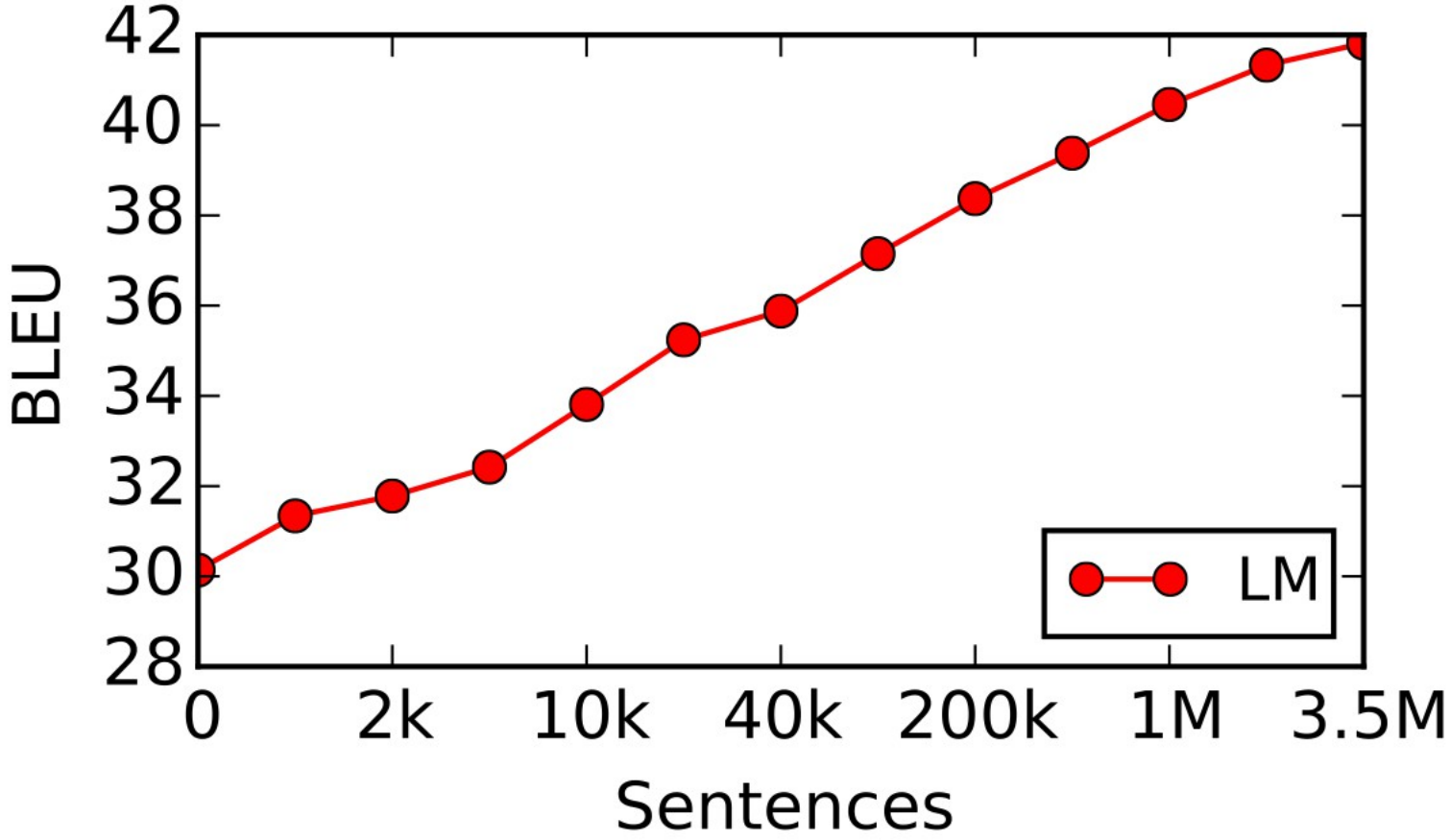
$$\underset{E}{\operatorname{argmax}}\, P(E|F) \approx \underset{E}{\operatorname{argmax}} \sum_i w_i \varphi_i(F, E, D)$$

原言語文F、目的言語文E、導出D
に対して定義される素性関数

素性関数

フレーズ翻訳確率	単語数	言語モデル	
$\varphi_{TM} = \sum_{\langle f, e \rangle \in D} \log P(f|e)$	$\varphi_W = \lvert E \rvert$	$\varphi_{LM} = \log P(E)$	など…

言語モデルの重要性



大きい言語モデル → 高い精度

小さい言語モデル → 低い精度

複数の目的言語モデルを使った翻訳

原言語		目的言語1		目的言語2
京都议定书	?? →	Kyoto Treaty	↔	京都条約
	→	Kyoto Sanction	↔	京都制裁
	→	Kyoto Protocol	↔	京都議定書

$$\varphi_{LM1} = \log P(E_1) \qquad \varphi_{LM2} = \log P(E_2)$$

データが足りないため
弱いモデル

データが多いため
強いモデル

E2 の情報を使って、
E1 の精度を上げられるか？

同期文脈自由文法 [Chiang 05]
(Synchronous Context Free Grammars; SCFG)

翻訳における同期文脈自由文法

联合国 → the UN
京都 → Kyoto

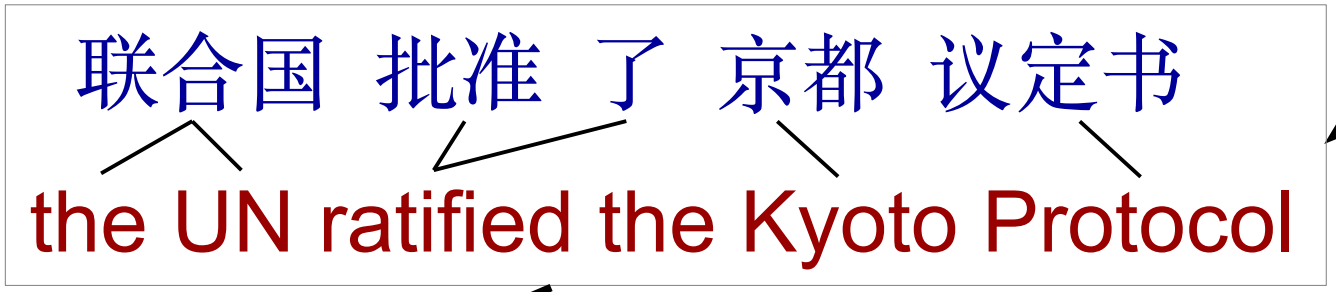
両言語の単語列

X_0 批准 了 X_1 → X_0 ratified the X_1
X_0 议定书 → X_0 protocol

変数を含むルール

同期文脈自由文法の学習

対訳データ：



単語対応

フレーズ抽出

联合国 → the UN
京都 议定书 → Kyoto protocol
京都 → Kyoto
议定书 → protocol
联合国 批准 了 京都 议定书
→ the UN ratified the Kyoto Protocol

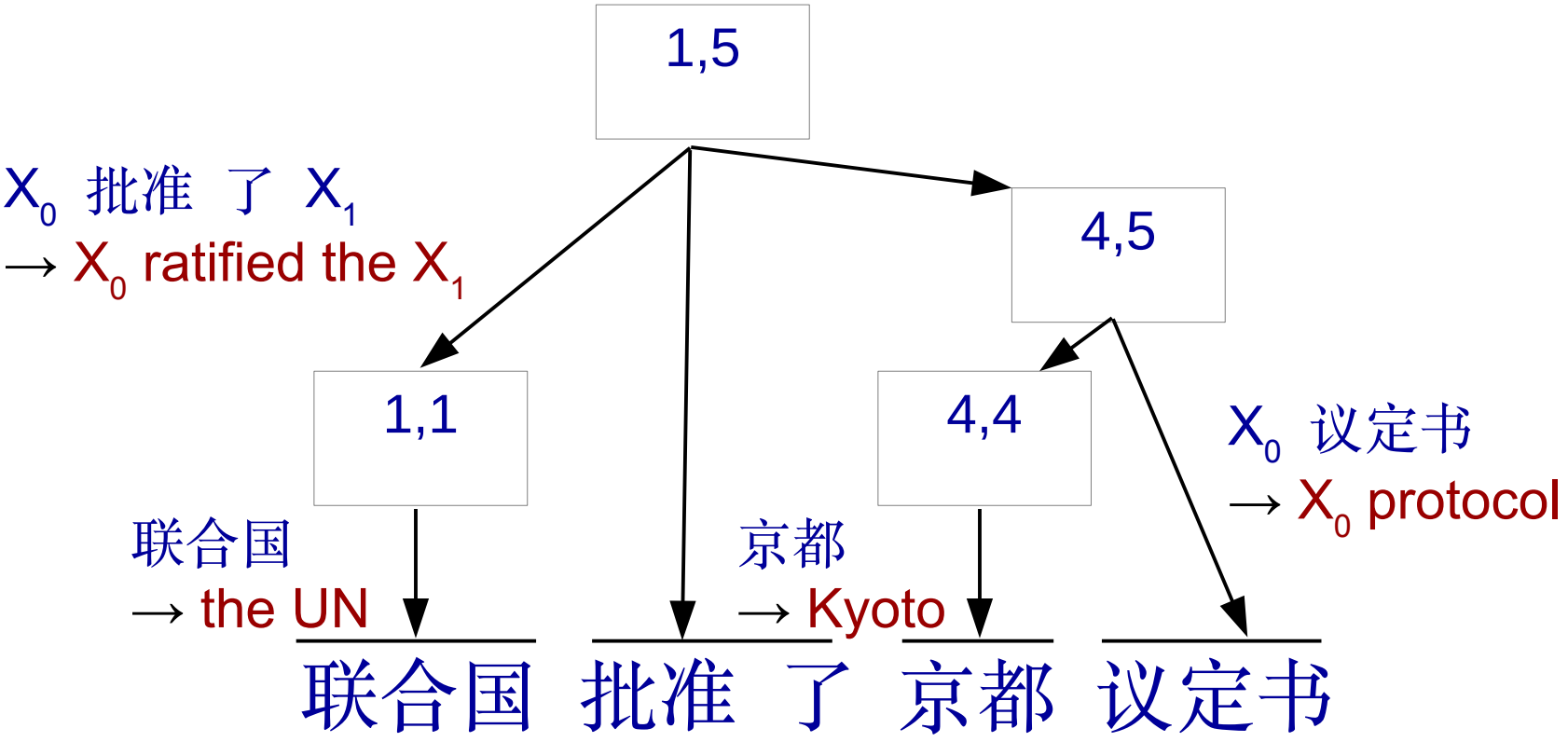
ルール作成

X_0 议定书 → X_0 protocol
京都 X_0 → Kyoto X_0
X_0 批准 了 X_1
→ X_0 ratified the X_1

ルールのスコア付け

翻訳モデル

同期文脈自由文法を用いた訳出



the UN ratified the Kyoto protocol

複数同期文脈自由文法
(Multi-Synchronous Context Free Grammars; MSCFG)

複数同期文脈自由文法のルール

联合国 → the UN | 国連

京都 → Kyoto | 京都

X_0 批准 了 X_1

→ X_0 ratified the X_1 | X_0 が X_1 を 批准した

X_0 议定书

→ X_0 protocol | X_0 議定書

<u>SCFG:</u> 1 つの目的言語を考慮した同期ルール

<u>MSCFG:</u> 1 つ**以上**の目的言語を考慮した同期ルール

MSCFG の学習・訳出

対訳データ

联合国 批准 了 京都 议定书
the UN ratified the Kyoto Protocol
国連 が 京都 議定書 を 批准した

→ MSCFG

単言語データ

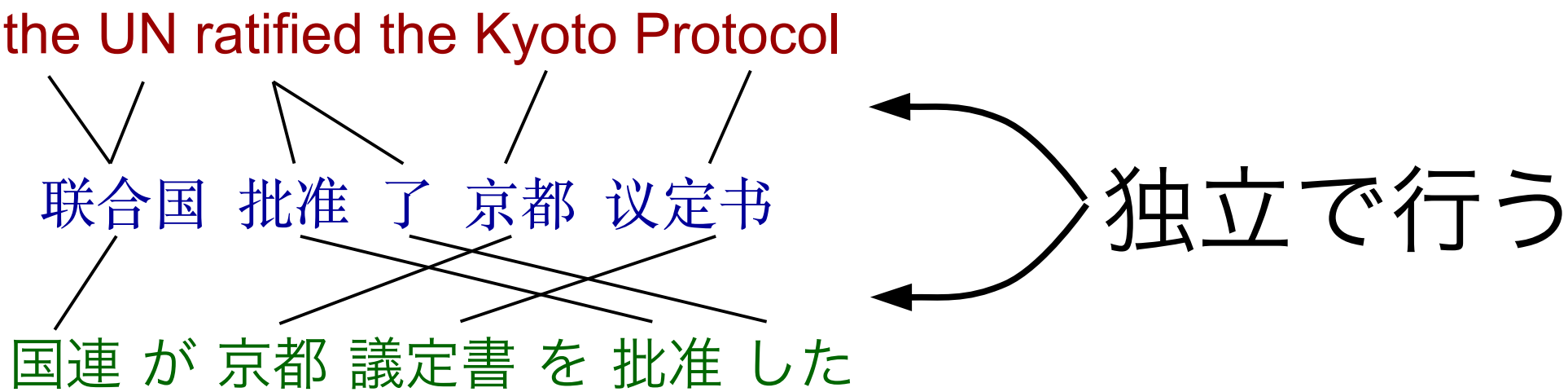
the UN ratified the Kyoto Protocol → 言語モデル 1

国連 が 京都 議定書 を 批准した → 言語モデル 2

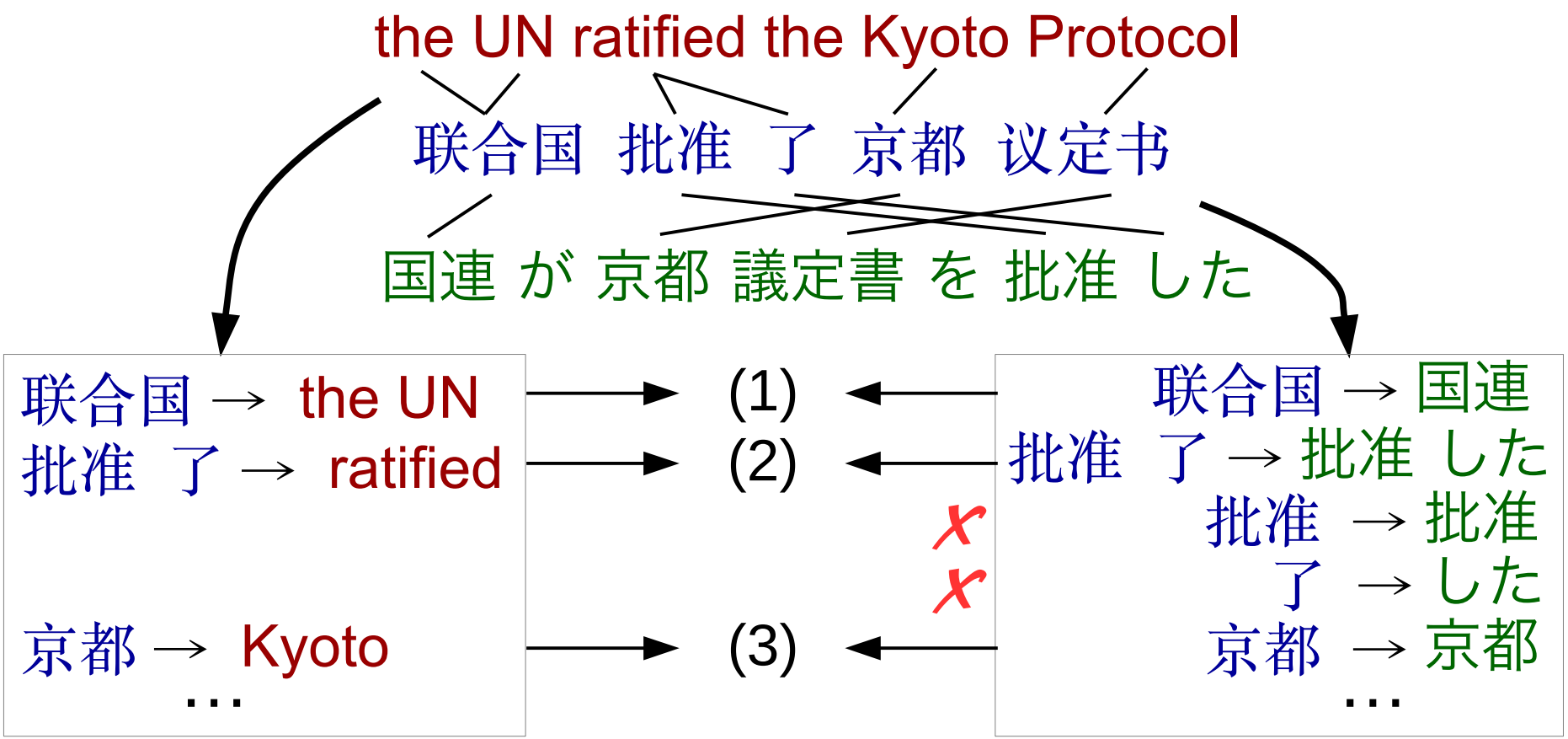
変更が必要なプロセス

- 単語の対応付け
- フレーズの抽出
- ルールのスコア付け
- 探索

単語の対応付け


the UN ratified the Kyoto Protocol
联合国 批准 了 京都 议定书
国連 が 京都 議定書 を 批准 した
独立で行う

フレーズ抽出



(1)	联合国	→ the UN	国連
(2)	批准 了	→ ratified	批准 した
(3)	京都	→ Kyoto	京都

SCFG/MSCFG における素性

SCFG

翻訳確率 (1)

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_1 \rangle} \log P(\boldsymbol{f} | \boldsymbol{e}_1)$$

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_1 \rangle} \log P(\boldsymbol{e}_1 | \boldsymbol{f})$$

語彙化翻訳確率 (1)

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_1 \rangle} \log P_{lex}(\boldsymbol{f} | \boldsymbol{e}_1)$$

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_1 \rangle} \log P_{lex}(\boldsymbol{e}_1 | \boldsymbol{f})$$

単語数 (1)

$$|E_1|$$

言語モデル (1)

$$\log P(E_1)$$

ルール数

$$|D|$$

MSCFG

翻訳確率 (2)

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_2 \rangle} \log P(\boldsymbol{f} | \boldsymbol{e}_2)$$

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_2 \rangle} \log P(\boldsymbol{e}_2 | \boldsymbol{f})$$

翻訳確率 (1+2)

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle} \log P(\boldsymbol{f} | \boldsymbol{e}_1, \boldsymbol{e}_2)$$

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle} \log P(\boldsymbol{e}_1, \boldsymbol{e}_2 | \boldsymbol{f})$$

語彙化翻訳確率 (2)

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_2 \rangle} \log P_{lex}(\boldsymbol{f} | \boldsymbol{e}_2)$$

$$\sum_{\langle \boldsymbol{f}, \boldsymbol{e}_2 \rangle} \log P_{lex}(\boldsymbol{e}_2 | \boldsymbol{f})$$

単語数 (2)

$$|E_2|$$

言語モデル (2)

$$\log P(E_2)$$

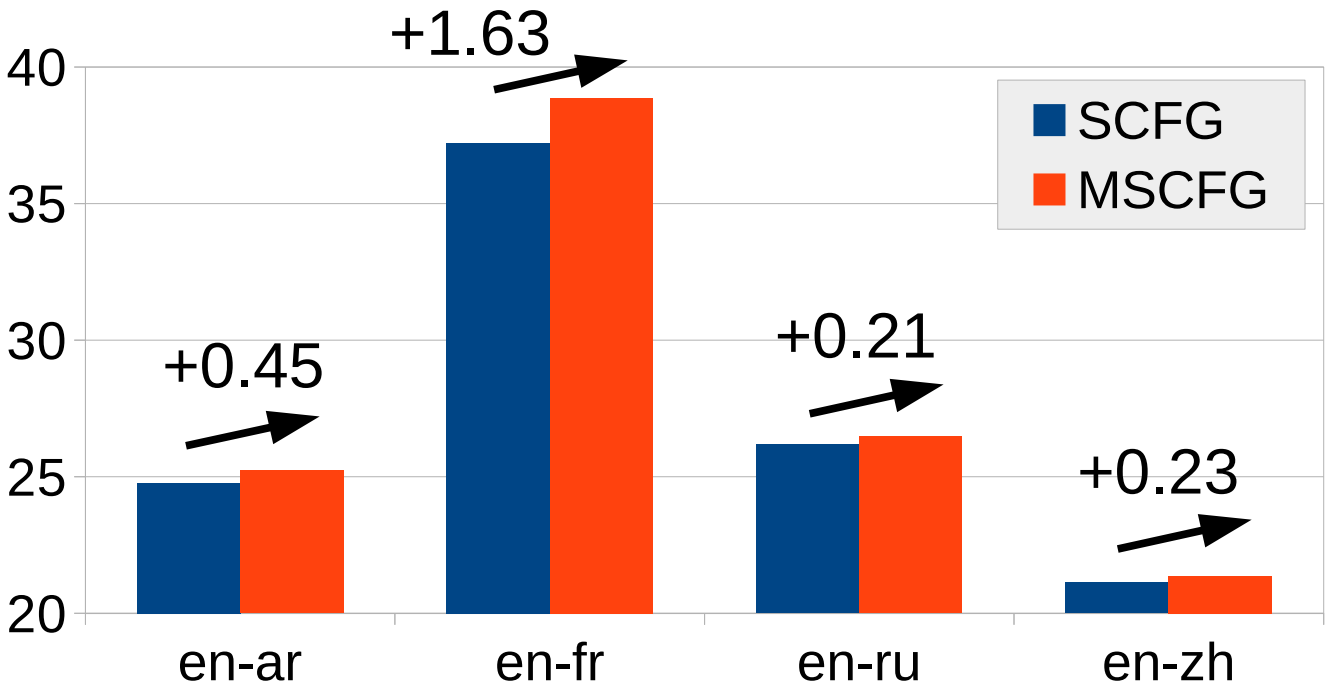
実験結果

複数の目的言語を考慮することで翻訳精度が向上するか？

国連コーパス：

対訳、E1 言語モデルデータ → 10 万文
E2 言語モデルデータ → 400 万文

F: 英　E1: 亜・仏・露・中　E2: 西



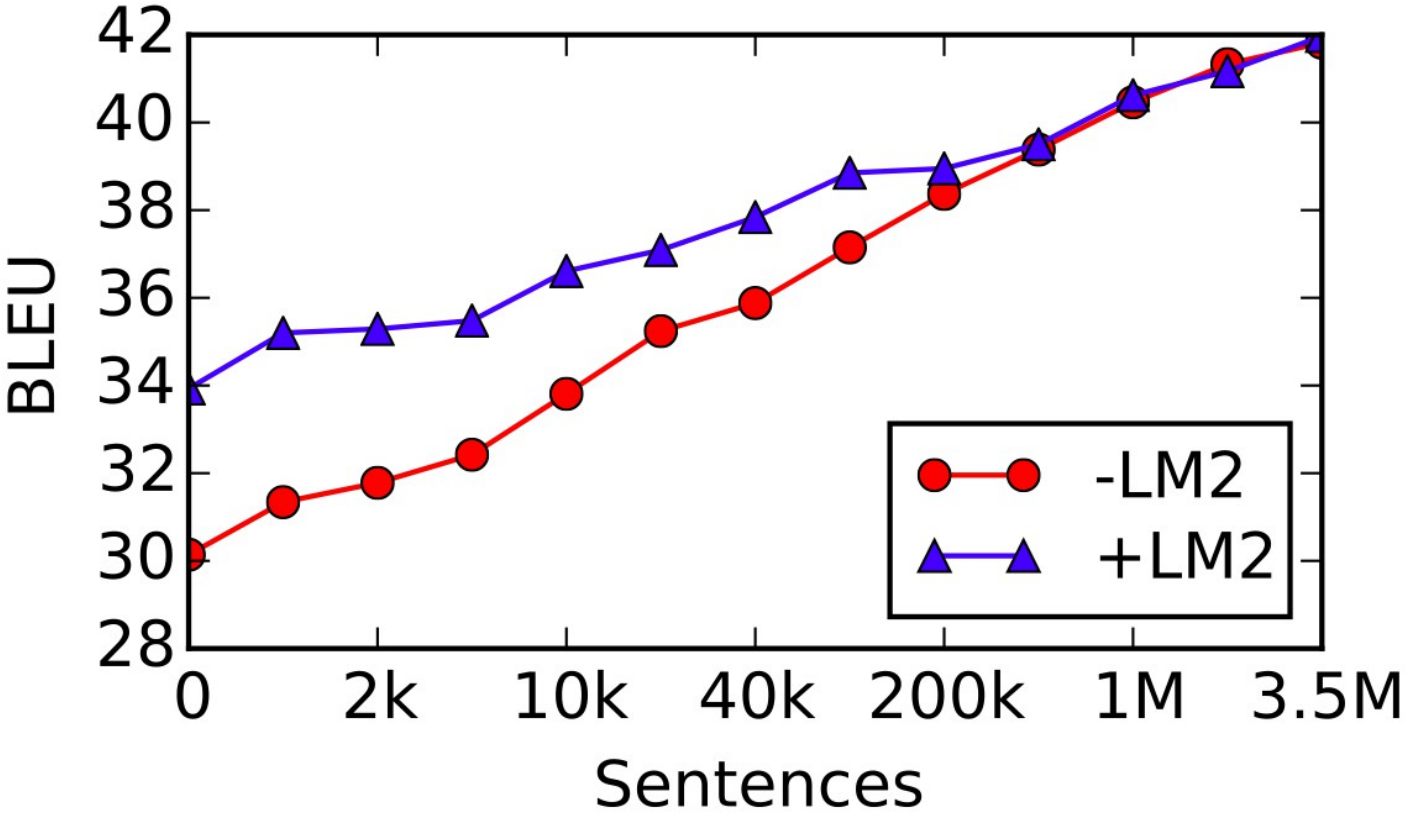
4 つの言語対における精度向上

類似している言語では効果的

向上幅は目的言語１の言語モデルの強さに依存するか？

対訳→ 10 万文　E1 言語モデルデータ → **変動**
E2 言語モデルデータ → 400 万文

F: 英　E1: 仏　E2: 西

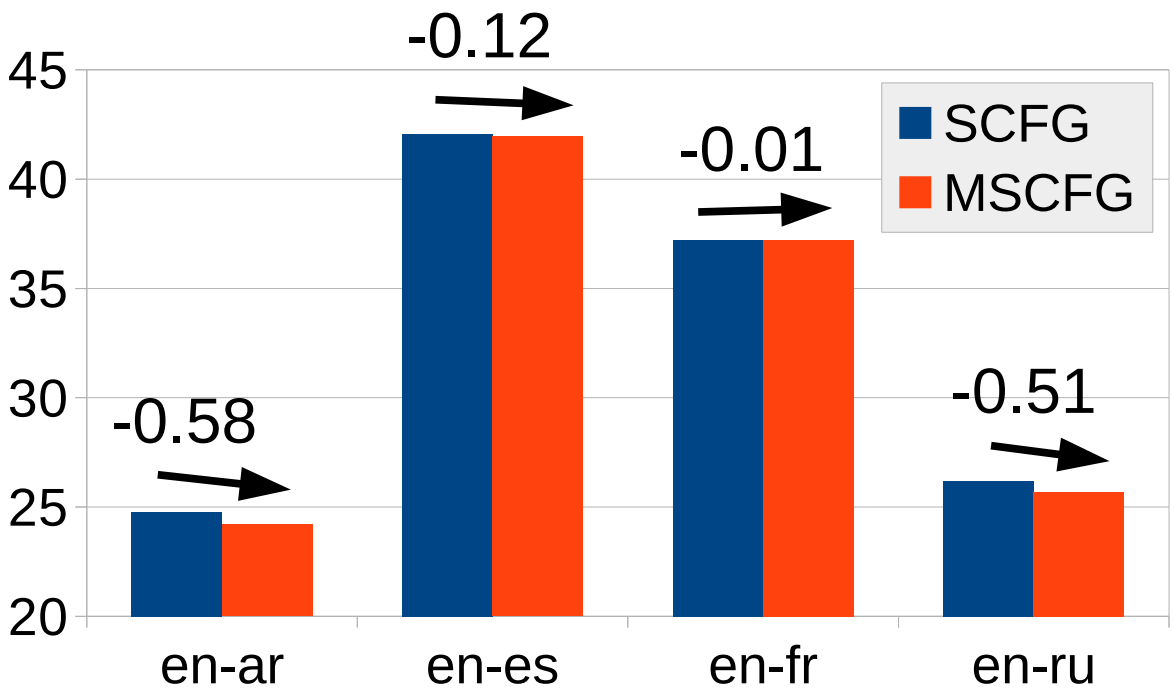


目的言語１の資源が少ない場合に効果的

目的言語２の性質に依存するか？

対訳、　E1 言語モデルデータ → 10 万文
　　E2 言語モデルデータ → 400 万文

F: 英　　E1: 亜・西・仏・露　　E2: 中



目的言語２が
他と異なると
悪化することも

今後の課題

今後の課題

- 複数の目的言語を用いた単語の対応付け

- 両方の目的言語の精度の最適化

- 言い換えなど、他分野への適応

ソースなど近日公開予定
http://www.phontron.com/project/naacl2015

付録

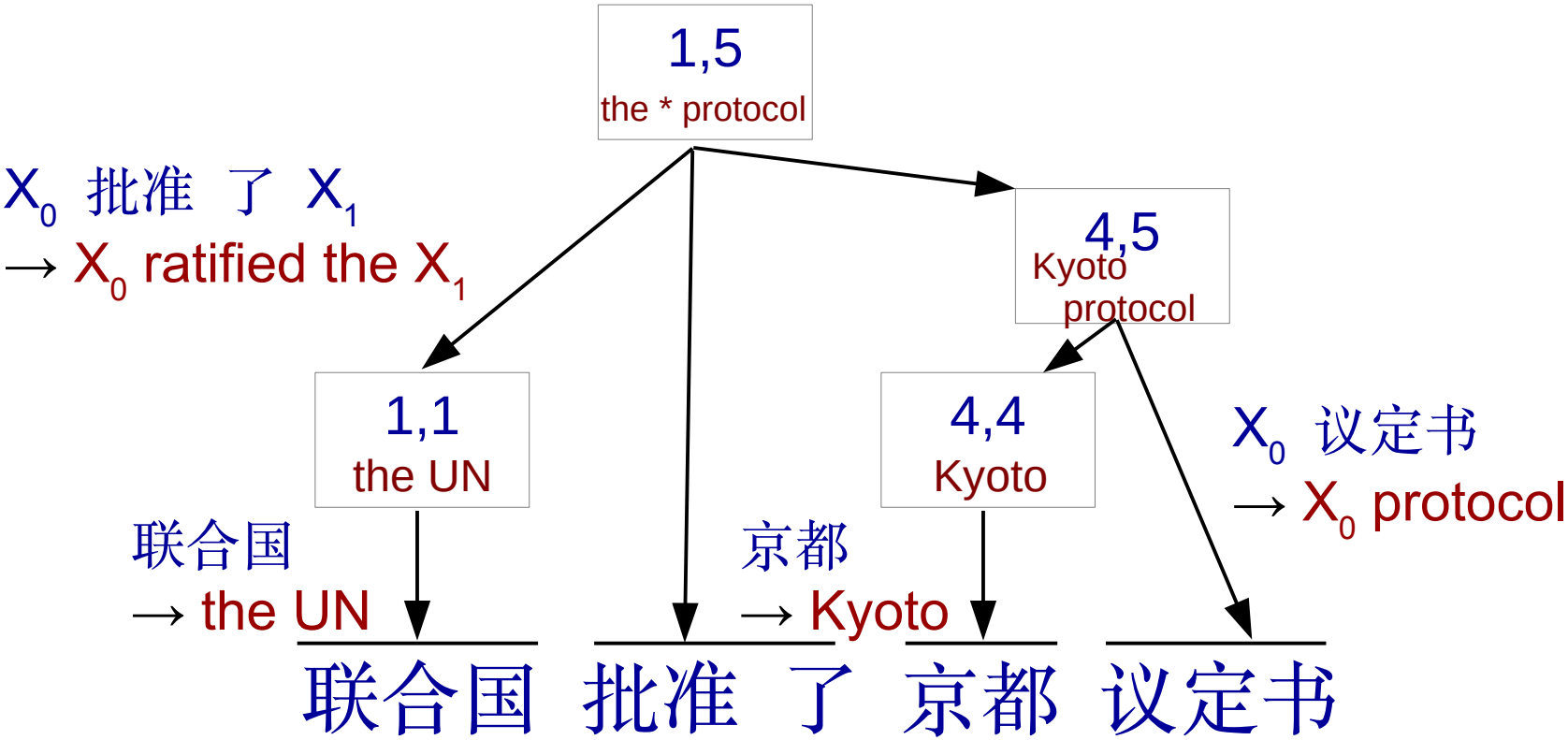
英→仏 + 西における改善例

『前置詞「de」「du」「de la」「des」の使い分けはフランス語を学ぶ人にとって非常に難しい…』
出典：About.comフランス語講座

単語	回数	単語	回数
de	70	)	10
du	29	parle	9
,	17	anglais	9
une	14	jour	8
pour	12	ordre	8
l'	12	le	6
des	10	et	5
au	10	que	5
(	10	dans	4
les	10	comité	4

http://french.about.com/od/grammar/a/de-vs-du-de-la-des.htm

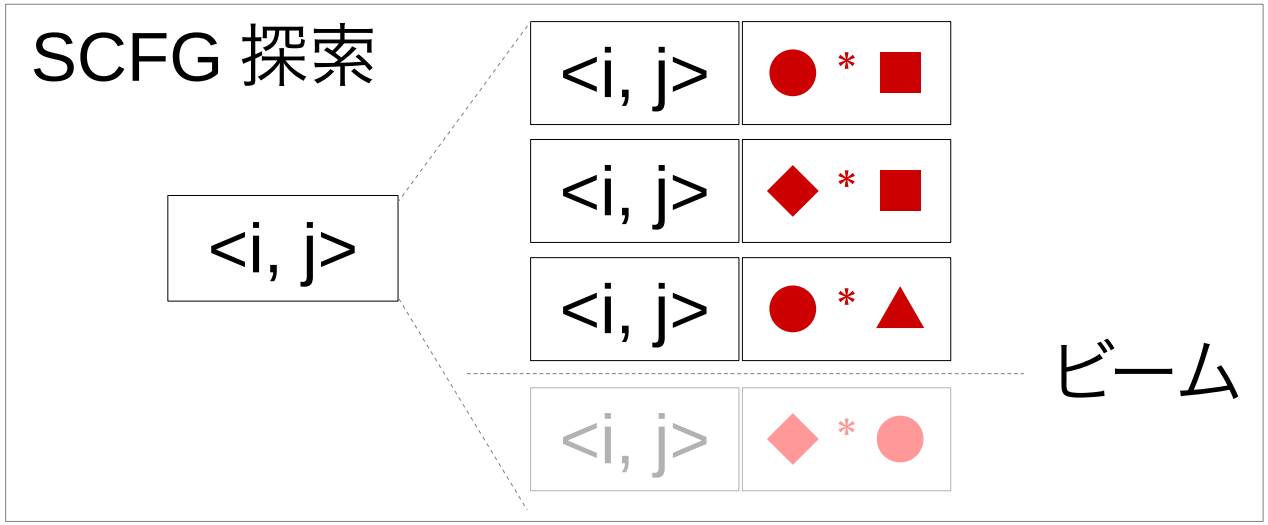
同期文脈自由文法を用いた訳出

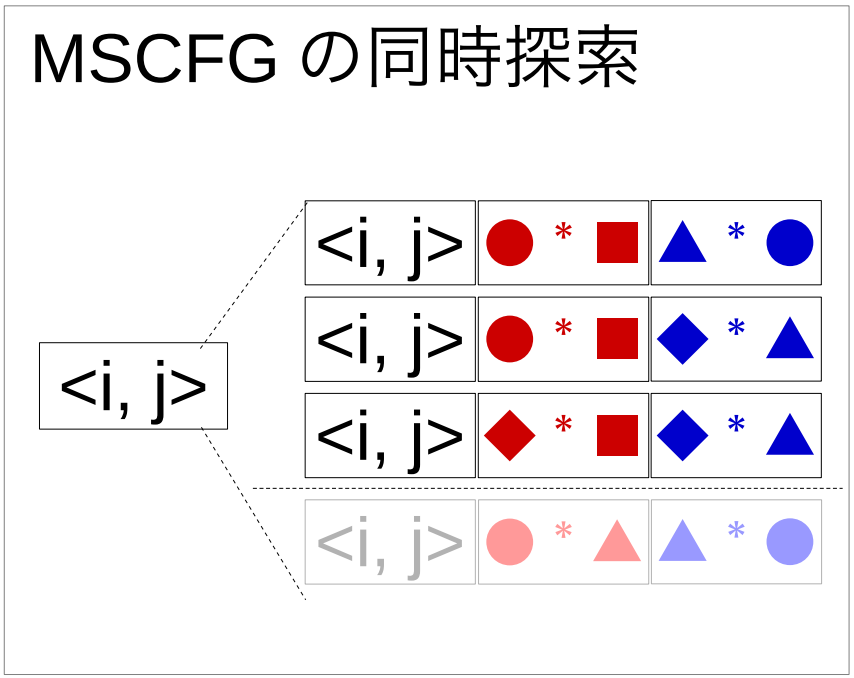


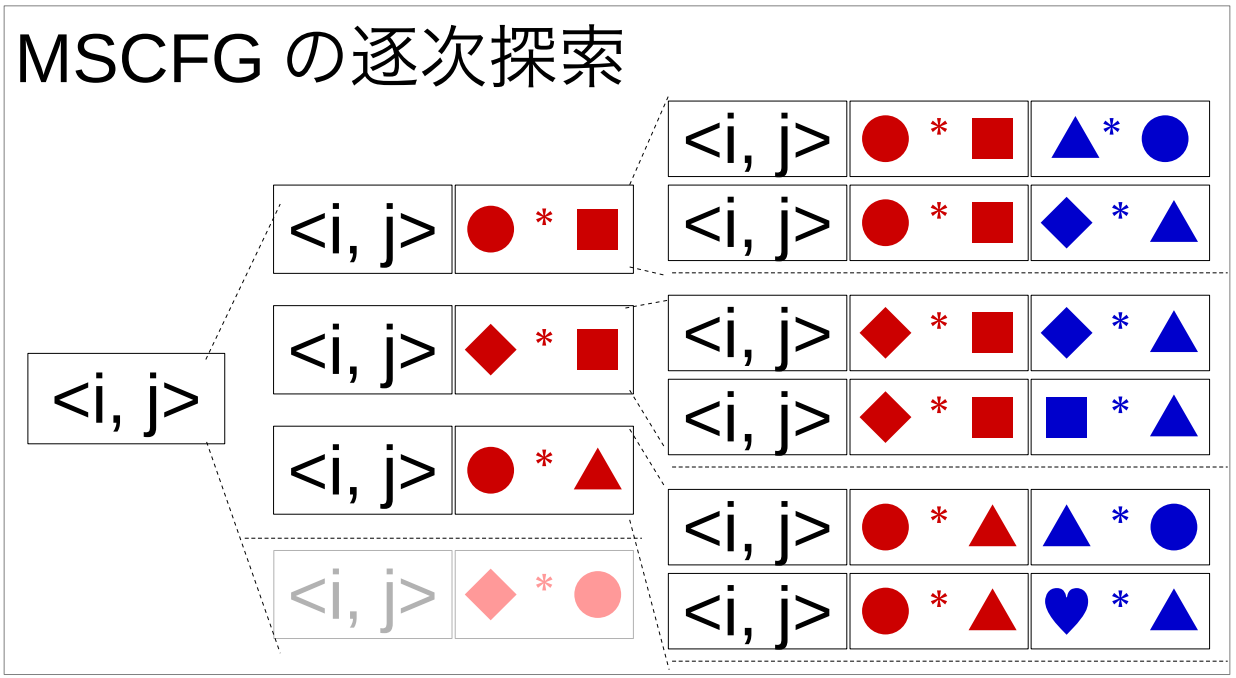
the UN ratified the Kyoto protocol

言語モデルを追加するため、
両側の単語を記憶

探索


SCFG 探索
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
ビーム


MSCFG の同時探索
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>


MSCFG の逐次探索
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>
<i, j>