Introduction of Pattern Based Statistical Machine Translation (PBSMT)

(≠ Phrase Based SMT)

鳥取大学大学 村上仁一

平成26年 12月13日

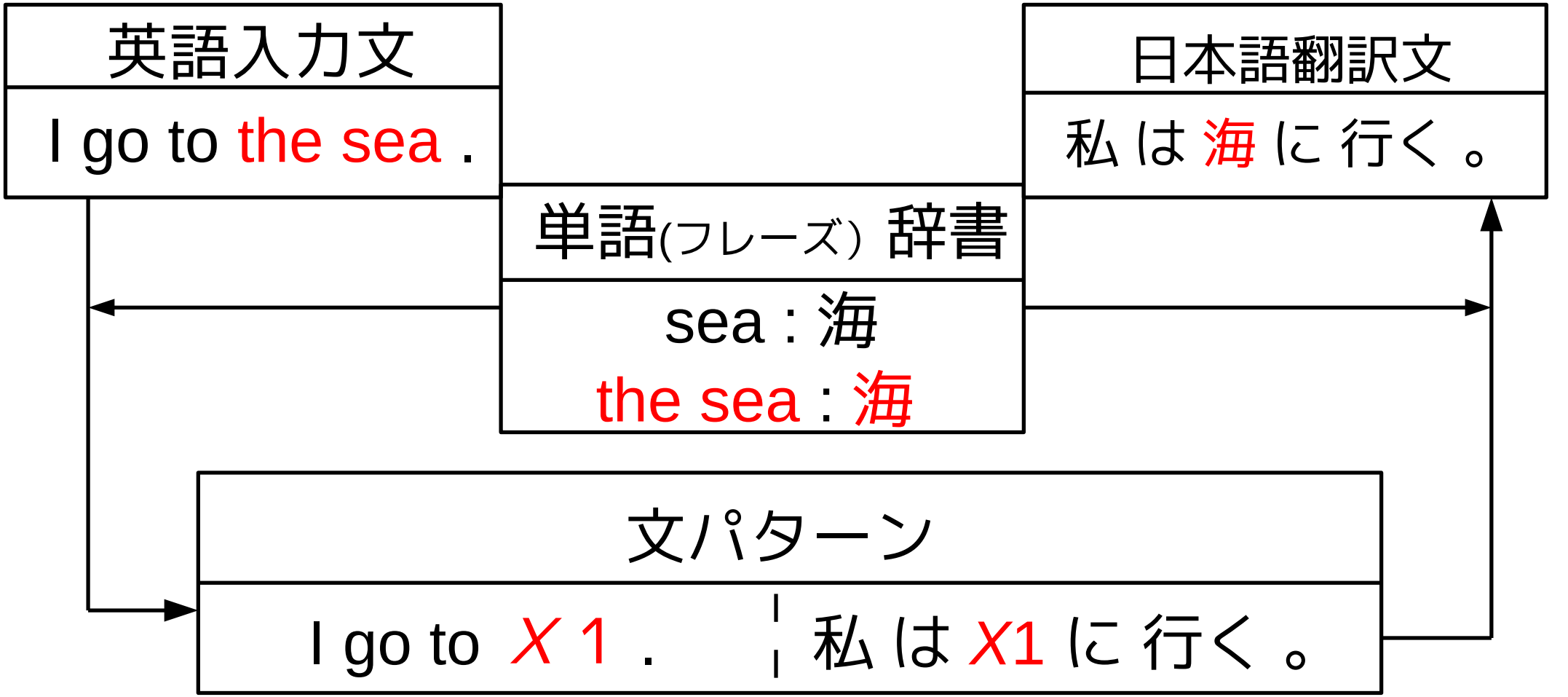
Pattern Based Machine Translation

Long long time ago

First Machine Translation

やまと　電気試験所　1959
「I like music」→「ワレガ　オンガクヲ　コノム」

英日パターン翻訳の概要



Pattern Based Machine Translation

長所：プログラムのimplementが比較的容易

短所： パターンと単語辞書を，人手で作成
（システムが高コスト）
特徴： 翻訳精度とカバー率：トレードオフ
曖昧性の解消：
基本：シソーラス，（単語意味属性）
参考：頻度（確率）

最近の翻訳　　SMT

パラレルコーパスから統計的に処理
（基本はベイズ推定）

Word-Based　　:Giza++
Phrase-Based　　:Och's heuristic
(Grow Diag Final, etc ...)
Hierarchical Phase-Based : (Syntax-Base)(Tree Structure)

長所：パラレルコーパスがあれば，容易に作成可能
短所：翻訳精度の限界

曖昧性の解消：　　基本：確率
参考：シソーラス，（品詞）

翻訳性能　日英翻訳（英日翻訳）

現状：

ルールベース翻訳　＞　SMT

原因（句の精度）

パターンベース翻訳:文全体を考慮して翻訳

SMT,EBMT　:各単語を組合せて翻訳

（要素合成）

［異論はあると思うけど．．．（木構造）］

目標：

統計的手法を用いたパターン翻訳

(Pattern Based SMT)

（ALT由来の歴史）

パターン翻訳の問題点と解決策

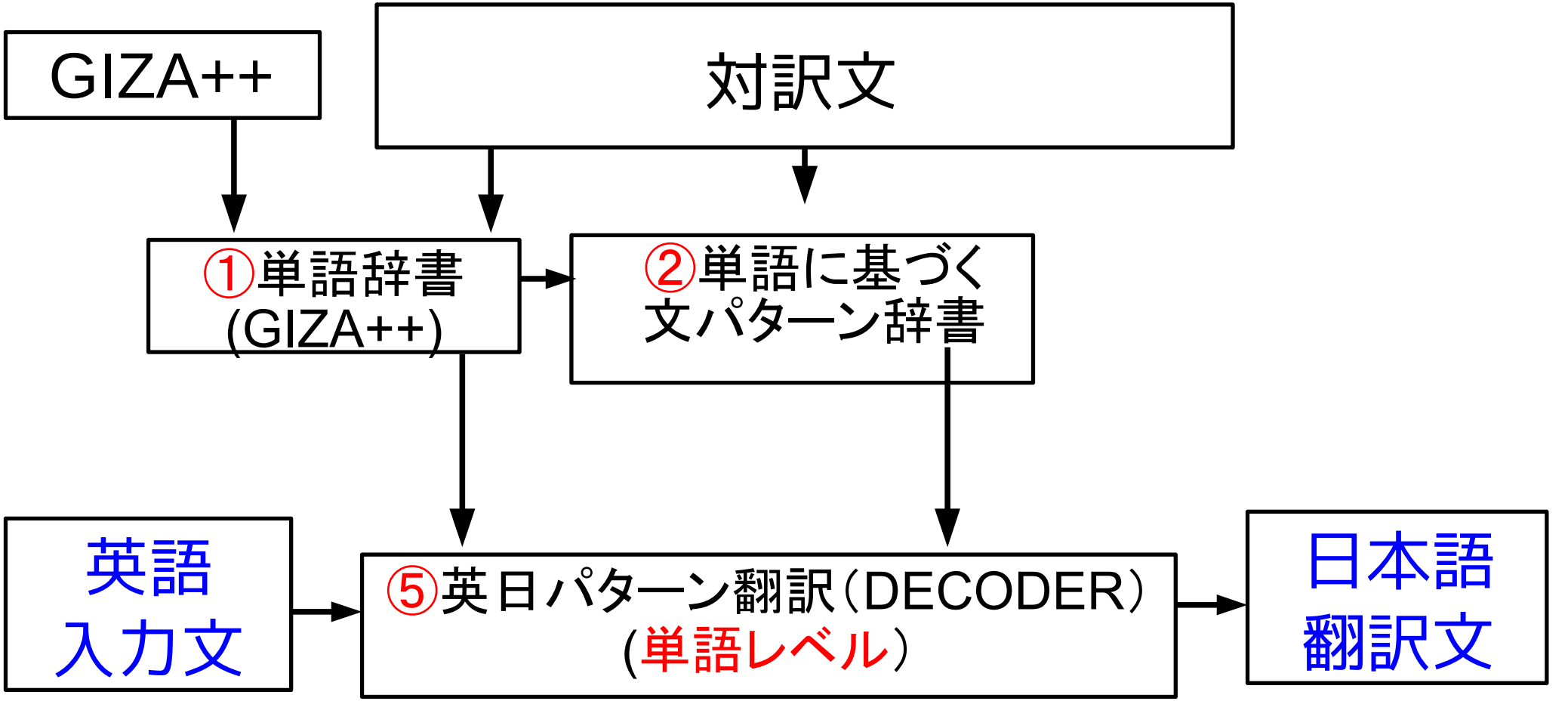
・コスト：高い → 単語辞書・文パターンの人手作成

解決策：単語辞書と文パターンを自動作成　(GIZA++)

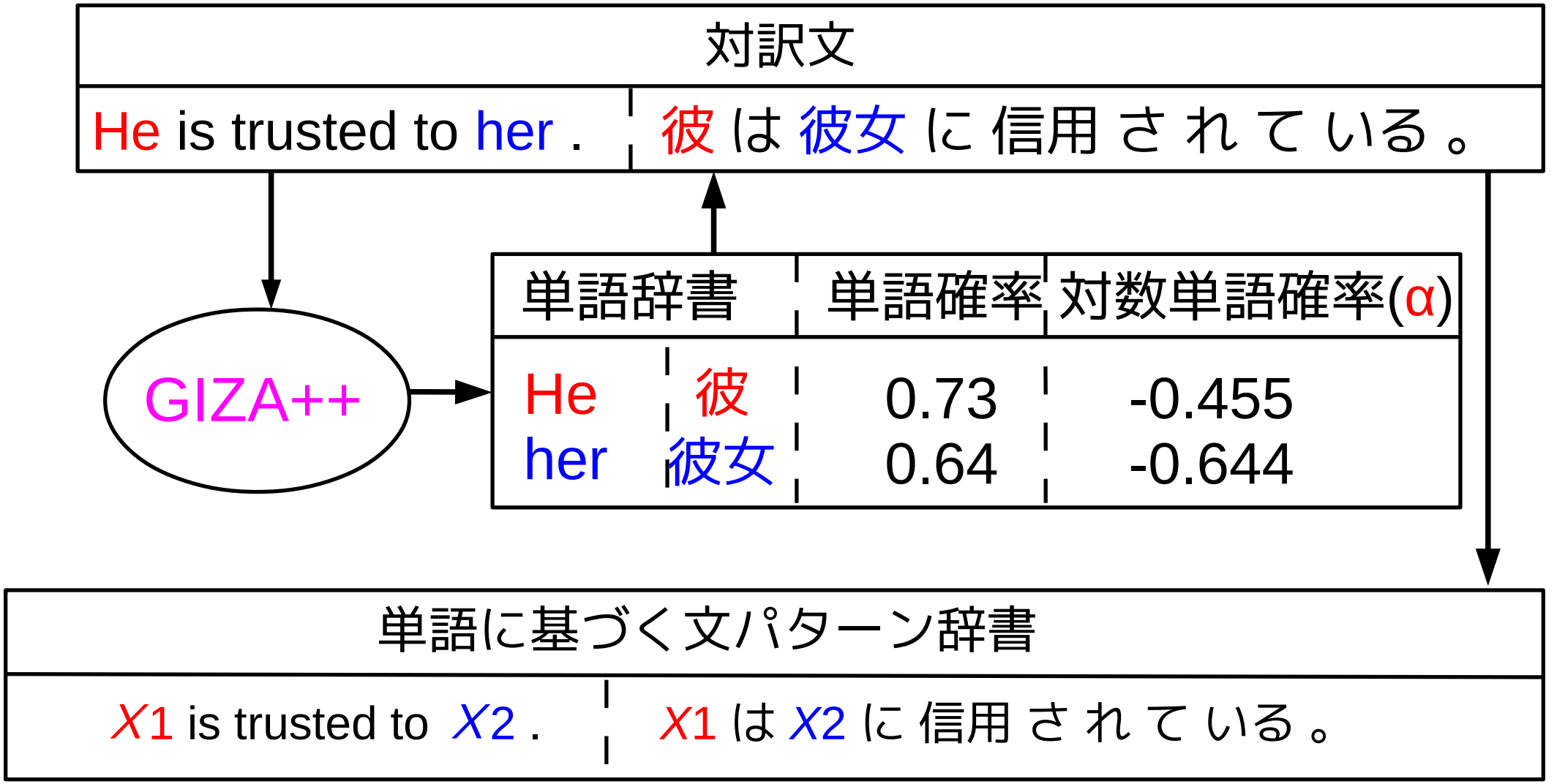
・翻訳精度 ⟷ カバー率

解決策：翻訳精度 → 字面の多い文パターンの選択
：カバー率 → 大量のフレーズ辞書と
句に基づく文パターン辞書の作成

提案手法の概要（単語レベル）PBSMT



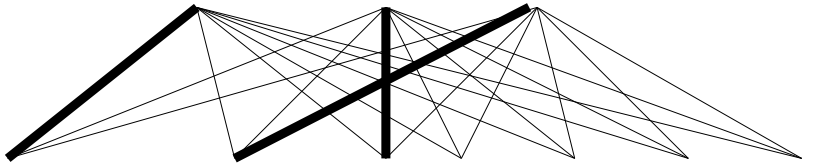
①単語辞書 ②単語に基づく文パターン辞書



文パターン対数確率(β)

句に基づく文パターンの例

*X*1 is trusted to *X*2 .



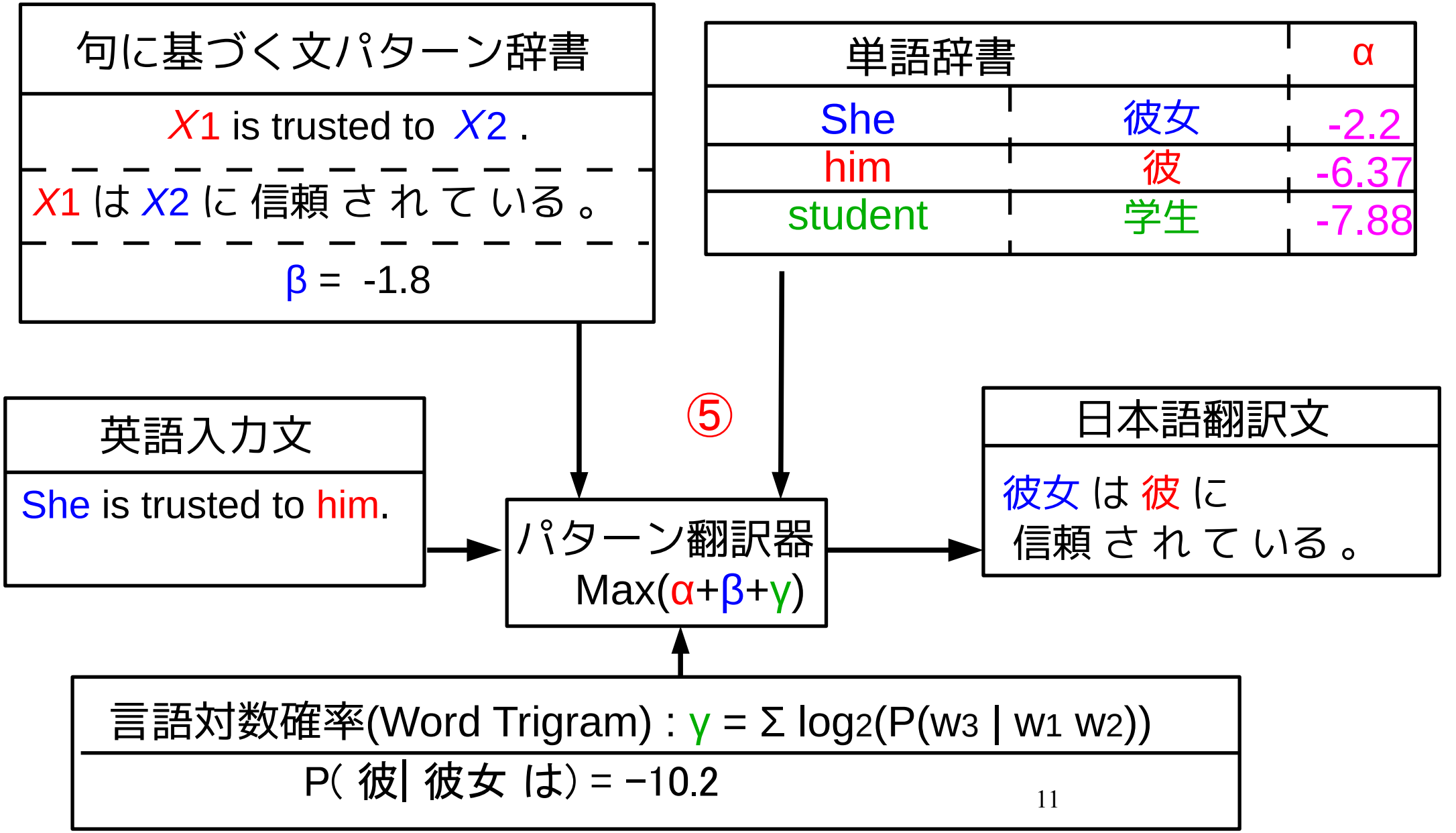
*X*1 は *X*2 に 信頼 さ れ て いる 。

文パターン対数確率(β)の計算

文パターン対数確率(β)
= log2 (“is | は” の翻訳確率
+ log2 (“trusted | 信頼” の翻訳確率)
+ log2 (“to | に” の翻訳確率)
= -0.44-1.02-0.32
= -1.8

GIZA++		単語確率
is	は	0.21
is	に	0.51
is	信頼	0.12
is	さ	0.07
⋮	⋮	⋮
trusted	は	0.21
trusted	に	0.51
trusted	信頼	0.12
⋮	⋮	⋮
to	は	0.21
to	に	0.51
to	信頼	0.12
⋮	⋮	⋮

DECODER(単語レベル)



実験条件(言語対数確率)

入力文 : 日本語文 100,000文

実験条件・結果(辞書の作成)

入力文(単文) : 対訳文 100,000文

単語辞書	21,439単語
単語に基づく文パターン辞書	16,385,504パターン

パターン数の増加の原因

列車 の スピード が 上がっ た 。 ||| The train raised its speed .

列車 の スピード N00 上がっ た 。 ||| Ｎ00 train raised its speed .
列車 の N00 が 上がっ た 。 ||| The train raised its Ｎ00 .
列車 の N00 N01 上がっ た 。 ||| Ｎ01 train raised its Ｎ00 .
N00 の スピード が 上がっ た 。 ||| The Ｎ00 raised its speed .
N00 の スピード N01 上がっ た 。 ||| Ｎ01 Ｎ00 raised its speed .
N00 の N01 が 上がっ た 。 ||| The Ｎ00 raised its Ｎ01 .
N00 の N01 N02 上がっ た 。 ||| Ｎ02 Ｎ00 raised its Ｎ01 .

日英翻訳（単語レベルPBSMT）

入力文：日本語入力文 10文

実験結果

→ 4文の英語翻訳文

カバー率：そこそこ？かなり悪い？
翻訳精度：そこそこ？かなり良い？
(mosesより確実に良い)

翻訳例（単語レベル）

Sentence_id = 0 "彼女 は 我 を 通し た 。 "
Reference "She had her own way . "
Pattern_id = 145 "N00 は 我 を 通し た 。 "
Pattern_id = 145 "Ｎ00 had his own way . "
Original JP = "彼 は 我 を 通し た 。 "
Original EN = "He had his own way . "
Decode_id = 0 "< 彼女 > は 我 を 通し た 。 "
Output "She had his own way . "
TAG 8 "< She > had his own way . "
TRIGRAM -118.832984 VARIABLE -17.032043 PATTERN -12176.252744
SUM -12312.117771
PATTERN -2021.372778 -1022.240279 1 1 1 1 1 1 0.857100
N00 "彼女 " "She " -0.803003 -0.794262 5237 4423 3035

翻訳例（単語レベル）

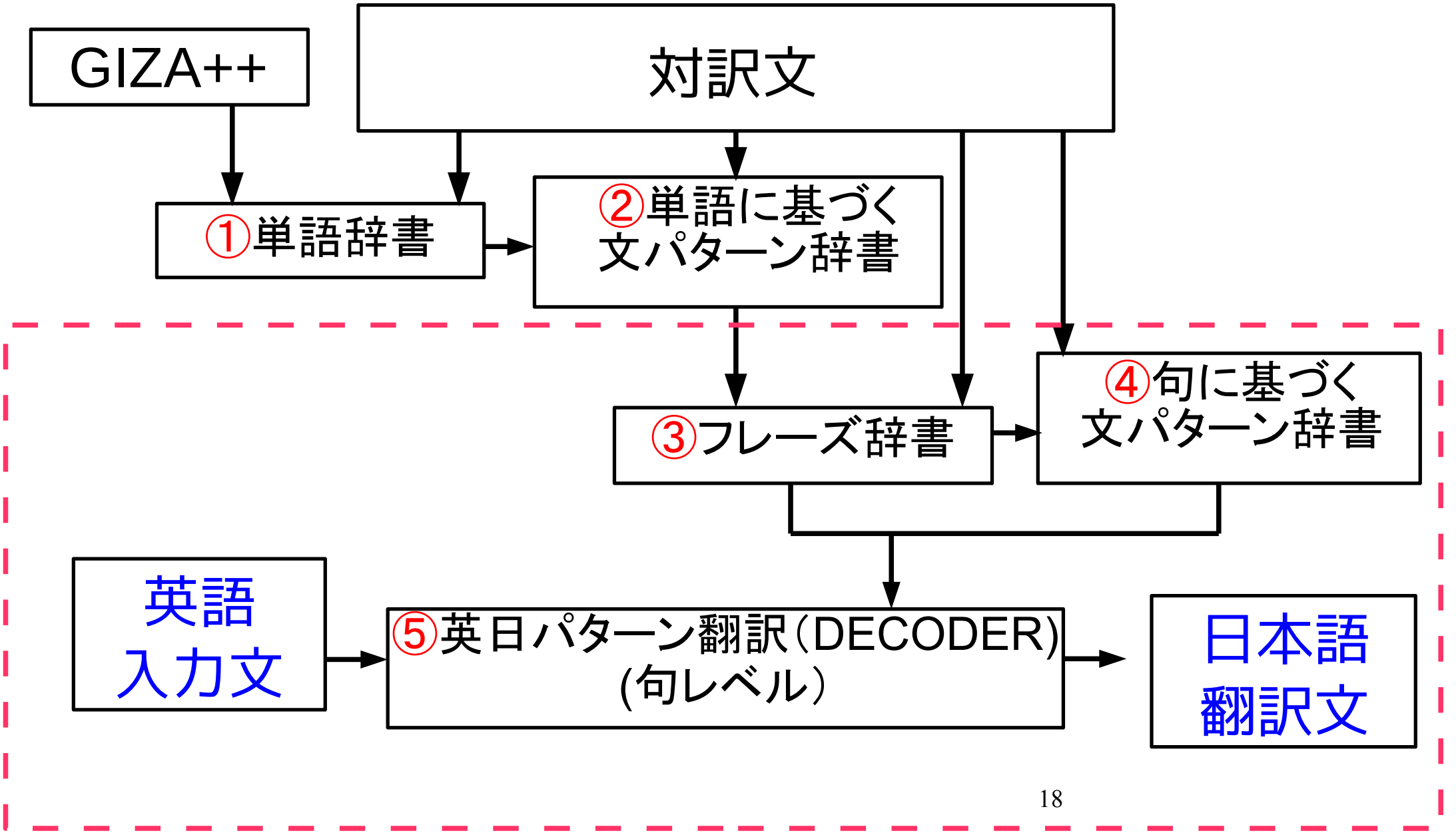
Sentence_id = 0 "彼 は 有罪 の 宣告 を 受け た 。 "
Reference "He was convicted . "
Pattern_id = 175 "N00 は N01 の 宣告 を 受け た 。 "
Pattern_id = 175 "The N00 was N01 to death . "
Original JP = "被告 は 死刑 の 宣告 を 受け た 。 "
Original EN = "The accused was sentenced to death . "
Decode_id = 0 "< 彼 > は < 有罪 > の 宣告 を 受け た 。 "
Output "The He was guilty to death . "
TAG 9 "The < He > was < guilty > to death . "
TRIGRAM -2211.434113 VARIABLE -35.183444 PATTERN -4199.051280
SUM -6445.668837
PATTERN -33.817419 -1011.163444 1 4 1 6 1 1 0.777800
N00 "彼 " "He " -0.714797 -0.546066 15970 13696 9893
N01 "有罪 " "guilty " -0.974909 -1.331845 48 48 29

単語レベルPBSMT

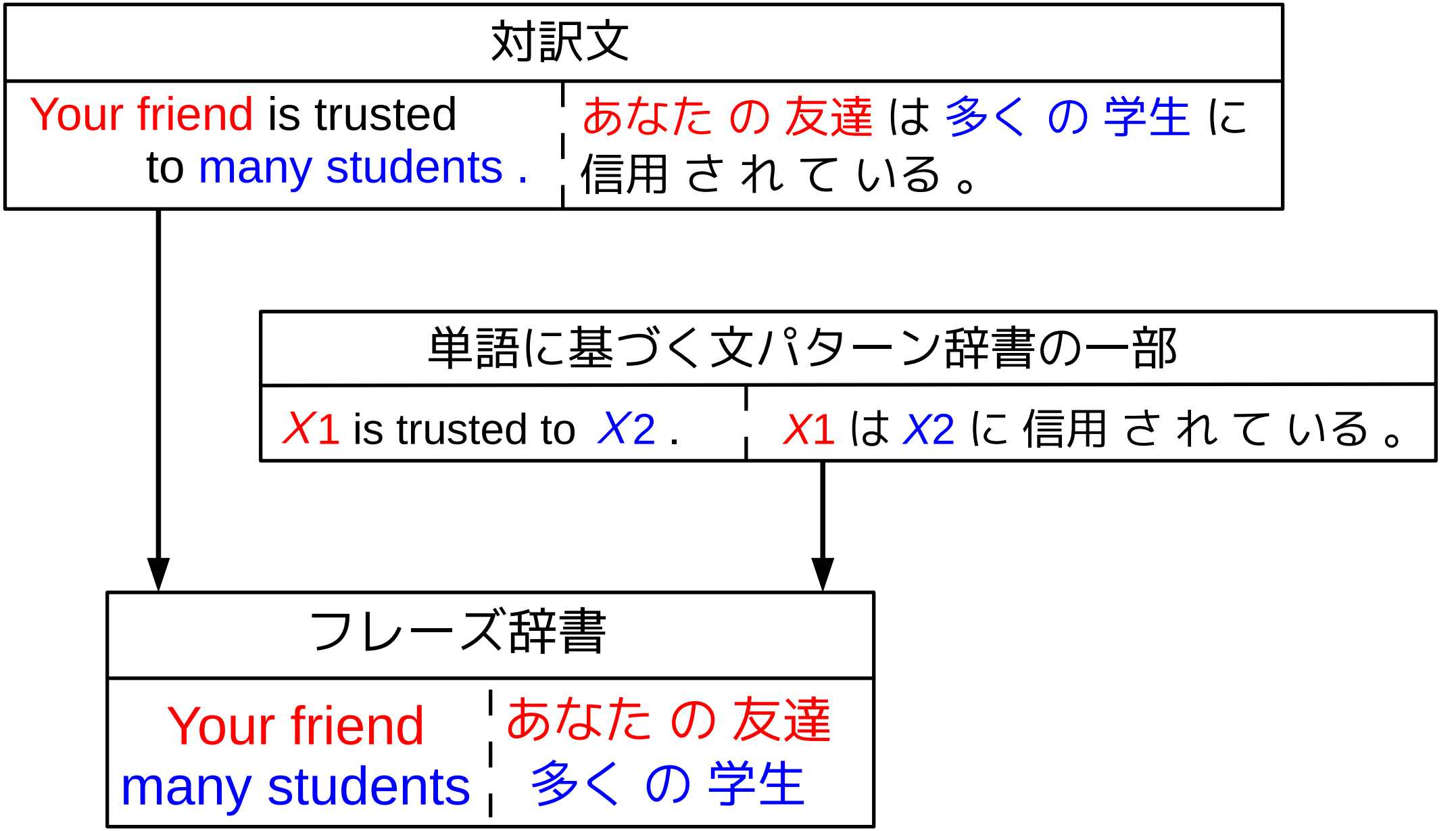
カバー率：そこそこ？かなり悪い？
翻訳精度：そこそこ？かなり良い？
(mosesより確実に良い)

カバー率の向上
単語レベル　→　句レベル

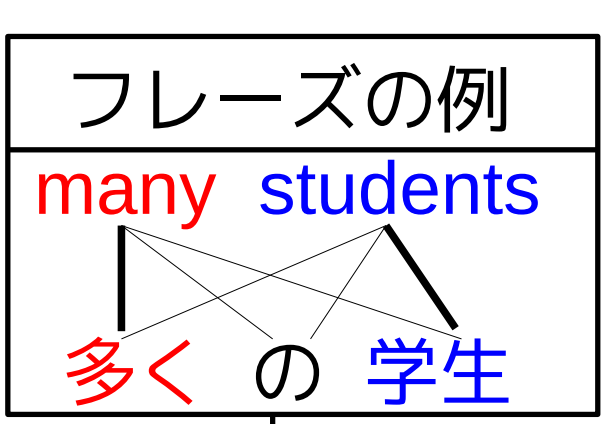
提案手法の概要(句レベルPBSMT)



③フレーズ辞書


対訳文
Your friend is trusted to many students .
あなた の 友達 は 多く の 学生 に 信用 さ れ て いる 。
単語に基づく文パターン辞書の一部
X1 is trusted to X2 .
X1 は X2 に 信用 さ れ て いる 。
フレーズ辞書
Your friend
many students
あなた の 友達
多く の 学生

③フレーズ対数確率(α)

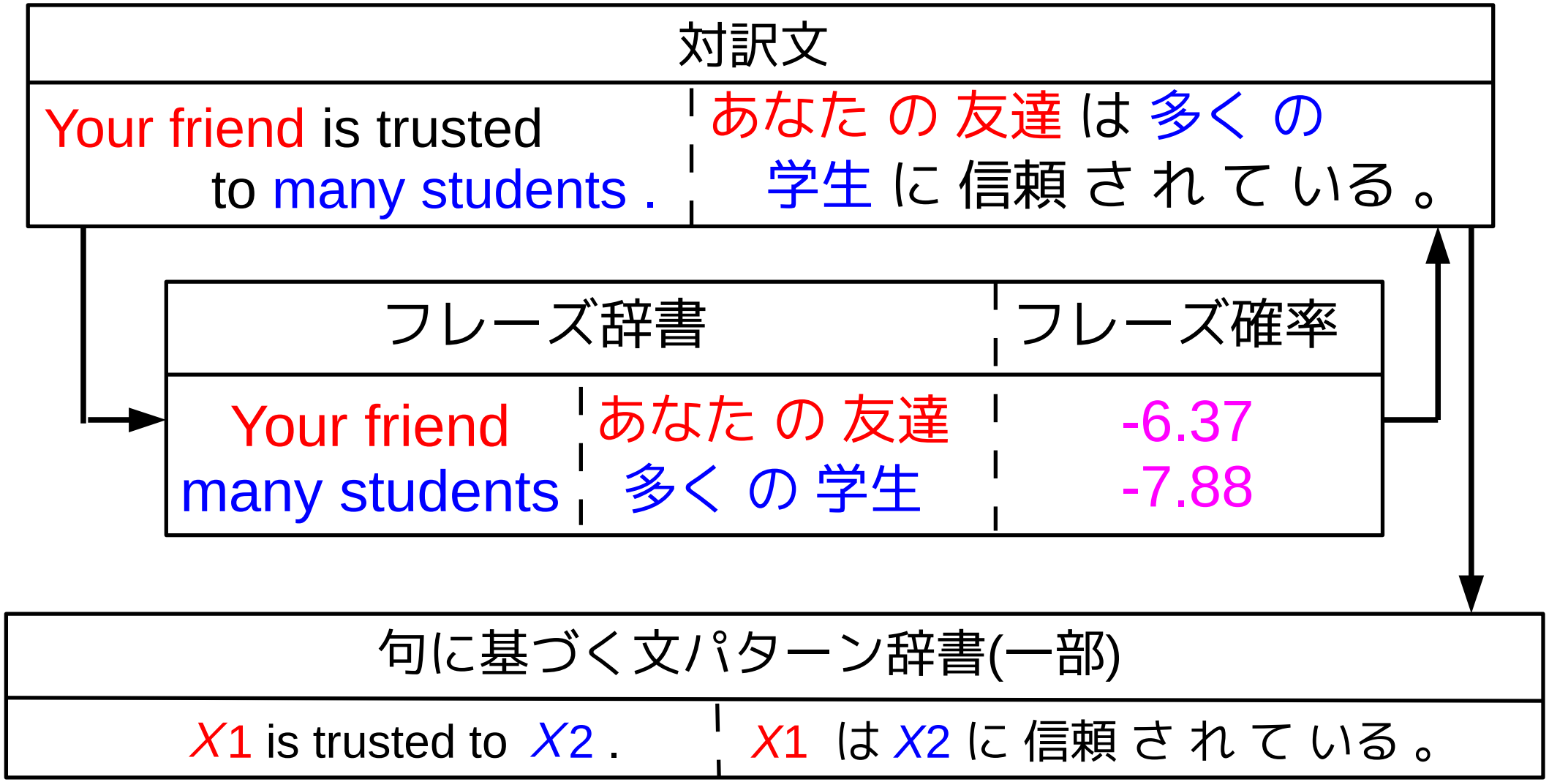


GIZA++		単語確率
many	多く	0.81
many	の	0.03
many	学生	なし
students	多く	0.008
students	の	0.02
students	学生	0.86

フレーズ対数確率(α)の計算

フレーズ対数確率 (α) = $\log_2$(“many | 多く” の翻訳確率)
+ $\log_2$(“students | 学生” の翻訳確率)
= -7.88

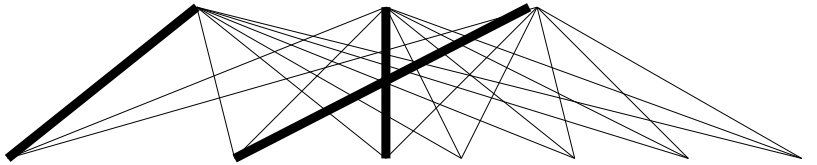
④句に基づく文パターン辞書



④文パターン対数確率(β)

句に基づく文パターンの例

*X*1 is trusted to *X*2 .



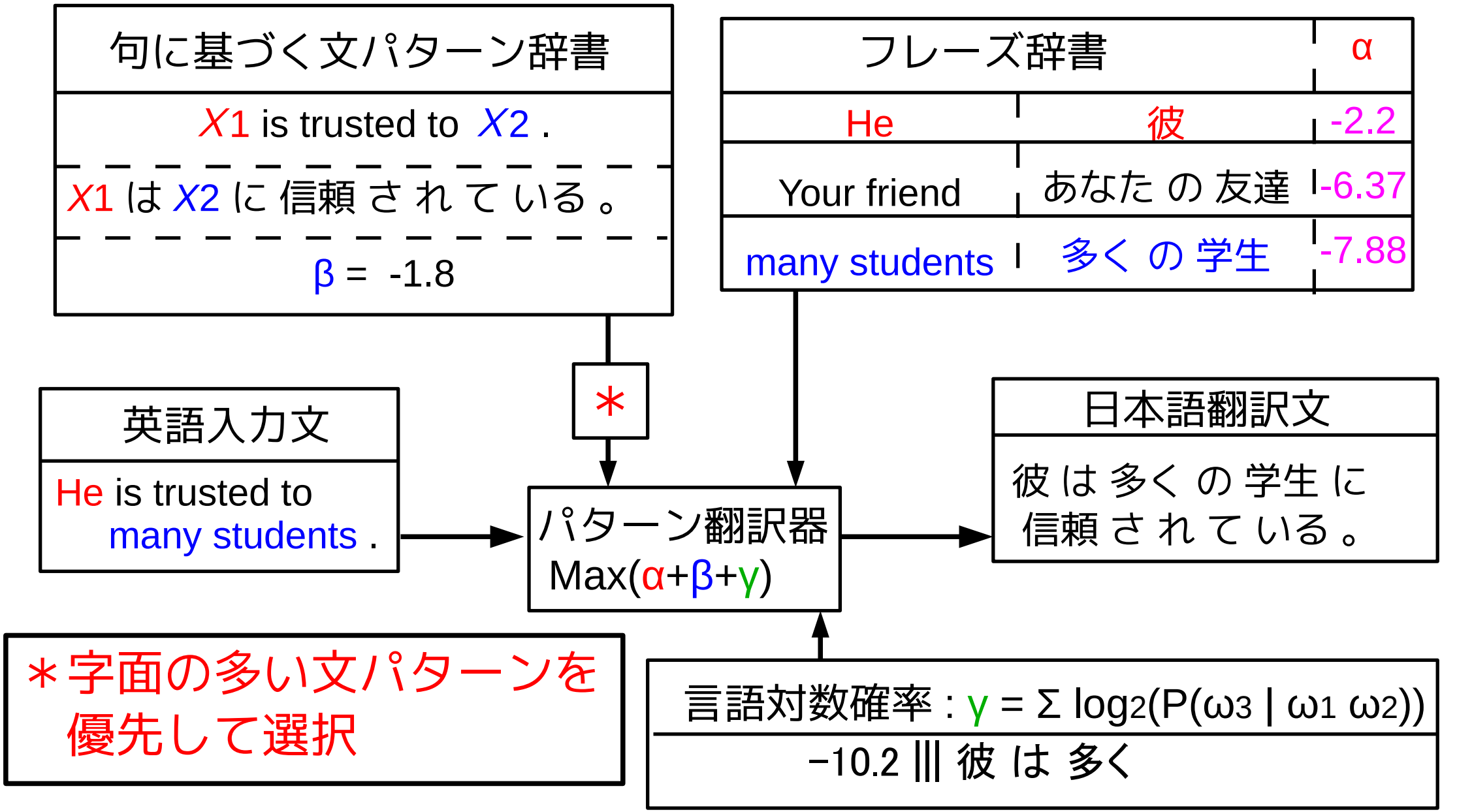
*X*1 は *X*2 に 信頼 さ れ て いる 。

文パターン対数確率(β)の計算

文パターン対数確率(β)
= log2 (“is | は” の翻訳確率
= log2 (“trusted | 信頼” の翻訳確率)
= log2 (“to | に” の翻訳確率)
= -0.44-1.02-0.32
= -1.8

GIZA++		単語確率
is	は	0.21
is	に	0.51
is	信頼	0.12
is	さ	0.07
⋮	⋮	⋮
trusted	は	0.21
trusted	に	0.51
trusted	信頼	0.12
⋮	⋮	⋮
to	は	0.21
to	に	0.51
to	信頼	0.12
⋮	⋮	⋮

⑤DECODER(句レベル)



実験条件・結果（辞書の作成）

入力文(単文) : 対訳文 100,000文

単語辞書	21,439単語
単語に基づく文パターン辞書	16,385,504パターン
フレーズ辞書	230,985,771フレーズ (8883パターンより)
句に基づく文パターン辞書	4,509,385,260パターン （フレーズ辞書2,000,000より）

（とにもかくにも，大量）

日英翻訳 （句レベルPBSMT）

入力文 : 日本語入力文 10文

実験結果

→ 9文の英語翻訳文

カバー率：かなり良い.
翻訳精度：そこそこ？かなり良い？
(mosesより確実に良い)

Sentence_id = 0 "信号 が 青 より 赤 に 変わっ た 。 "
Reference "The signal changed from green to red . "
Pattern_id = 5792 "N00 が 青 N01 N02 に N03 た 。 "
Pattern_id = 5792 "The N00 jumped N01 N03 to N02 . "
Original JP = "信号 が 青 から 赤 に ぱっと 変わっ た 。 "
Original EN = "The traffic light jumped from green to red . "
Decode_id = 0 "< 信号 > が 青 < より > < 赤 > に < 変わっ > た 。 "
Output "The traffic light jumped from green to red . "
TAG 11 "The < traffic light > jumped < from > < green > to < red > . "
TRIGRAM -20.001958 VARIABLE -3.906401 PATTERN -3.701207 SUM -27.609567
PATTERN -20.625607 -5.803825 1 1 1 2 3 1 0.888900
N00 "信号 " "traffic light " -5.022379 -13.986114 57 6 6
221416 7582 1986 73 2 2 4.848608
N01 "より " "from " -5.468747 -8.332586 568 4125 48
1695376 22028893 3509 50 0 0 29.067371
N02 "赤 " "red " -0.830075 -2.160989 53 154 38
181621 1076976 12431 56 94 42 27.245856
N03 "変わっ " "green " -7.864134 -6.442975 169 64 3
917629 323078 1117 0 40 0 2.522527

Sentence_id = 0 "彼 は 仕事 で 京都 へ 行っ た 。 "
Reference "He went to Kyoto on business . "
Pattern_id = 666 "彼 は 仕事 N00 N01 行っ た 。 "
Pattern_id = 666 "He went N01 N00 business . "
Original JP = "彼 は 仕事 で ロンドン へ 行っ た 。 "
Original EN = "He went to London on business . "
Decode_id = 0 "彼 は 仕事 < で 京都 > < へ > 行っ た 。 "

Output "He went to Kyoto on a business . "

TAG 10 "He went < to > < Kyoto on a > business . "
TRIGRAM -31.635600 VARIABLE 46.966509 PATTERN -28.411444 SUM -13.080535
PATTERN -26.412138 -5.416448 1 7 1 1 67 1 0.777800
N00 "で 京都 " "Kyoto on a " -4.935776 -10.092929 6 2 2 38382 19369 735 0 0 0 1.963585
N01 "へ " "to " -1.436010 -4.722648 1707 15515 936 5630098 72416789 184120 0 0 0 634.525208

Sentence_id = 0

"エイズ の 確実 な 治療 法 は まだ わかっ て い ない 。 "

Reference "No sure cures for AIDS are known yet . "

Pattern_id = 206 "N00 N01 N02 N03 は まだ わかっ て い N04 。 "

Pattern_id = 206 "We do N04 know N00 N03 N02 the N01 yet . "

Original JP = "その 殺人 事件 の 全ぼう は まだ わかっ て い ない 。 "

Original EN = "We do not know the whole story of the murder case yet . "

Decode_id = 0 "< エイズ の > < 確実 > < な > < 治療 法 > は まだ わかっ て い < ない > 。 "

Output

"We do not know much of a cure for that of the world yet . "

TAG 17 "We do < not > know < much of a > < cure for that > < of > the < world > yet . "

TRIGRAM -154.400645 VARIABLE 5.746834 PATTERN -9.370785 SUM -158.024597

PATTERN -24.586750 -14.907848 1 1 1 27 6 1 0.615400

N00 "エイズ の " "much of a " -102.186662 -13.372556 8 9 1 77040 6015 88 3 0 0 1.064933

N01 "確実 " "world " -5.087451 -9.111390 24 550 1 68101 6894238 3 13 133 0 1.002868

N02 "な " "of " -6.631931 -8.376980 5857 19451 1445 16487237 111906577 199042 8 0 0 977.337631

N03 "治療 法 " "cure for that " -9.033083 -112.001839 13 1 1 35943 1644 19 15 0 0 0.949617

N04 "ない " "not " -2.839885 -1.545133 6684 2340 1377 34780004 8427214 407899 70 0 0 968.787064

Sentence_id = 0 "ぶらんこ が 揺れ て いる 。 "
Reference "The swing is swinging . "
Pattern_id = 2554 "N00 が N01 て いる 。 "
Pattern_id = 2554 "The N00 are N01 . "
Original JP = "枝 が 風 に 揺れ て いる 。 "
Original EN = "The branches are swaying about in the wind . "
Decode_id = 0 "< ぶらんこ > が < 揺れ > て いる 。 "
Output "The little girls are swinging . "
TAG 8 "The < little girls > are < swinging > . "
TRIGRAM -118.931569 VARIABLE -7.870360 PATTERN -49.528231 SUM -176.330160
PATTERN -9.898090 -4.731406 756 315 68 2364 1336 69 0.833300
N00 "ぶらんこ " "little girls " -100.000000 -200.000000 3 1 1 30494 933 49 1 0 0 0.682555
N01 "揺れ " "swinging " -4.614721 -2.000000 84 10 7 570222 46890 1725 39 6 1 5.361762

Sentence_id = 0 "ぶらんこ が 揺れ て いる 。 "
Reference "The swing is swinging . "
Pattern_id = 2554 "N00 が N01 て いる 。 "
Pattern_id = 2554 "The N00 are N01 . "
Original JP = "枝 が 風 に 揺れ て いる 。 "
Original EN = "The branches are swaying about in the wind . "
Decode_id = 0 "< ぶらんこ > が < 揺れ > て いる 。 "
Output "The little girls are swinging . "
TAG 8 "The < little girls > are < swinging > . "
TRIGRAM -118.931569 VARIABLE -7.870360 PATTERN -49.528231 SUM -176.330160
PATTERN -9.898090 -4.731406 756 315 68 2364 1336 69 0.833300
N00 "ぶらんこ " "little girls " -100.000000 -200.000000 3 1 1 30494 933 49 1 0 0 0.682555
N01 "揺れ " "swinging " -4.614721 -2.000000 84 10 7 570222 46890 1725 39 6 1 5.361762

Pattern Based SMTの現状の問題

フレーズ辞書

変なフレーズが生成

ぶらんこ　|||　He

ぶらんこ ||| of ||| -100.000000 -100.000000 3 19451 2 1878 24020493 72 1 0 0 147.872000 5758.645016 0.791450
ぶらんこ ||| of the ||| -100.000000 -200.000000 3 5694 2 1878 7186804 42 1 2 0 96.662000 2024.412268 0.315356
ぶらんこ ||| of the circus ||| -100.000000 -300.000000 3 1 1 1878 307 1 1 0 0 56.832000 573.553112 -0.774277
ぶらんこ ||| of the circus are ||| -100.000000 -400.000000 3 1 1 1878 282 1 1 0 0 19.847000 733.590115 -1.151195
ぶらんこ ||| of the swing ||| -2.321928 -204.000000 3 1 1 1878 207 21 1 0 0 -28.040477 361.103121 0.194947
ぶらんこ ||| of the swing with ||| -2.321928 -304.000000 3 1 1 1878 147 11 1 0 0 -65.025477 521.139427 -0.192271

考察

Pattern Based SMT とHierarchical SMTの違い

思想　　文全体を考慮　vs　要素合成
(ネットワーク vs 木構造)
パラメータ　　少ない　　vs　多い
パラメータの信頼性
高い　　vs　低い
学習方法　　明確　　vs　一部ヒューリステック
(一部ヒューリステック)　　(Inside Outside)

今後の課題

生成される句（フレーズ）の数と精度
（手で作成された句との比較）

生成される句パターンの数と精度
（原文とのLevenshtein Distance）

翻訳速度

まとめ

統計的手法を用いたパターン翻訳 (PBSMT)

コスト：高い
→ 単語辞書と文パターンを自動作成

翻訳精度 ⟷ カバー率

→翻訳精度：字面の多い文パターンの選択
→カバー率：大量の単語辞書と文パターンの作成

高い翻訳精度