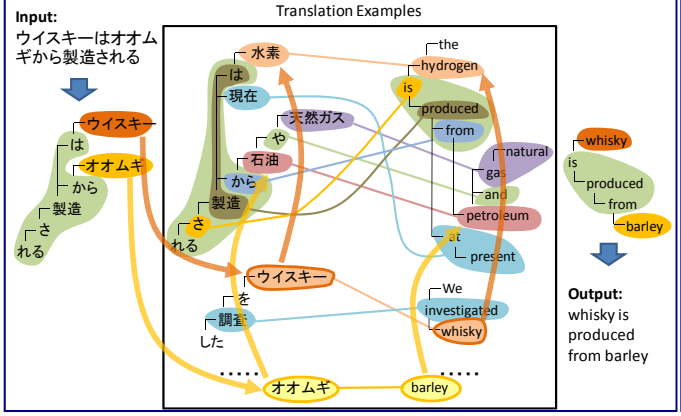
Fully Syntactic EBMT System of KYOTO Team in NTCIR-8

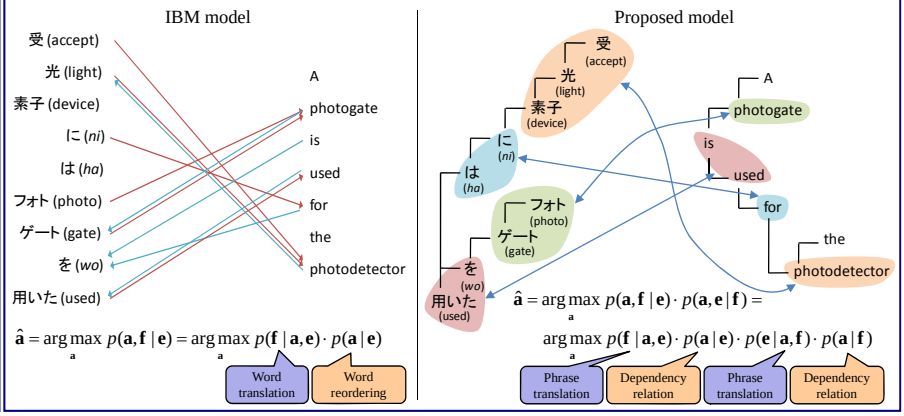
Toshiaki Nakazawa, Sadao Kurohashi

Graduate School of Informatics, Kyoto University

System Description



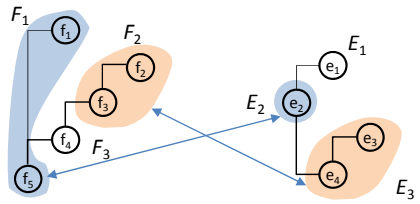
Alignment Model



Model Training

- Step 1: Estimate word translation prob.
  - p(琉球 | Ryukyu) = 0.7
  - p(大学 | university) = 0.6
  - Word base
- Step 2: Estimate phrase translation prob. and dependency relation prob.
  - E-step
    1. Create initial alignment
    2. Modify the alignment by hill-climbing
  - Generate possible phrases
  - M-step: Parameter estimation
  - p(c) = 0.4, p(c;c) = 0.3, p(p) = 0.2
  - Tree base
  - p(琉球 | Ryukyu) = 0.7, p(大学 | university) = 0.6, p(琉球 大学 | university of the Ryukyus) = 0.9

Phrase Translation Probability



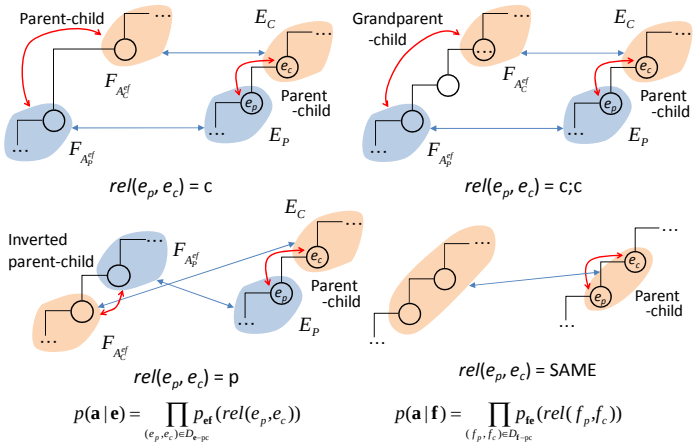
$\mathbf{A}^{fe}$:
$A_1^{fe}=2, A_2^{fe}=3, A_3^{fe}=0$

$\mathbf{A}^{ef}$:
$A_1^{ef}=0, A_2^{ef}=1, A_3^{ef}=2$

$$p(\mathbf{f}\,|\,\mathbf{a},\mathbf{e}) = \prod_{j=1}^{N} p(F_j\,|\,E_{A_j^{fe}}) = p(F_1\,|\,E_2)\cdot p(F_2\,|\,E_3)\cdot p(F_3\,|\,\mathrm{NULL})$$

$$p(\mathbf{e}\,|\,\mathbf{a},\mathbf{f}) = p(E_1\,|\,\mathrm{NULL})\cdot p(E_2\,|\,F_1)\cdot p(E_3\,|\,F_2)$$

Dependency Relation Probability

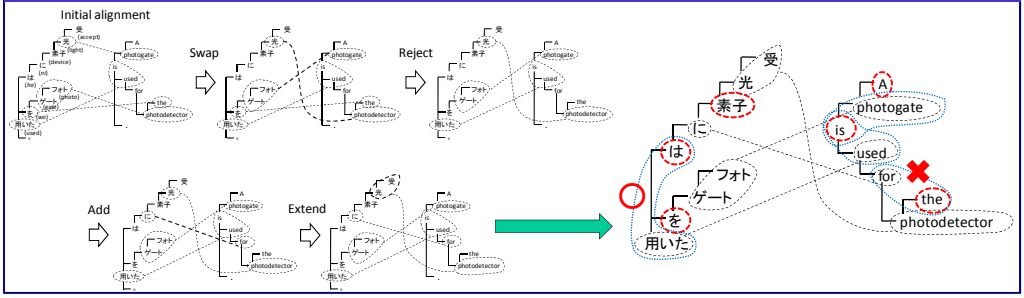


$rel(e_p, e_c) = \mathrm{c}$ $\quad rel(e_p, e_c) = \mathrm{c;c}$

$rel(e_p, e_c) = \mathrm{p}$ $\quad rel(e_p, e_c) = \mathrm{SAME}$

$$p(\mathbf{a}\,|\,\mathbf{e}) = \prod_{(e_p,e_c)\in D_{e\text{-}pc}} p_{\mathbf{ef}}(rel(e_p,e_c)) \qquad p(\mathbf{a}\,|\,\mathbf{f}) = \prod_{(f_p,f_c)\in D_{f\text{-}pc}} p_{\mathbf{fe}}(rel(f_p,f_c))$$

Possible Phrase Generation



Alignment Evaluation Results (Ja-En)

	Pre.	Rec.	AER
Step 1	86.20	44.54	41.24
Step 2 – 1	84.87	49.20	37.48
Step 2 – 2	86.19	58.71	29.98
Step 2 – 3	85.43	63.36	27.05
Step 2 – 4	82.68	65.31	26.83
Step 2 – 5	76.83	66.36	28.64
intersection	90.59	45.53	39.34
grow-final-and	80.00	60.48	31.00
grow-diag-final-and	77.86	61.93	30.92

Translation Evaluation Results

Direction	Training Corpus	Parsing Accuracy	Kyoto-U		RBMT		Moses	
			BLEU	Adeq.	BLEU	Adeq.	BLEU	Adeq.
Ja->En	About 1M	90% (En) 90% (Ja)	19.2	3.60	11.6	3.64	20.9	3.35
En->Ja			21.2	---	10.3	---	21.4	---
Ja->Zh	About 600K	70% (Zh) 90% (Ja)	14.6	---	8.9	---	17.2	---
Zh->Ja			16.4	----	6.9	---	21.8	---

Translation Samples

Input: 以下、添付図面に即してこの発明に係るコネクタの保持構造の最良の形態について説明する。
Output: A description will be hereinafter given of a best mode a holding structure of a connector according to the present invention with reference to the attached drawings .

Input: しかし、このような処理を行うためには、パイプライン式算術符号復号器52に連続して文脈インデックスを供給しなければならない。
Output: Context index must continuously , however be supplied to order such processing to the pipeline system arithmetic code decoder 52 .

Input: This invention relates to a purification method for organometallic compounds, especially trimethylaluminum.
Output: 本 発明 は 有機 金属 化合 物 特に トリメチル の 浄化 方法 に 関する 。

Input: FIG. 4 is a plan view for illustrating the dimensional shape of waveguides used to simulate the stability of the optical intensity in the case of making the parallel portion 12 short and in the case of making the portion 12 long.
Output: 図4はこの平行部12を短くすること場合及び例えば部12を長くする場合の光学濃度の安定性をシミュレートする用いられる導波路の次元形状を説明用の平面図である。

NTCIR-8 Results

	Intrinsic (BLEU)		Extrinsic		
	JE	EJ	BLEU	MAP	Recall@100
KYOTO	22.22	24.29	17.25	0.1909	0.5258
Moses	29.08	35.27	24.01	0.1943	0.5701