Kyoto-U: Syntactical EBMT System for NTCIR-7 Patent Translation Task

Toshiaki Nakazawa, Sadao Kurohashi

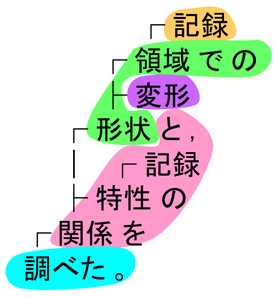
Graduate School of Informatics, Kyoto University

System Overview

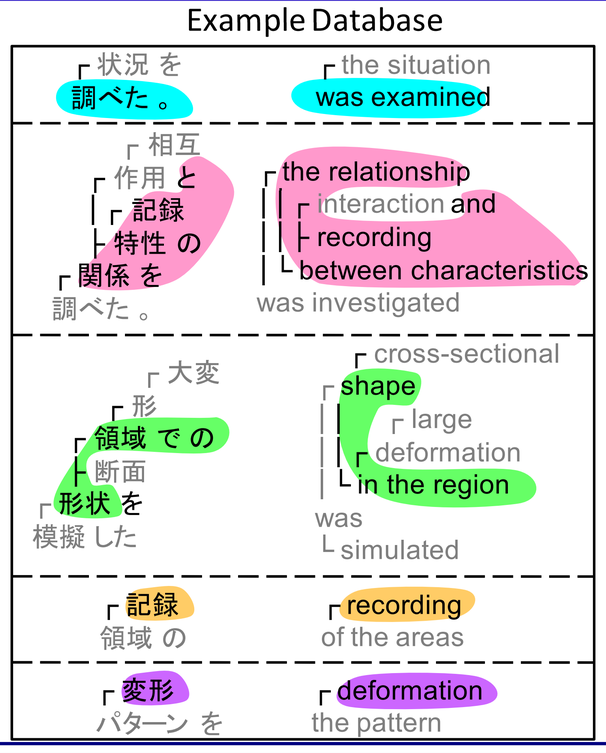
Input:
記録領域での変形形状と，記録特性の関係を調べた。

↓

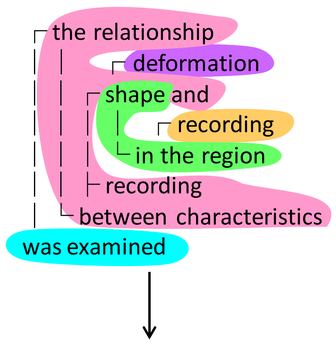
Input Dependency Tree



Example Database



Output Dependency Tree



↓

Output:
The relationship between deformation shape in the recording region and recording characteristics was examined .

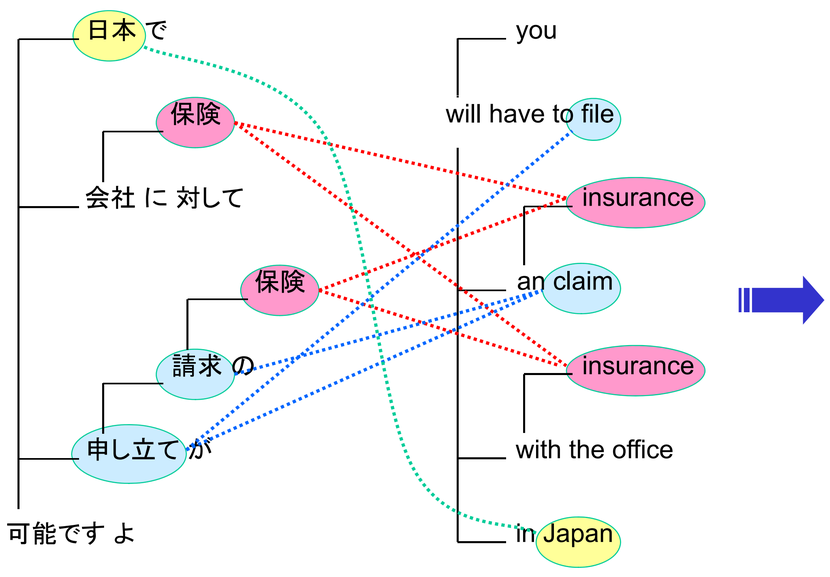
Structure-based Alignment

- **Dependency structure transformation**
  - **Japanese: Morphological analyzer JUMAN and dependency analyzer KNP**
  - **English: Nlparser (by Charniak) and hand-made rules defining head words for phrases**
- **Word/phrase correspondence detection**
  - **bilingual dictionaries**
  - **numeral normalization**
    二百十六万 ⇔ 2,160,000 ⇔ 2.16 million
  - **statistical substring alignment (Cromieres 2006)**
  - **transliteration (Katakana, NE)**
    ローズワイン ⇔ rosuwain ⇔ rose wine
    新宿 ⇔ shinjuku ⇔ shinjuku

$$\frac{P(s_j, s_e)}{P(s_j)P(s_e)} > \theta$$

- **Handling remaining words**

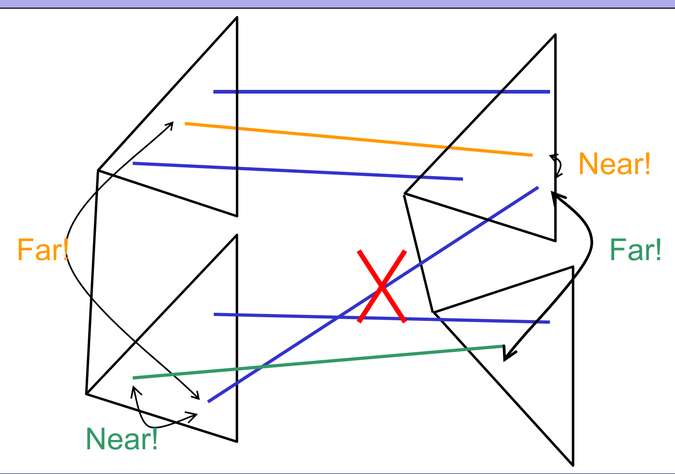
Alignment Disambiguation with Consistency Score & Dependency Type



$$\arg\max_{alignment} \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} cs(d_S(a_i,a_j), d_T(a_i,a_j))}{n(n-1)/2}$$

Consistency Score; Source-side Distance; Target-side Distance

n = # of correspondence candidates



Dependency Type Distance

Japanese		English	
用言:レベルC	6	S/SBAR/SQ …	5
用言:レベルB+ / B	5	VP/WHADVP	4
用言:レベルB- / A	4	WHADJP	
ノ格 / 連体	2	ADVP/ADJP/NP/PP/	3
文節内/用言:レベルA+	1	INTJ/QP/PRT/PRN	
Others	0	Others	1

- ***f*(·): consistency score**
  - ‘near-near’: positive
  - ‘far-far’: 0
  - ‘near-far’/’far-near’: negative
- ***d*(·): distance**
  - dependency type distance

Japanese -> English Intrinsic Evaluation Result

BLEU		Adequacy		Fluency		Average	
27.20	NTT	3.81	tsbmt	4.02	Japio	3.88	tsbmt
27.14	moses	3.71	Japio	3.94	Tsbmt	3.86	Japio
27.14	MIT	3.15	MIT	3.66	MIT	3.40	MIT
25.48	NAIST-NTT	2.96	NTT	3.65	NTT	3.30	NTT
24.79	NICT-ATR	2.85	Kyoto-U	3.55	moses	3.18	moses
24.49	KLE	2.81	moses	3.44	tori	3.10	Kyoto-U
23.10	tsbmt	2.66	NAIST-NTT	3.43	NAIST-NTT	3.04	NAIST-NTT
22.29	tori	2.59	KLE	3.35	Kyoto-U	3.01	tori
21.57	Kyoto-U	2.58	tori	3.28	HIT2	2.94	KLE
19.93	mibel	2.47	NICT-ATR	3.28	KLE	2.86	HIT2
19.48	HIT2	2.44	HIT2	3.09	mibel	2.78	NICT-ATR
19.46	Japio	2.38	mibel	3.08	NICT-ATR	2.74	mibel
15.90	TH	1.87	TH	2.42	FDU-MCandWI	2.13	TH
9.55	FDU-MCandWI	1.75	FDU-MCandWI	2.39	TH	2.08	FDU-MCandWI
1.41	NTNU	1.08	NTNU	1.04	NTNU	1.06	NTNU

Translation Result Example (BLEU: 21.62)

Input: 図 ４ に 示した メモリ アレイ の 配置 を 採用 する こと で 、下位 側 データバス ６２ および 上位 側 データバス ６４ は 、それぞれ 総 延長 を ５Ｌ に する こと が できる 。

Output: By adopting the arrangement shown in FIG. 4 of the memory array , data lower bus 62 side data bus 64 can be made a total length between can be elongated respectively into the 5L .

Ref: The use of the memory-array arrangement shown in FIG . 4 allows each of a lower data bus 62 and an upper data bus 64 to have the total length of 5L .

English -> Japanese Intrinsic Evaluation Result

BLEU		Adequacy		Fluency		Average	
30.58	moses	3.53	tsbmt	3.69	moses	3.60	tsbmt
29.15	NICT-ATR	2.90	moses	3.67	tsbmt	3.30	moses
28.07	NTT	2.74	NTT	3.54	NTT	3.14	NTT
22.65	Kyoto-U	2.59	NICT-ATR	3.20	NICT-ATR	2.89	NICT-ATR
17.46	tsbmt	2.42	Kyoto-U	2.54	Kyoto-U	2.48	Kyoto-U

- After resolving the defect of not caring whether a child node is a pre-child or post-child, the BLEU score rose to 24.02 from 22.65.

Translation Result Example (BLEU: 24.11)

Input: in FIG. 3A which corresponds to Example 1 the crowning shape is set in the vicinity of the lower limit

Output: 下限 近傍 に 実施 例 １ に 対応 する 図 ３ クラウン 形状 は 、設定 されて いる 。

Ref: 実施 例 １ に 相当 する 図 ３ａ で は 、クラウニング 形状 を 下限 近傍 に 設定 した 。

Conclusion

- Translation result showed that our EBMT system is competitive to the state-of-the-art SMT systems
- Using syntactical information must be useful for structurally different language pairs such as Japanese and English
- Patent sentences often have typical expressions, mathematical or chemical formulas and so on, so we may need to adopt some pre-processes to avoid parsing errors to handle such peculiar expressions properly