

# NAIST-NTT System Description for Patent Translation Task at NTCIR-7 -- Semi-supervised Learning of Bilingual Lexicon for Technical Terms from Wikipedia --

Mamoru Komachi (NAIST, Japan), Masaaki Nagata (NTT, Japan), Yuji Matsumoto (NAIST, Japan)

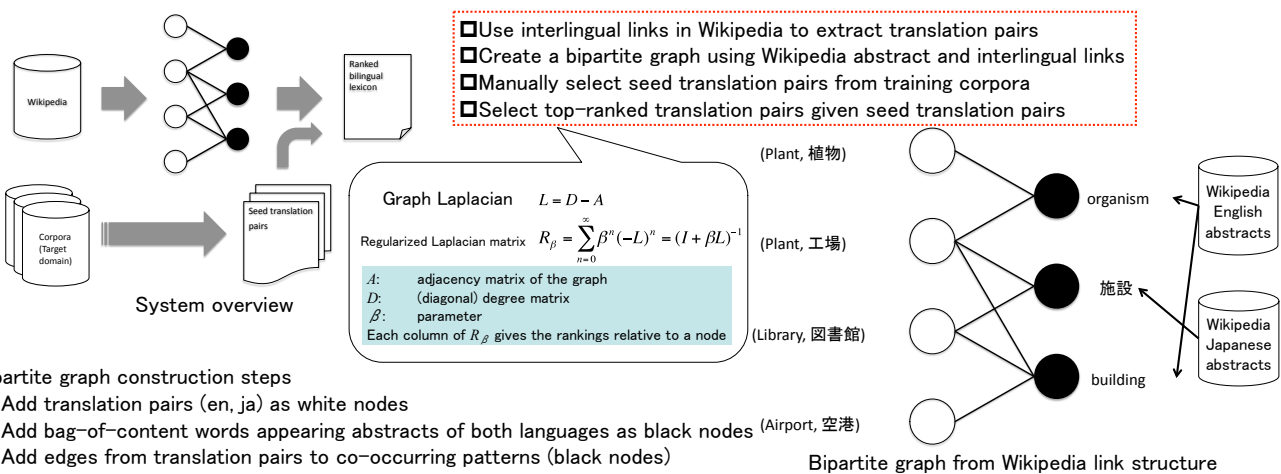
## Background

- The cost of hand-tagging resources is crucial for MT
  - Wikipedia has become one of the important source for knowledge acquisition
  - Wikipedia's link structure has attracted attention in NLP fields, and the link structure turns out to be useful for extracting bilingual lexicon from Wikipedia [Adafre and Rijke 2006; Erdmann 2008]
- This study aims at:
- Domain adaptation (word sense disambiguation per domain) of term extraction
  - Automatic refinement of translation pairs extracted from Wikipedia

## Approach

Assumption: one sense per domain [Thelen and Riloff 2002]

→ Use "relatedness" measure in link analysis to select the most relevant sense to the domain



## Experiment

- Extract bilingual lexicon from Wikipedia: 222,739 translation pairs (197,770 after filtering)
- Split the bilingual lexicon randomly into 8 sub-lexicons, and choose 5 seeds for each sub lexicon (total  $8 * 5 = 40$  seeds)
- After applying the regularized Laplacian kernel, collect the top 10%, 50% and 75% of the ranked list for each sub lexicon
- Accumulate sub lexicons by taking the intersection of the 8 collected lists to obtain final bilingual lexicon

Random samples of extracted lexicon

Wikipedia	# of words (OOV coverage)	samples
10%	11,970 (1.9%)	(natural selection, 自然選択説), (scrabble, スクラブル), (phase transition, 相転移), (diamond, ダイアモンド), (videocassette recorder, ビデオテープレコーダ)
50%	75,420 (7.7%)	(movement for multiparty democracy, 複数政党制民主主義運動), (fentanyl, フェンタニル) [an opioid analgesic], (sigma sagittarii, シュタキ) [the second brightest star system in the constellation Sagittarius], (shintaro abe, 安倍晋太郎) [the former prime minister of Japan], (nippon television, 日本テレビ放送網)
75%	113,277 (11.5%)	(pride final conflict 2003, pride grandprix 2003 決勝戦) [a mixed martial arts event held by PRIDE Fighting Championships], (uglyness, 醜), (palma il vecchio, バルマイイル・ヴェッキオ) [an Italian painter], (jean gilles, ジャン・ジール) [a French composer; a French soldier], (amiloride, アミロライド) [a potassium-sparing diuretic]
100%	197,770 (13.5%)	(brilliant corners, ブリリアント・コーナース) [an album by a jazz musician], (charly mottet, シャーリー・モテ) [a French former professional cyclist], (deep purple in rock, ディープパープル・インロック) [an album by an English rock band], (june 2003, 「最近の出来事」2003年6月) [navigational entry for events happened in June 2003], (moanin', モーニン) [a jazz album]
filtered	24,969	(1.1年) [year], (UTC+9, UTC+9) [Japanese side contains only alphanumeric characters], (Aera, AERA) [case-insensitive match] (大岡越前, 大岡越前) [garbage in English side], (image:himeji castle frontview.jpg, himeji castle frontview.jpg) [Wikipedia format navigational links], (user:eririninrin, eririninrin) [Wikipedia specific entries]

Performance of Wikipedia dict.

	Single-ref		Fmlrun-int	
	JE	EJ	JE	EJ
Baseline (WMT08)	26.39	28.25	25.34	27.19
Wikipedia (10%)	N/A	27.47	N/A	N/A
Wikipedia (50%)	N/A	27.46	N/A	N/A
Wikipedia (75%)	N/A	27.42	N/A	N/A
Wikipedia (100%)	26.48	27.28	25.48	28.15

□ Dataset: NTCIR-7 Patent Translation Task (1.8 million parallel sentences)

□ Add the extracted bilingual lexicon to the training corpus to learn the translation probability between translation pairs

□ The extracted bilingual lexicon slightly improves BLUE score

□ However, increasing the extracted lexicon constantly degrades BLEU score for EJ on single-reference setting → Need re-examination

Sample seeds

English	Japanese
Thermal spray	溶射
Epoxy	エポキシ樹脂
Single crystal	単結晶
Laser cooling	レーザー冷却
Centrifugal compressor	遠心圧縮機

## Conclusion and Future Work

- Demonstrated that a large scale bilingual dictionary can be extracted from Wikipedia
- Improved the quality of extracted bilingual lexicon by applying a graph kernel
- Plan to integrate graph-based word sense disambiguation into statistical machine translation framework