Expanding Paraphrase Lexicons by Exploiting Lexical Variants

Atsushi Fujita (NICT, Japan; *fujita@paraphrasing.org*) and Pierre Isabelle (NRC, Canada)

- Given a paraphrase lexicon → Obtain a large number of new paraphrase pairs
  - Generalization: Obtain paraphrase patterns by generalizing corresponding words
  - Instantiation: Collect phrases that match the paraphrase patterns from monolingual corpus
- Promising results
  - Large coverage: 100x expansion is possible, although the rate depends on the size of corpora
  - High accuracy: comparable to state-of-the-art, without harnessing rich resources

Motivation

- How to create a paraphrase lexicon with both high coverage and high accuracy?

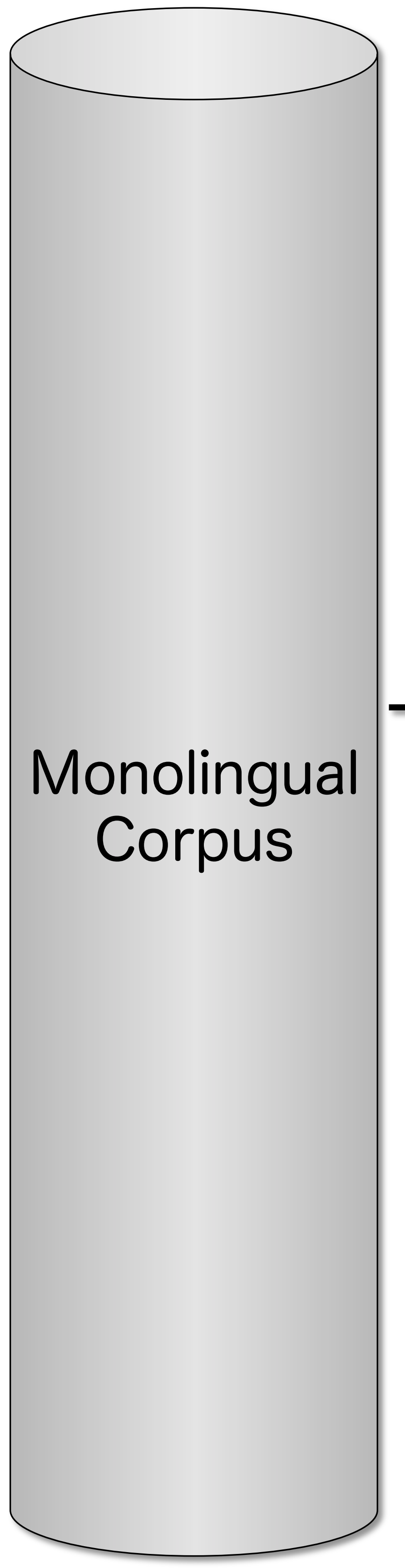
	Corpus	Method	Coverage	Accuracy
(A)	Monolingual	Dist. sim. e.g., [Lin+, 01]	☺ Largest	☹ Lower than the others
(B)	Bilingual/multilingual parallel	Pivoting e.g., [Bannard+, 05]	☹ Limited	☺ Relatively high
(C)	Monolingual parallel	Alignment e.g., [Barzilay+, 01]	☹ Limited	☺ Relatively high
(D)	Monolingual comparable	Alignment e.g., [Hashimoto+, 11]	☹ Limited	☺ Relatively high

Need clean up

Need expansion

- Prefer unsupervised and language-independent way

Exploiting Lexical Correspondences within Paraphrases



S_{Seed}: seed paraphrase pairs

airports in Europe ⇔ European airports
amendment of regulation ⇔ amending regulation
should be noted that ⇔ is worth noting that

↓ *Step 1. Learning Paraphrase Patterns*

Paraphrase patterns

X:ϕ in *Y*:ϕ ⇔ *Y*:an *X*:ϕ
X:ment of *Y*:ϕ ⇔ *X*:ing *Y*:ϕ
should be *X*:ed that ⇔ is worth *X*:ing that

↓ *Step 2. Harvesting New Paraphrase Pairs*

S_{LV}: new paraphrase pairs

cohesion in Europe ⇔ European cohesion
democracy in Europe ⇔ European democracy
increase in Haiti ⇔ Haitian increase
transportation in suburb ⇔ suburban transportation
economy in Uruguay ⇔ Uruguayan economy
amendment of documents ⇔ amending documents
amendment of protocol ⇔ amending protocol
investment of resources ⇔ investing resources
recruitment of engineers ⇔ recruiting engineers
should be highlighted that ⇔ is worth highlighting that
should be reiterated that ⇔ is worth reiterating that
should be stated that ⇔ is worth stating that

1-1. Collecting affix patterns

is aimed at achieving ⇔ aims to achieve

↓ (i) Extracting all candidates

Word$_1$	Word$_2$	Affix$_1$	Affix$_2$	Stem
aimed	aims	X:ed	X:s	aim
aimed	achieve	X:imed	X:chieve	a
achieving	aims	X:chieving	X:ims	a
achieving	achieve	X:ing	X:e	achiev

(ii) Filtering candidates [Gaussier, 99]

Affix$_1$	Affix$_2$	STEM(5)	Retain?
X:chieve	X:imed	0	No
X:chieving	X:ims	0	No
X:ed	X:s	69	Yes
X:ing	X:e	330	Yes

# of unique stems having at least 5 characters in length

1-2. Generating paraphrase patterns

is aimed at achieving ⇔ aims to achieve → [Fujita+, 12] leaves this as is → is aimed at achieving ⇔ aims to achieve → is *X*:ed at *Y*:ing ⇔ *X*:s to *Y*:e

amendment of regulation ⇔ amending regulation → amendment of *X* ⇔ amending *X* → *Y*:ment of *X*:ϕ ⇔ *Y*:ing *X*:ϕ

development of tourism ⇔ developing tourism → development of *X* ⇔ developing *X*

Newly generalized

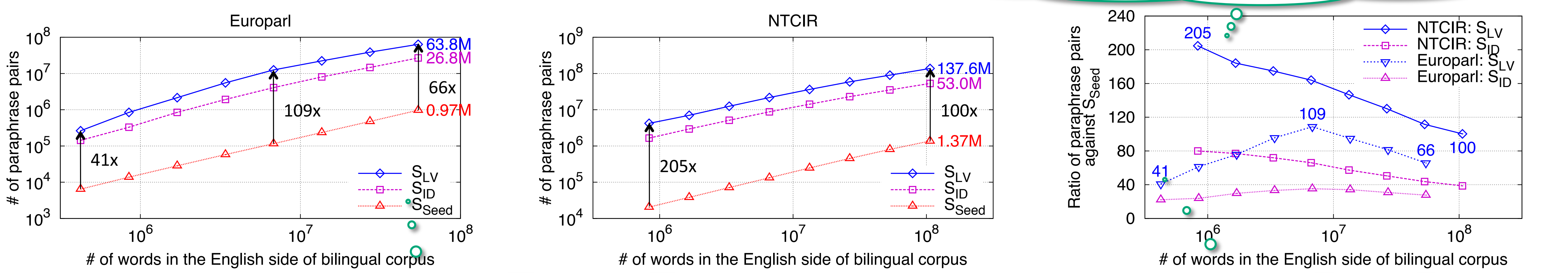
Turned identical

- Searching monolingual corpus for new instances
  - Both lexical and sub-lexical matching
  - High-level parallelization: both corpus and patterns
- Assessing new instance with yet another similarity
  - Adjacent word 1-4 grams of phrase appearances
  - High-level parallelization: both corpus and phrases

Quantity: Large Multiple of Seed Paraphrases

- English experiments using the approach (B) for obtaining S_{Seed}
  - (a) Europarl (En-Fr, 55.7 MW) + News Crawl 2013 (1.20 BW)
  - (b) NTCIR Patent: Parallel (En-ja, 107 MW) + NTCIR En Monolingual 2006-2007 (1.36 BW)
  - Moses, MeCab, SyMGIZA++, UniNe stoplists

High leverage when Bi <<<< Mono



S_{ID}: generalize only identical words [Fujita+, 12]

Leverage is limited in some condition

Quality: Reasonably High

- Human evaluation of paraphrase substitutions
  - Apply paraphrases in (a) to WMT "newstest" data
  - 200 phrases * 3-best paraphrases chosen by 5g LM
  - Unit-wise, 2-phased, classification-based protocol [Fujita, 13]

	N	Grammar	Meaning	Both
S_{Seed}	66	0.85	0.91	0.76
S_{ID}	339	0.84	0.78	0.66
S_{LV}	534	0.74	0.78	0.59
Total	600	0.75	0.79	0.61
CCG+LM [Callison-Burch, 08]		0.68	0.61	0.55

☺ The first decision was given following a complaint by the Quebec Common Front of persons receiving social assistance.

The first decision was given following a complaint by the Quebec Common Front of persons who receive social assistance.

☺ Investment is intended to improve the availability of locomotives and the rail network, even in the face of extreme weather conditions.

Investment is aimed at improving the availability of locomotives and the rail network, even in the face of extreme weather conditions.

☹ The safety issue was considered sufficiently serious for all affected parties to be informed.

The safety issue was sufficient consideration serious for all affected parties to be informed.

☹ Federal Security Service now spread a big network of fake sites and there are tons of potential buyers of military weapons.

Federal Security Service now spread a big network of fake sites and there are tons of a potential buyer of military weapons.

Future Work

- Effective combination of heterogeneous paraphrase lexicons
- Application to further large S_{Seed} and monolingual corpus
- Application to various languages
- Integration into NLP applications, e.g., MT