Bootstrapping A Statistical Speech Translator From A Rule-Based One

Manny Rayner, Paula Estrella
Pierrette Bouillon

Goals of Paper

- “Relearning Rule-Based MT systems”
  - Goal: bootstrap statistical system from rule-based one
  - Question 1: can we do it at all?
  - Question 2: if so, can we add robustness?
  - E.g. Dugast et al 2008 with SYSTRAN
- Can we do it with a small-vocabulary high-precision speech translation system?
  - Key problem: shortage of training data
  - Must also bootstrap statistical speech recognition
  - How do the two components fit together?

Basic Method

(For both recognition and translation)

- Use rule-based system to make training data
- Train on generated data
- Produce statistical version of system

Outline

- Goals of paper
- ➢ MedSLT
- Bootstrapping a statistical recogniser
- Bootstrapping an interlingua-based SMT
- Putting it together
- Conclusions

MedSLT (1)

- Open Source medical speech translator for doctor-patient examinations
- Unidirectional communication (patient answers non-verbally, e.g. nods or points)
- System deployed on laptop/mobile device

English MedSLT examples

where is the pain

is the pain in the front of the head

do you often get headaches in the morning

does bright light give you headaches

do you have headaches several times a day

does the pain last more than an hour

MedSLT (2)

- Multilingual
  - Here, use EN →FR and EN → JP versions
- Medium vocabulary
  - 400-1100 words, depending on language
- Grammar-based: uses Open Source Regulus platform
  - Grammar-based recognition
  - Interlingua-based translation
- Safety-critical application
  - Check correctness before speaking translation
  - Use “backtranslation” to check

Backtranslation

- Source: Do you have headaches at night?
- B/trans: Do you experience the headaches at night?
- Target: Vos maux de tête surviennent-ils la nuit?
- Target: Yoru atama wa itamimasu ka?

Outline

- Goals of paper
- MedSLT
- ➢ Bootstrapping a statistical recogniser
- Bootstrapping an interlingua-based SMT
- Putting it together
- Conclusions

Bootstrapping a Statistical Recogniser

(Hockey, Rayner and Christian 2008)

- Recognition in MedSLT
  - Grammar-based language model
  - built using data-driven method
- Seed corpus used to extract relevant part of resource grammar
- Resulting grammar compiled to CFG form

Two ways to build a statistical recogniser

- Direct
  - Seed corpus → statistical recogniser
- Indirect
  - e.g. (Jurafsky et al 1995, Jonson 2005)
  - Use the grammar to generate a larger corpus
  - Seed corpus →
    grammar →
    corpus →
    statistical recogniser

Refinements to generation idea

- Generate using Probabilistic CFG
  - Better than plain CFG
- “Interlingua filtering”
  - Use interlingua to remove strange sentences

Example: CFG generated data

what attacks of them 're your duration all day

have a few sides of the right sides regularly frequently hurt

where 's it increased

what previously helped this headache

have not any often ever helped

are you usually made drowsy at home

what sometimes relieved any gradually during its night

's this severity frequently increased before helping

when are you usually at home

how many kind of changes in temperature help a history

Example: PCFG generated data

does bright light cause the attacks

are there its cigarettes

does a persistent pain last several hours

is your pain usually the same before

were there them when this kind of large meal helped joint pain

do sudden head movements usually help to usually relieve the pain

are you thirsty

does nervousness aggravate light sensitivity

is the pain sometimes in the face

is the pain associated with your headaches

Example: PCFG generated data with interlingua filtering

does a persistent pain last several hours

do sudden head movements usually help to usually relieve the pain

are you thirsty

does nervousness aggravate light sensitivity

is the pain sometimes in the face

have you regularly experienced the pain

do you get the attacks hours

is the headache pain better

are headaches worse

is neck trauma unchanging

Experiment: CFG/PCFG, different sizes of corpus, filtering

Version	corpus	WER	SER
Grammar-based	948	21.96%	50.62%
Stat, seed corpus	948	27.74%	58.40%
Stat, CFG generation	4281	49.0%	88.4%
Stat, PCFG generation	4281	25.98%	65.31%
Stat, PCFG generation	497 798	24.38%	59.88%
Stat, PCFG, filter	497 798	23.76%	57.16%

Bootstrapping statistical recognisers: conclusions

- Indirect method for building recogniser better than direct one
  - PCFG generation is essential
  - Interlingua filtering gives further small win
- Original grammar-based recogniser still better than all statistical variants

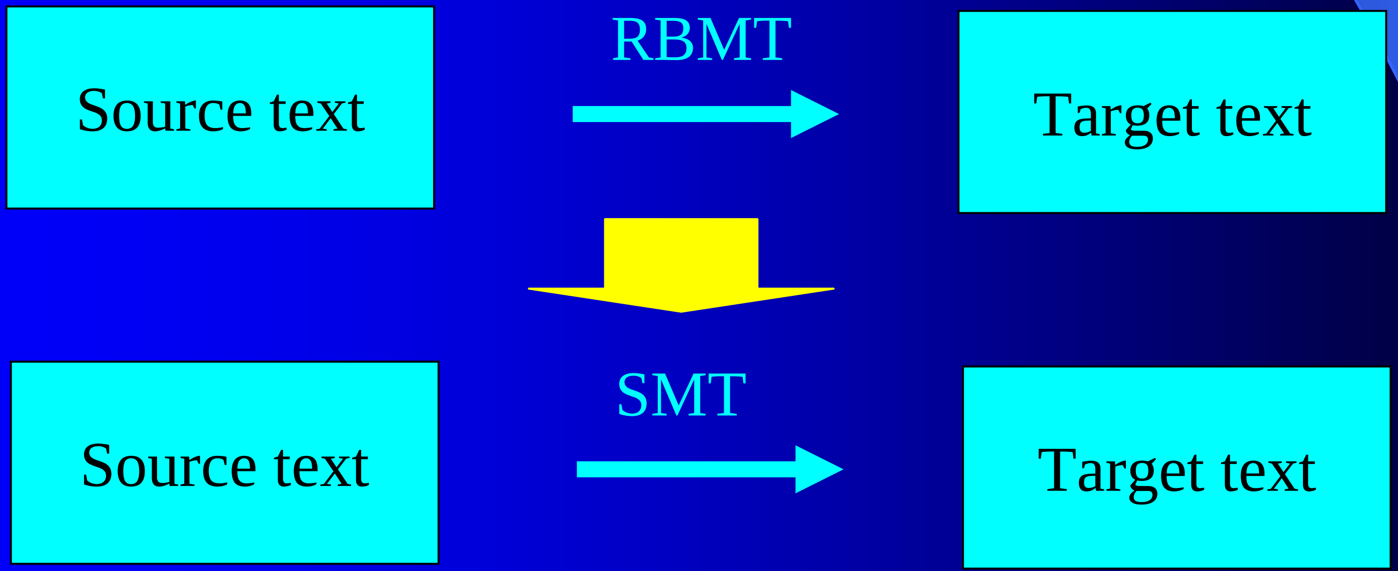
Outline

- Goals of paper
- MedSLT
- Bootstrapping a statistical recogniser
- ➢ Bootstrapping an interlingua-based SMT
- Putting it together
- Conclusions

“Relearning RBMT”

(Rayner, Estrella and Bouillon 2010)

- Similar to recognition: use rule-based system to generate training data



Source text

RBMT

Target text

Source text

SMT

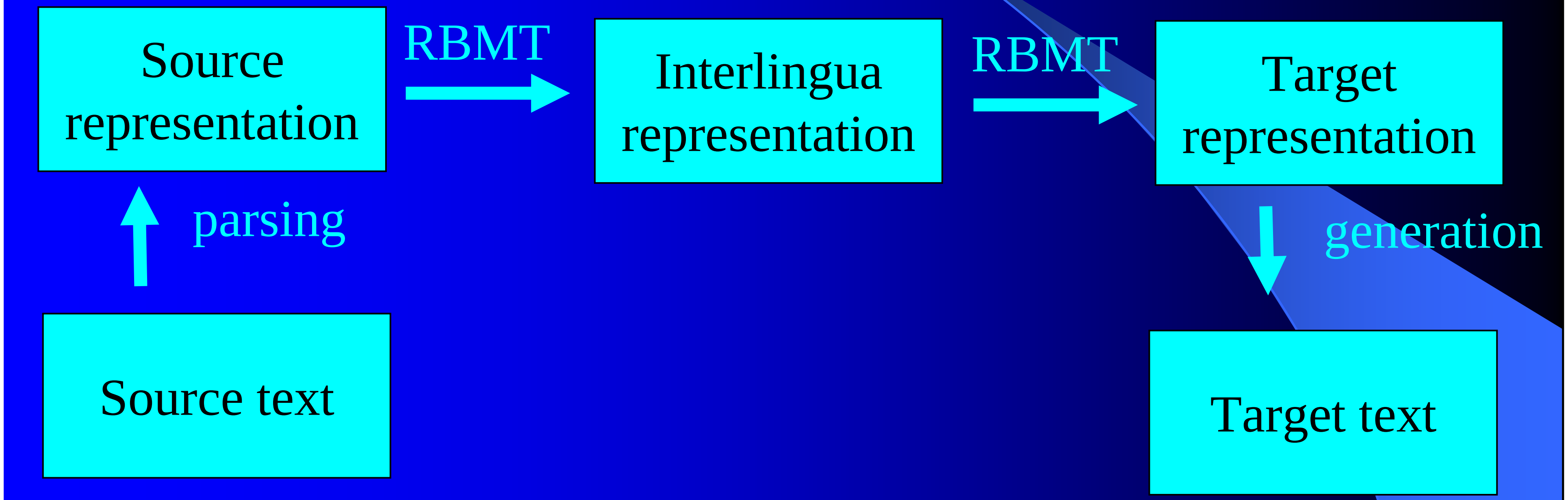
Target text

Naive approach

(Rayner et al 2009)

- Naive approach is unimpressive
- If bootstrapped SMT translation different from RBMT translation, usually wrong
- Very poor for English → Japanese
  - Better for English → French
- Tops out quickly, then no improvement

"Relearning Interlingua-Based Machine Translation"



Source representation

RBMT

Interlingua representation

RBMT

Target representation

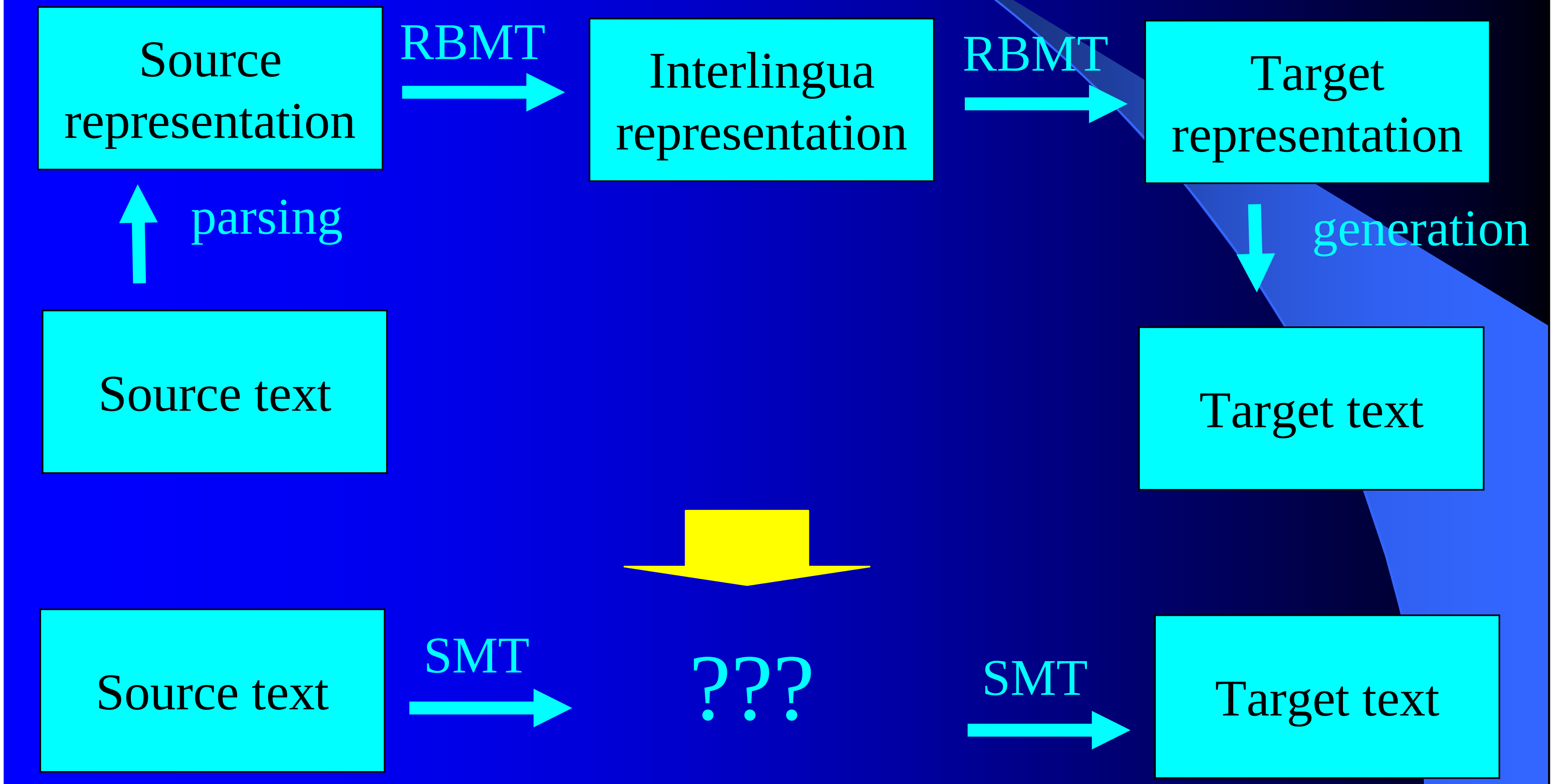
parsing

generation

Source text

Target text

“Relearning Interlingua-Based Machine Translation”



Source representation

RBMT

Interlingua representation

RBMT

Target representation

parsing

generation

Source text

Target text

Source text

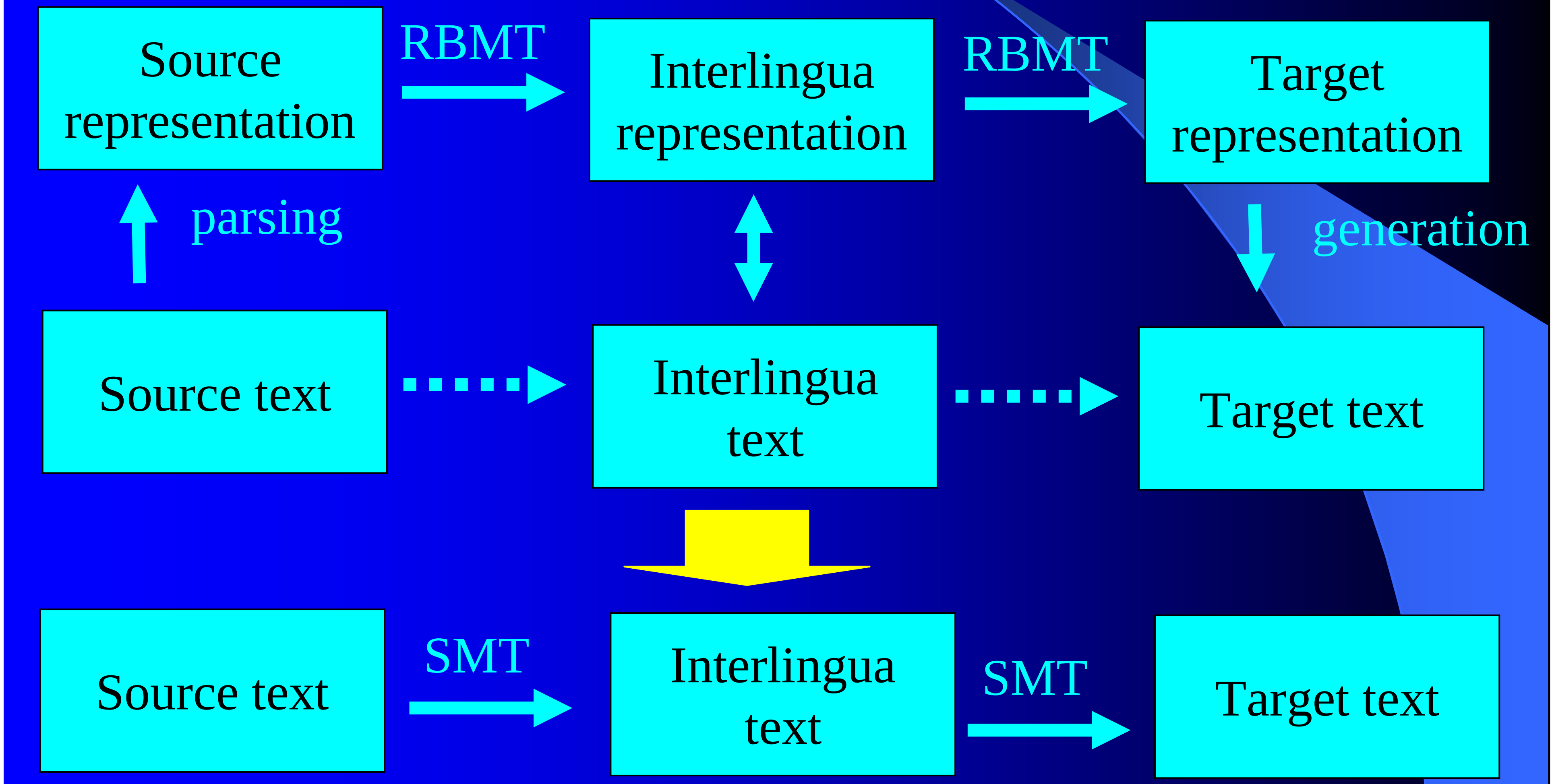
SMT

???

SMT

Target text

"Relearning Interlingua-Based Machine Translation"



Source representation

RBMT

Interlingua representation

RBMT

Target representation

parsing

generation

Source text

Interlingua text

Target text

Source text

SMT

Interlingua text

SMT

Target text

“Interlingua text”

- What is “interlingua text”?
- How can we use it to relearn an interlingua-based system as an SMT?
- Think of interlingua as a language
  - Define using formal grammar
  - Associate text form with representation
  - Text form is simplified/telegraphic English

Interlingua and Text Form

English sentence: “Does the pain spread to the jaw?”

Interlingua representation

[null=[utterance_type,ynq], arg1=[symptom, pain],
null=[state, radiate], null=[tense,present]],
to_loc=[body_part, jaw]]

Interlingua Text (English version)

“YN-QUESTION pain radiate PRESENT jaw”

Can also have versions of interlingua text based on other languages…

Different Forms of Interlingua Text

EN	does the pain last for more than one day
IN/E	YN-QUESTION pain last PRESENT duration more-than one day
JP	ichinichi sukunakutomo itami wa tsuzukimasu ka
IN/J	more-than one day duration pain last PRESENT YN-QUESTION

Bootstrapping an interlingua-based SMT

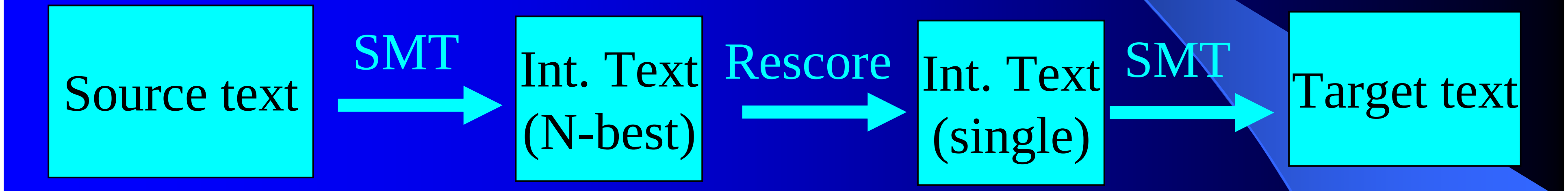
- Randomly generate source data
- Translate using EN-FR and EN-JP RBMT
- Save interlingua in EN and JP text forms
- Train SMT models using Moses etc

Exploiting interlingua text

- Rescoring
  - – Do Source → Interlingua in N-best mode
  - – Prefer well-formed interlingua text
- Reformulation
  - – Split up EN-JP as EN-IN/E + IN/J-JP
  - – SMT translation only between languages with similar word-orders

Processing pipelines (can also combine both ideas)

- SMT + rescoring + SMT



Source text

SMT

Int. Text (N-best)

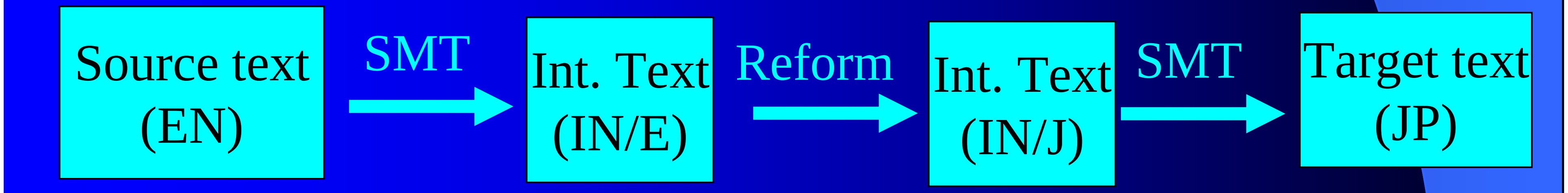
Rescore

Int. Text (single)

SMT

Target text

- SMT + interlingua-reformulation + SMT



Source text (EN)

SMT

Int. Text (IN/E)

Reform

Int. Text (IN/J)

SMT

Target text (JP)

Experiments

- Evaluate relative performance of different processing pipelines
- Evaluate on held-out part of generated data
  - Measure agreement with RBMT translation
  - GEAF 2009 paper: when SMT and RBMT different, SMT often worse and hardly ever better
- Evaluate on real out-of-coverage data
  - Use human judges

Results on generated data

(Metric: agreement with original RBMT system)

Configuration	EN → FR	EN → JP
Plain RBMT	(100%)	(100%)
Plain SMT	65.8%	26.8%
SMT + SMT	76.6%	10.5%
SMT + int-reformulation + SMT	---	74.1%
SMT + int-rescoring + SMT	78.5%	10.8%
SMT + int-rescore + int-reform + SMT	---	78.5%

Results on real OOC text data

Processing pipeline:
SMT + rescoring (+ reformulation for JP) + SMT

358	out-of-coverage utterances
245	well-formed interlingua
81	good backtranslation
76/81	good translations (French)
71/81	good translations (Japanese)

Summary (translation)

- Goal: relearn small RBMT system as SMT
- Not trivial if high precision required
- Much better results if we use interlingua
- Key idea: text form of interlingua
  - Use interlingua to reorder SMT output
  - Use interlingua to handle word-order problems
- Good results on EN-FR and EN-JP
  - Good agreement with RBMT (in-coverage data)
  - Adds non-trivial robustness (out-of-coverage data)

Outline

- Goals of paper
- MedSLT
- Bootstrapping a statistical recogniser
- Bootstrapping an interlingua-based SMT
- ➢ Putting it together
- Conclusions

Putting it together

- Combine (for both EN → FR and EN → JP)
  - best bootstrapped statistical recognition module
  - best bootstrapped MT module
- Compare different versions

Versions

- Original RBMT system
  - Rule-based recognition + rule-based MT
- Bootstrapped statistical system
  - Statistical recognition + statistical MT
- Hybrid system
  - Rule-based if it gives a result OTHERWISE bootstrapped statistical

Comparing versions

- Show pairs of results to bilingual judges
  - Statistical versus rule-based
  - Hybrid versus rule-based
- Ask which version judge prefers
  - If one result is null, other must be useful
  - Bad translation is worse than no translation
- Get backtranslation judgements
  - Which examples would be discarded?

Results (EN → FR)

Comparison	Judged by		
	J1	J2	Agree
Rules v Stat (all)	**261-43**	**259-43**	**247-33**
Rules v Stat (g. b/trans)	**69-25**	**71-27**	**62-20**
Hybrid v Rules (all)	**29-180**	**30-181**	**25-177**
Hybrid v Rules (g. b/trans)	18-12	19-15	15-12

Results (EN → JP)

Comparison	Judged by		
	J1	J2	Agree
Rules v Stat (all)	125-98	**147-96**	**101-47**
Rules v Stat (g. b/trans)	**61-25**	**66-41**	**49-21**
Hybrid v Rules (all)	49-62	30-81	**23-55**
Hybrid v Rules (g. b/trans)	17-8	19-9	14-8

Hybrid versus rule-based with backtranslation

- Small increase in recall
- Loss of precision seems more important
- Typical bad example (EN → FR)

Do you take medicine for your headaches? →
Avez-vous vos maux de tête quand vous prenez des médicaments?
("Do you have headaches when you take medicine?")

Summary and conclusions

- Method for bootstrapping statistical speech translation system from rule-based one
- Central problems:
  - Safety-critical application
  - Not much training data available
- Exploiting interlingua makes bootstrapped version much more competitive
- Hybrid version increases recall a little but degrades precision

Bottom line

- Generally applicable methods
- Might be useful for bootstrapping statistical speech translators in some domains
- For safety-critical applications like medicine, no reason to think statistical is better than rule-based

Thank you!