Hierarchical Hybrid Translation between English and German

Yu Chen, Andreas Eisele

DFKI GmbH, Saarbrücken, Germany

May 28, 2010

Outline

- Introduction
- Architecture
- Experiments
- Conclusion

SMT vs. RBMT

[K. Chen & H. Chen, 1996]

Rule-based machine translation (RBMT)

- Advantages
  1. easy to build an initial system
  2. based on linguistic theories
  3. effective for core phenomena
- Disadvantages
  1. rules are formulated by experts
  2. difficult to maintain and extend
  3. ineffective for marginal phenomena

Statistical machine translation (SMT)

- Advantages
  1. numerical knowledge
  2. extracts knowledge from corpus
  3. reduces the human cost
  4. mathematically grounded model
- Disadvantages
  1. no linguistic background
  2. search cost is expensive
  3. hard to capture long distance phenomena

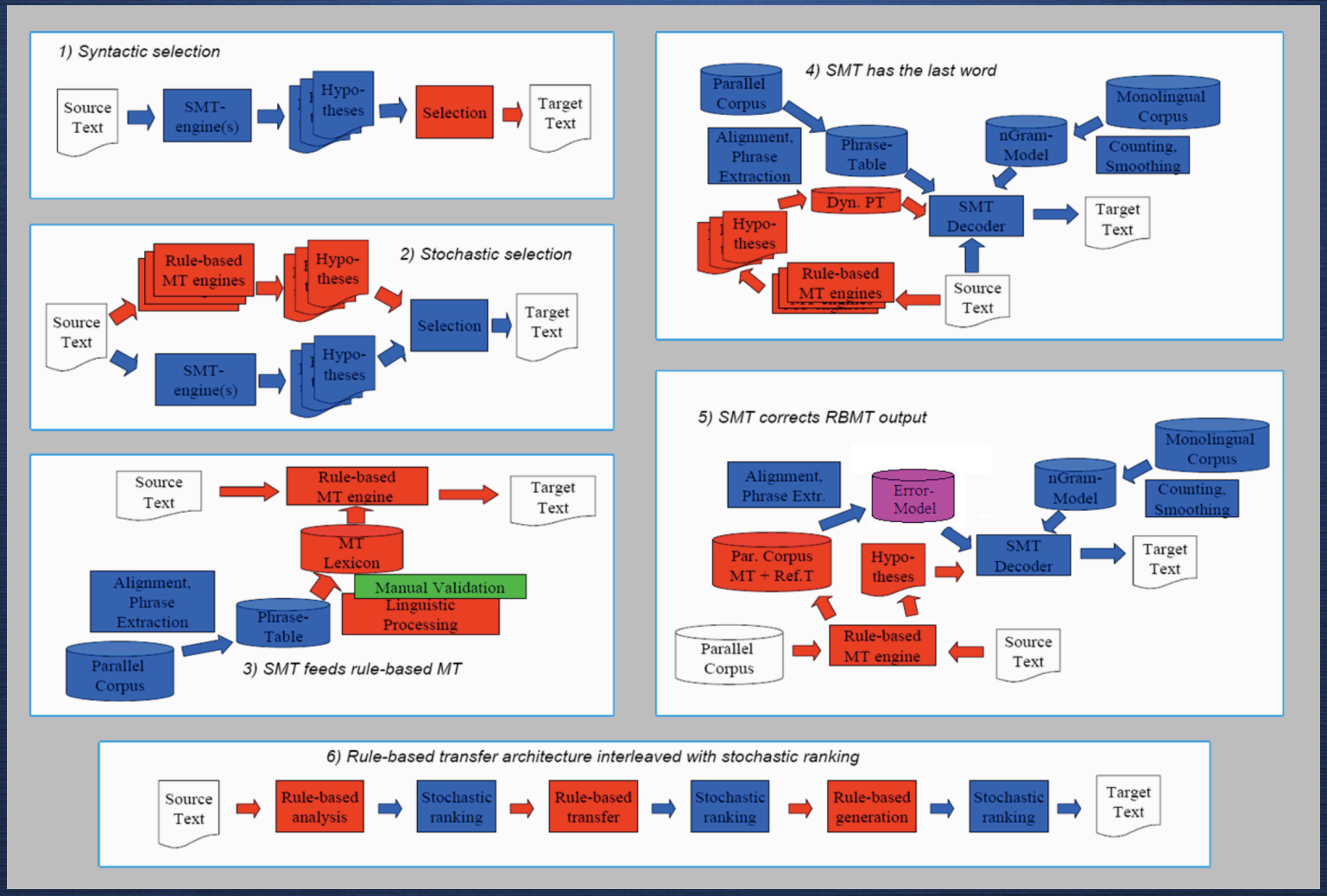
HYBRID MACHINE TRANSLATIONS

- RBMT & SMT: complementary

	RBMT	SMT
Syntax, Morphology	++	- -
Structural semantics	++	- -
Lexical semantics	-	+
Lexical adaptivity	- -	+
Lexical reliability	+	-

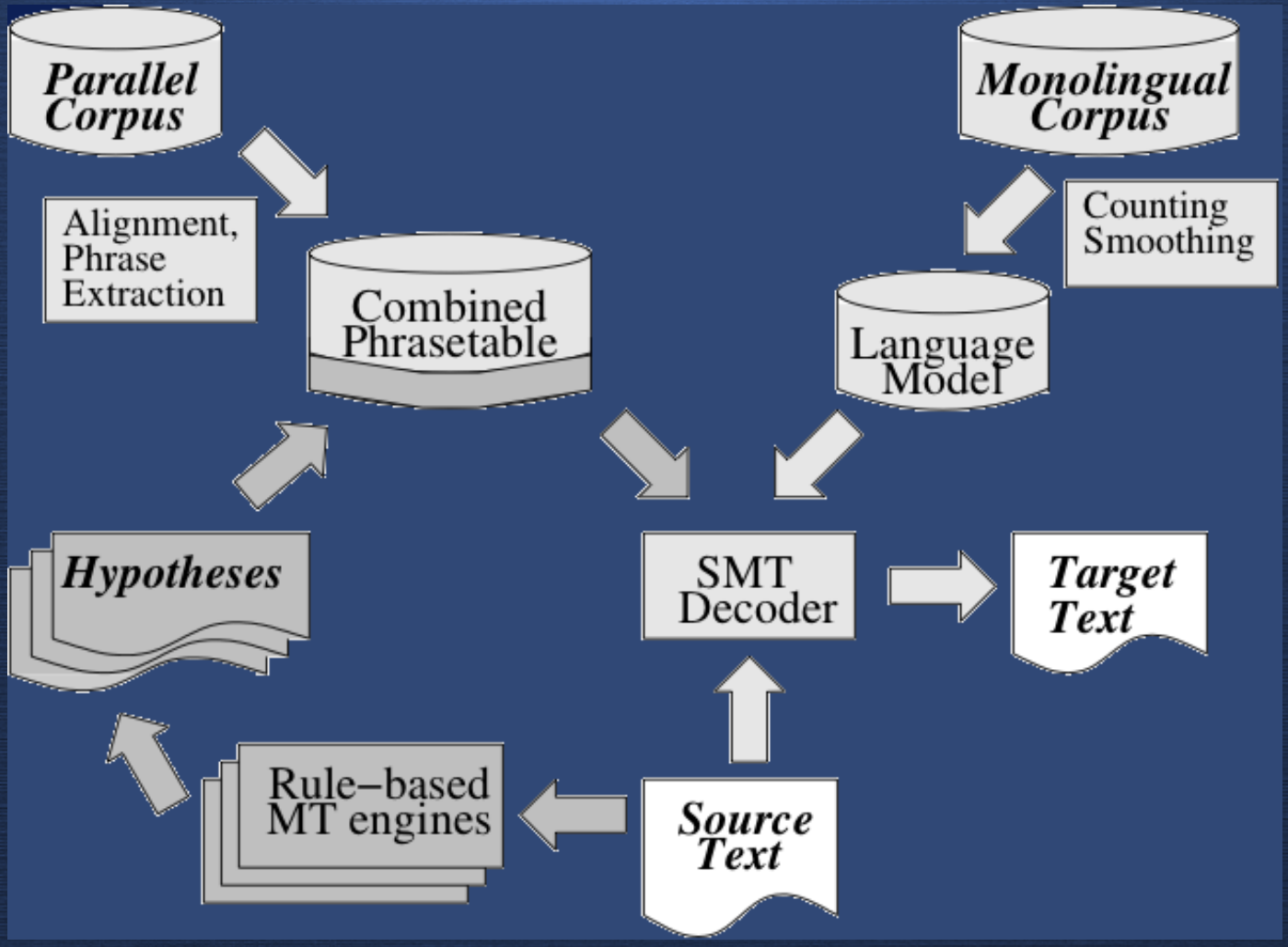
- Integrated approaches for better results

VARIOUS WAYS TO COMBINE SMT AND RBMT



[explored in EuroMatrix (Plus) since 2006]

HYBRID ARCHITECTURE CONSIDERED HERE



Phrase-based SMT + RBMT

[Eisele et al. 2008]

- Integration approach with PBSMT
  - Extract translation correspondences from RBMT outputs
  - Construct phrase tables using the extracted information
  - Combine phrase tables from different sources
  - Produce the final translations with a SMT decoder...
  - ...re-purposed to combine snippets from different sources
- Problems
  - Multiple RBMT systems required for better performance
  - Only outperforms SMT baseline in out-of-domain tests
  - Good structures from RBMT not exploited by SMT

HIERARCHICAL SMT + RBMT

How to utilize the ***implicit linguistic knowledge*** contained in RBMT translations in a robust way?

HIERARCHICAL SMT + RBMT

How to utilize the ***implicit linguistic knowledge*** contained in RBMT translations in a robust way?

- Our solution: Hierarchical phrases
  - Preserve some useful structures from RBMT outputs
  - No detailed linguistic analysis required

RBMT PHRASE TABLE

Close to the standard SMT training

- RBMT translations
  - Only on the development/test set
- Word alignment
  - From the input text to the RBMT translation
  - Large word alignment model from the SMT baseline system as the base model
- Phrase extraction
  - Standard rule extraction procedure with suffix arrays
  - Higher feature values than the normal SMT models

COMBINED PHRASE TABLE

Phrase table extension with RBMT features

source	target	SMT features			RBMT features		
zum	at the	1.98	1.89	2.43	1.95	1.82	2.12
der X_1 , die	the X_1 which	1.25	1.78	1.67	1.05	1.48	1.42
der X_1 der X_2	of the X_1 of the X_2	1.39	1.12	1.86	1.58	1.06	1.50
landesgrenzen	boundaries	1.15	1.75	1.11	1.0	1.0	1.0
X_1 abgeschlossen sein	X_1 be finalised	1.84	1.70	1.85	1.0	1.0	1.0
fakten X_1 der X_2	facts X_1 against the X_2	1.04	1.04	3.61	1.0	1.0	1.0
nach den	after that	1.0	1.0	1.0	1.11	2.10	2.12
auf der X_1	on which X_1	1.0	1.0	1.0	1.36	1.42	2.13
die X_1 von X_2	who X_1 of X_2	1.0	1.0	1.0	1.38	1.27	1.92

Training

- Tuned on combined PT with development set
- Translate test set with combined PT

Experiment setup

- Data
  - Training: Europarl-v4
  - Development: WMT 2007
  - Testing: WMT 2008 (EP)
  - Out-of-domain: News (NC)
- Tools
  - RBMT: Lucy
  - Hierarchical SMT: Joshua
  - Alignment
    - Training set: Berkeley aligner
    - Hypothesis alignment: GIZA++
  - MERT: ZMERT

BLEU SCORES

	de-en		en-de	
	EP	NC	EP	NC
Lucy	16.40	17.02	11.23	13.01
Moses	27.27	16.66	19.42	10.27
+Lucy	27.26	16.06	19.19	12.35
Joshua	27.51	16.24	20.69	10.48
+Lucy	**27.52**	**17.69**	**20.89**	**13.21**

Examples

In-domain

Source	Ich möchte Sie daran erinnern, dass sich unter unseren Verbündeten entschiedene Befürworter dieser Steuer befinden.
Reference	Let me remind you that our allies include fervent supporters of this tax.
Lucy	I would like to remind you of there being decisive proponents of this tax among our allies.
Moses	I would like to remind you that under our allies are strong supporters of this tax.
+Lucy	I would like to remind you that there are among our allies in favour of this tax.
Joshua	I would like to remind you that , under our allies are strong supporters of this tax.
+Lucy	I would like to remind you that there are strong supporters of this tax among our allies.

EXAMPLES

OUT-OF-DOMAIN

Source	So kooperieren die Hochschulen schon aus Tradition mit den Nachbarländern.
Reference	The university-level institutions' cooperation with the neighboring countries, for instance, is part of a tradition.
Lucy	So the colleges co-operate already from tradition with the neighbor countries closely.
Moses	So the universities from tradition cooperate closely with the neighbouring countries.
+Lucy	So the colleges co-operate closely with the neighbouring already from tradition.
Joshua	So cooperate closely with the neighbouring the universities from tradition.
+Lucy	So the universities, already from tradition, co-operate closely with the neighbouring countries.

Alignment for RBMT outputs

Systems	Test Set	∅	Europarl
Moses+Lucy	Europarl	19.37	19.19
Moses+Lucy	News	12.50	12.38
Joshua+Lucy	Europarl	20.83	20.89
Joshua+Lucy	News	13.17	13.21

SUMMARY

- Integrating a RBMT system with hierarchical SMT system

  Features
  - Hierarchical rule extraction from RBMT outputs
  - Phrase table extension with all features

  Results
  - Improvement over both sub-systems
  - Superior to hybrid system based on PBSMT

Future Work

- Other variants
  - Larger in-domain language models
- Better alignment and rule extraction
  - Build more reliable alignment model for RBMT outputs
  - Internal alignments from RBMT
- Deeper integration