

Architectures for hybrid/multi-engine systems

Teresa Herrmann – Silke Theison

Overview

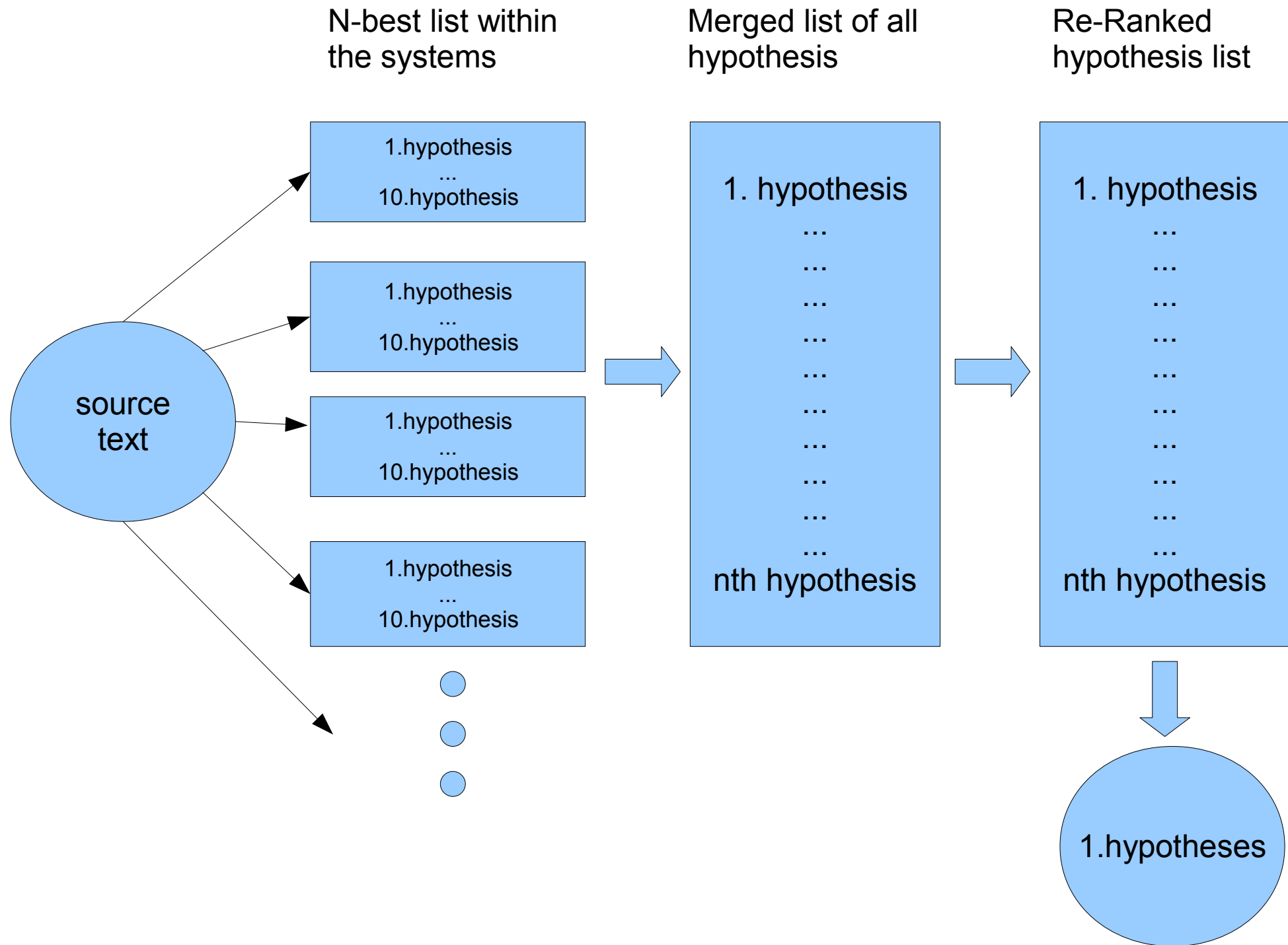
- Combining Output from different Machine Translation Systems
- Multi-Engine MT with an Open Source Decoder for SMT
- Postediting of Systran
 - Introduction to Systran
 - Dugast et al. (2007)
 - Simard et al. (2007)
- Logon- hybrid deep MT: Oepen et al. (2007)
 - Basic Linguistic System
 - Hybrid Architecture
- Conclusion

Combining outputs from different multiple MT systems

- Three different combination methods
 - Sentence-level
 - Phrase-level
 - Word-level

Sentence Level

- Re-ranking a merged N-best list
- Hypothesis Confidence Estimation
 - Generalized linear models
 - Confidence P_{ij} for system i generating hypothesis j
 - $\log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \sum_{l=1}^L w_{ij} x_{ijl}$ w_{ij} =weight x_{ijl} = l th feature
 - **Features** (Rank the Systems N-best list, Sentence Posterior with system-specific, System's total score, Number of Words in the hypothesis, ...)
- Hypothesis Re-Ranking
 - Weights are estimated for each feature
 - Re-score weights with the help of a 5-gram Language Model
 - Re-rank merged N-best list using the new weights

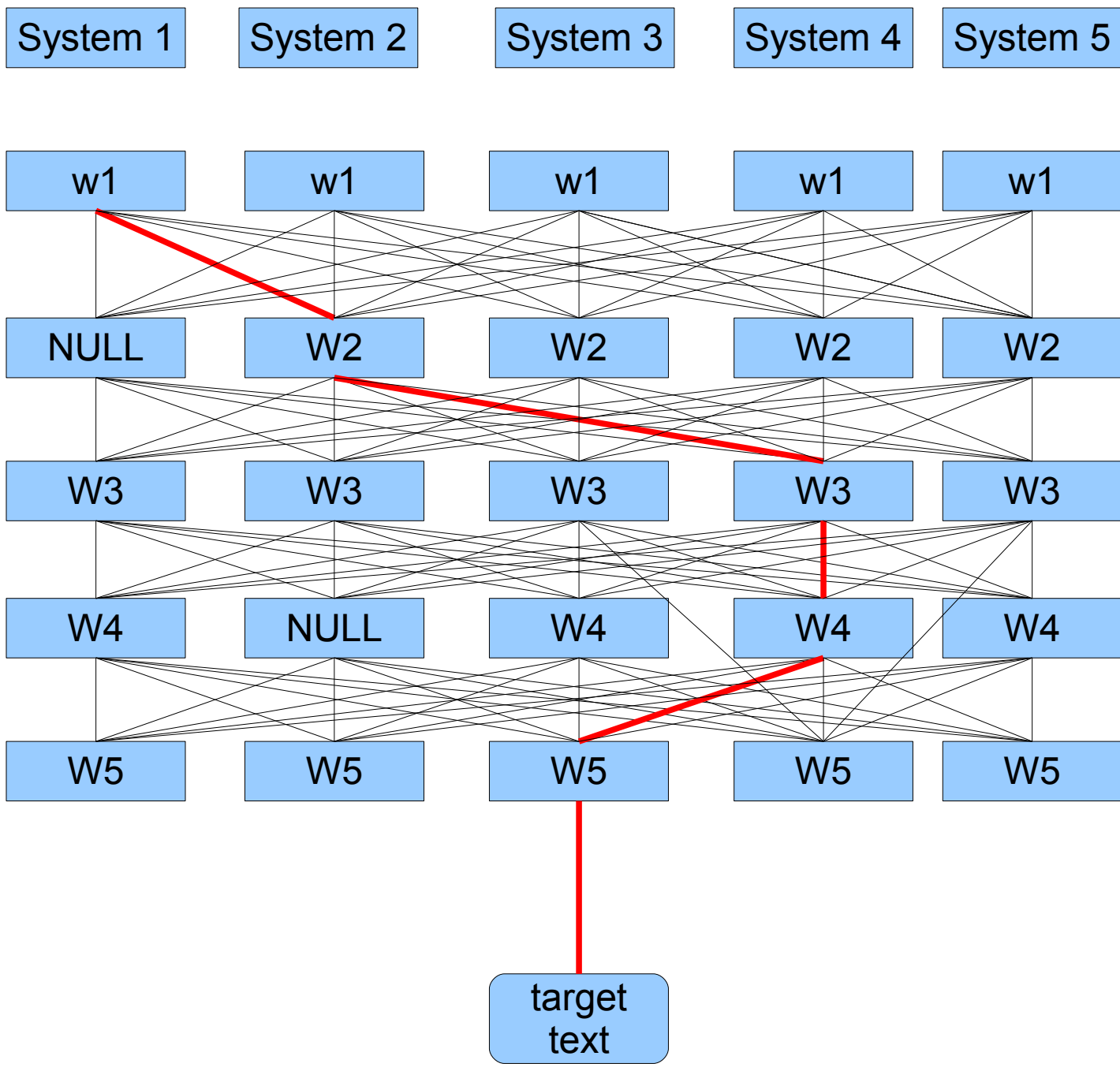


Phrase Level

- Extracting new Phrase Translation Tables
 - Target-to-source phrase alignments
 - Phrase confidence estimation
 - e.g. score is raised when several systems agree
- Re-decode (Pharaoh)
 - New PTT and Language Model

Word Level

- Confusion network
 - Generation
 - Choose skeleton
 - > the hypothesis that agrees most with the average of all hypothesis is chosen as the skeleton
 - Align n=10 hypothesis of each system with the skeleton
 - Decoding
 - Finding the highest scoring path through the network
- Tuning
 - no tuning
 - system weights
 - tuning
 - on 6 systems
 - on 8 systems



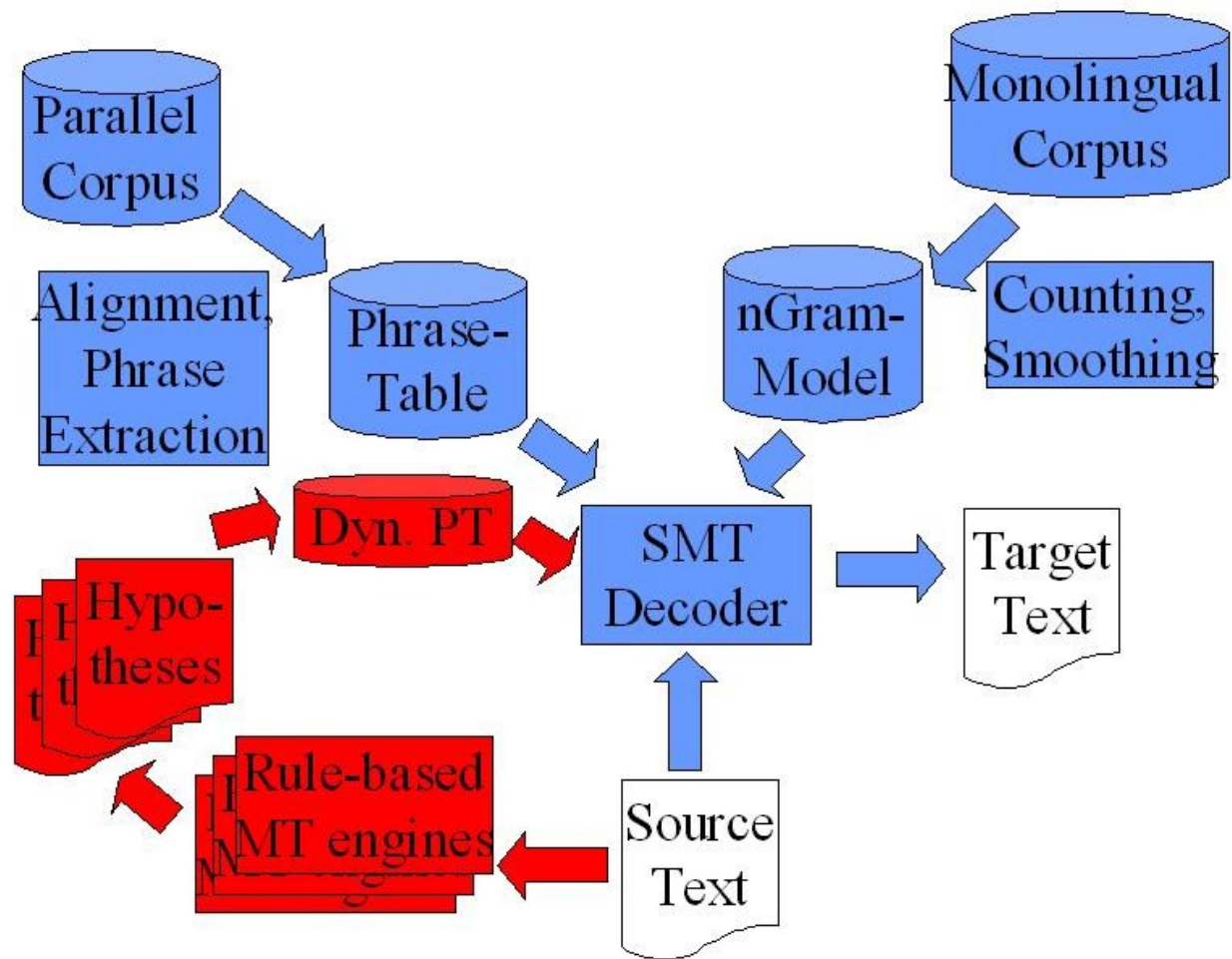
Multi-Engine MT with an Open Source Decoder for SMT

- Combine rule-based machine translation (RBMT) with statistical machine translation (SMT)
 - Translate input via RBMT
 - Alignment of RBMT output with input
 - Create additional phrase tables on RBMT alignment
 - Enhance original phrase table with new phrase tables
 - Include extra columns into the phrase table (one for each knowledge source, i.e. translation system)
 - Translate input via SMT with enhanced phrase-table

Results

Arabic	Newswire		Newsgroups	
	TER	BLEU	TER	BLEU
system A	42.98	49.58	59.73	20.36
system B	43.79	47.06	61.55	18.08
system C	43.92	47.87	60.81	18.08
system D	40.75	52.09	59.25	20.28
system E	42.19	50.86	59.85	19.73
system F	44.30	50.15	61.74	20.61
phrcomb	40.45	53.70	59.90	21.49
sentcomb	41.56	52.18	60.21	19.77
no weights 6	39.33	53.66	58.15	20.61
TER 6	39.41	54.37	58.21	20.85
TER 8	39.43	54.40	57.96	21.44

Chinese	Newswire		Newsgroups	
	TER	BLEU	TER	BLEU
system A	56.57	29.63	68.61	13.20
system B	56.30	29.62	69.87	12.33
system C	59.48	31.32	69.37	13.91
system D	58.32	33.77	67.61	16.86
system E	58.46	32.40	69.08	15.08
system F	56.79	35.30	68.08	16.31
phrcomb	56.50	35.33	68.48	15.88
sentcomb	56.71	36.24	69.50	16.11
no weights 6	53.80	36.17	66.87	15.90
BLEU 6	54.34	36.44	66.50	16.44
BLEU 8	54.86	36.90	66.45	17.32



SYSTRAN

- traditional rule-based system
- constant evolution since the founding of the company in 1968
 - evolution principles
 - provide deterministic output
 - incremental translation quality
- linguistic resources
 - simple and multiword lexical entries
 - customized disambiguation rules

SYSTRAN

- coverage: 80 language pairs, 22 source languages
- used as
 - basis of major portals (Google, Yahoo!, BabelFish, ...)
 - customized systems for corporate customers
 - desktop applications
 - applications for PDA devices

Most recently

- growing interest in making use of available corpora
 - dictionary improvement
 - word sense disambiguation
 - selection of alternatives using a language model

PORTAGE

- statistical phrase-based MT system
- operation on 3 main phases
 - preprocessing
 - tokenization
 - translation suggestions generated by rules
 - decoding
 - produce translation hypotheses
 - rescoring
 - error-driven
 - reorder the list of n-best translations
 - train a model for optimizing BLEU score
 - includes IBM model 2 probabilities in both directions
 - choose best final hypothesis

Portage Decoding

- search for hypotheses t with highest probabilities of being translations of current source sentence s according to a model for $P(t|s)$
- model for $P(t|s)$
 - trigram language model(s)
 - SRILM toolkit
 - phrase translation model(s)
 - symmetrized IBM model 2 word alignments for phrase pair induction
 - distortion model
 - similar to Koehn (2004) + final cost to account for sentence endings
 - word-length feature

Statistical post-editing on SYSTRAN

- combination of a SYSTRAN system with a „statistical post-editing“ (SPE) system
- Dugast et al. (2007)
- participation in Shared task of ACL 2007 Workshop on Statistical MT

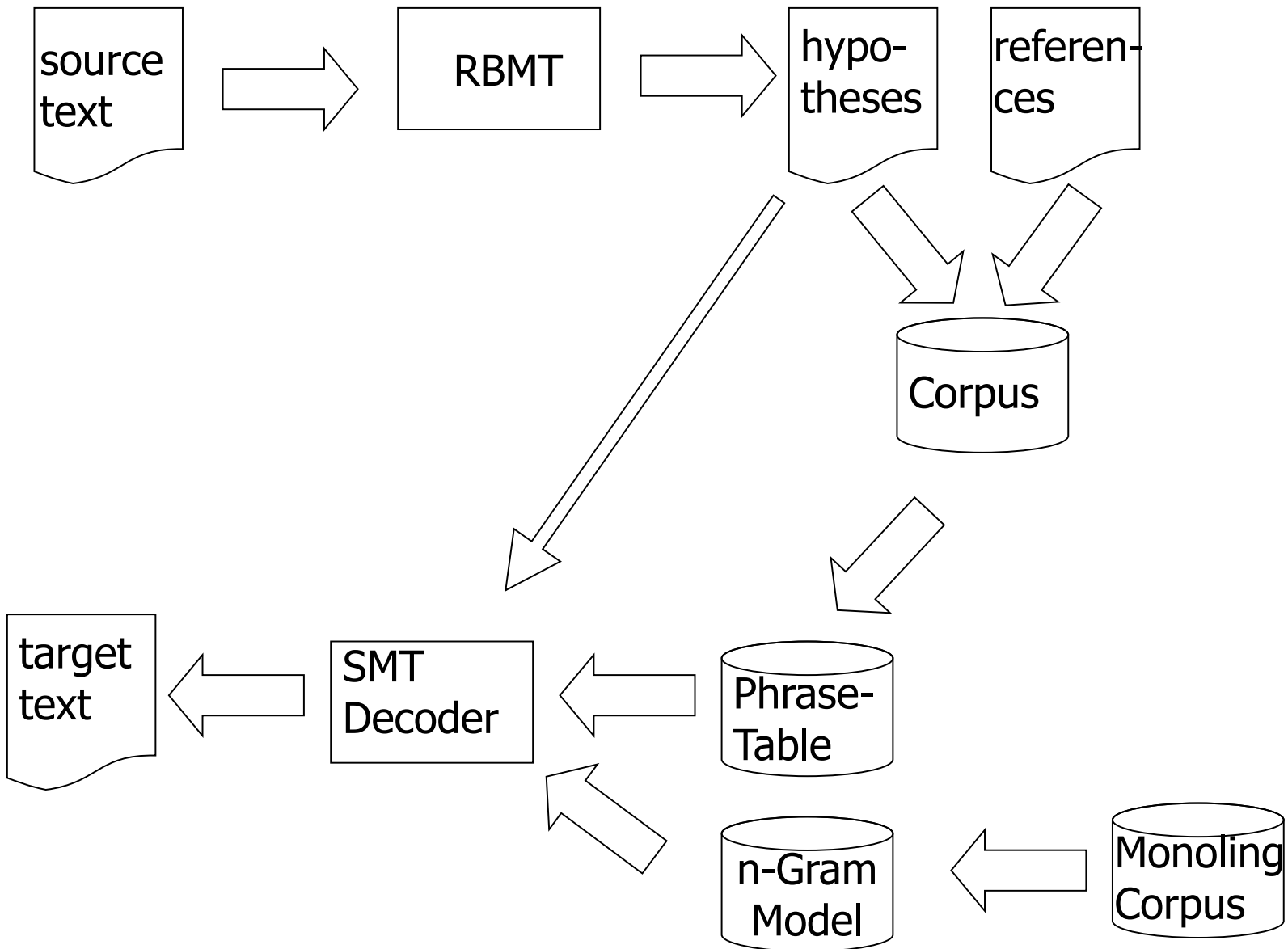
Experiments

- objective
 - train SMT system on parallel corpus of SYSTRAN translations and references
- achievements
 - increase of BLEU score (10 points)
 - improvement of translation fluency
- evaluation
 - qualitative analysis of contributions of SPE

Experiments

- SYSTRAN + Portage (En<>Fr)
- SYSTRAN + Moses (De>En, Es>En)
- Linguistic evaluation of pure SYSTRAN output and SYSTRAN + SPE

Architecture



Impact of SPE

- quite high: almost all sentences post-edited
 - Word Change Rate (WCR) relatively small
 - SPE output close to raw SYSTRAN output
 - SPE output structure completely based on SYSTRAN output
- only post-editing not complete reshuffling of translation

Classification of SPE changes

	<i>#improv</i>	<i>#de-grad</i>	<i>#improv / #degrad</i>	<i>#equiv</i>
termchg all	90	32	3	33
termchg_nfw	1	0		0
termchg_term	59	7	8	29
termchg_loc	15	1	15	1
termchg_mean	15	24	1	3
gram all	44	38	1	8
gram_det	20	3	7	4
gram_prep	12	9	1	1
gram_pron	0	1	0	2
gram_tense	2	8	0	0
gram_number	4	4	1	0
gram_gender	2	8	0	0
gram_other	4	5	1	1
punct/digit/case	8	7	1	1
wordorder_short	0	1	0	0
wordorder_long	0	0		0
style	3	1	3	1

Relative Improvements

	SYSTRAN PORTAGE En>Fr	SYSTRAN Moses De>En	SYSTRAN Moses Es>En
termchg all	+22%	+46%	+46%
termchg_nfw	0%	+3%	+1%
termchg_term	+19%	+42%	+45%
termchg_loc	+8%		
termchg_mean	-6%		
gram all	+2%	+4%	+12%
gram_det	14%	+2%	+4%
gram_prep	2%	+1%	+5%
gram_pron	-1%	+1%	+4%
gram_tense	-4%	+1%	-0%
gram_number	0%	None	None
gram_gender	-4%	n/a	n/a
gram_other	-1%	None	None
punct/digit/case	1%	-1%	-1%
wordorder_short	-1%	+1%	+1%
wordorder_long	0%	None	+1%
style	1%	+3%	+2%

Analysis of Results

- term changes
 - mostly improved
 - local choice of word sense or alternative translation of words and locutions
 - degradations
 - changes of sentence meaning by lexical modification
- grammatical changes
 - improvements
 - determiners
 - prepositions
 - improvement/degradation ratio low
 - global improvements, but many unacceptable degradations

Analysis of Results

- long distance
 - no changes observed
 - local reordering negative for En>Fr, negligible for other language pairs
- morphology
 - French: accounts for 25% of degradations
 - no control mechanism for morphology in SPE
- acceptance criterion for new SYSTRAN version (8 improv for 1 degrad) not met

Intermediate Conclusions

- good results for
 - automatic scoring
 - linguistic analysis
- further improvements
 - adding linguistic control mechanisms
 - linguistic constraints in decoding process
 - available in translation output
- need for learning new terminology
 - exploit phrasetales built on parallel corpora

Intermediate Perspectives

- integration of data-driven mechanisms within translation engines
 - word sense disambiguation
 - selection of alternative translations
 - specific local phenomena (determination)
- 2 possible approaches
 - specialize statistical layer by breaking it down into components
 - execute on a narrow and accurate area
 - integrated into rule-based system
 - refinements of global SPE approach
 - introduce linguistic constraints

Statistical phrase-based post-editing on SYSTRAN and PORTAGE

- automatic post-editing (APE) strategy
- Simard et al. (2007)
- participation in Shared task of the Second Workshop on Statistical MT

Automatic Post-Editing Strategy

- initially translate input into target language using rule-based MT system
 - SYSTRAN
- automatically post-edit output using a statistical phrase-based system
 - PORTAGE
 - system learns to correct typical errors of rule-based system
 - performs domain-specific corrections and adaptations to the output

Motivation

- repetitive nature of errors of rule-based systems
 - train statistical MT system to correct systematic errors
 - „source language“: output of rule-based system
 - „target language“: reference translations (human)
 - training material for APE layer usually domain-specific
- automatic adaption of rule-based system to specific domain

Resources

- training and test corpora
 - specific and unusual context
 - 3 parallel views
 - source language text
 - machine translation
 - manually post-edited machine translation
 - small corpus
 - Fr>En: 500k source language words
 - En>Fr: 350k
- Questions
 - scale up?
 - independent machine translations and references?

PORTAGE: Statistical Phrase-based Post-Editing

- Configurations
 - use of two distinct phrasables
 - Europarl
 - News Commentary training corpus
 - multiple phrase-probability feature functions
 - joint probability estimate
 - frequency-based conditional probability estimate
 - variants thereof
 - 4-gram language model
 - combined Europarl + News Commentary
 - 3-gram language model
 - mini-corpus of test-relevant sentences (IR techniques)
 - 5-gram truecasing model
 - combined Europarl + News Commentary

Training corpus – 2 variants

- dependent texts
 - raw MT output
 - manually post-edited versions of translations
- independent texts
 - raw MT output
 - independent human reference

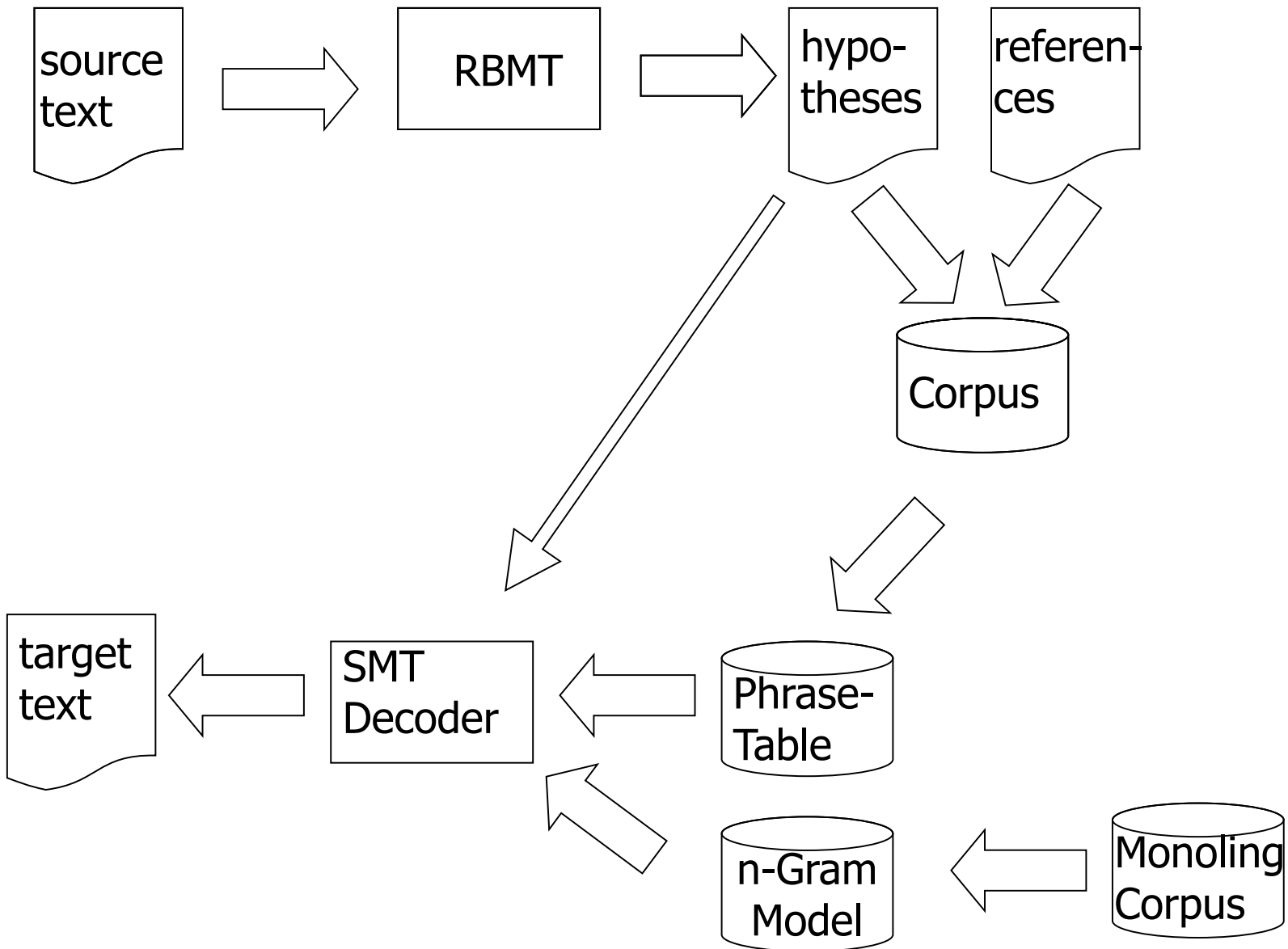
Training corpus – 2 variants

- dependent texts
 - raw MT output
 - manually post-edited versions of translations
- independent texts
 - raw MT output
 - independent human reference

2 Systems

- domain specific
 - Europarl
 - News Commentary
- configurations
 - different for domains
 - adapted language model
 - rest identical

Architecture



Resulting BLEU Scores

	en \rightarrow fr	fr \rightarrow en
<hr/>		
Europarl (>32M words/language)		
SYSTRAN	23.06	20.11
PORTAGE	31.01	30.90
SYSTRAN+PORTAGE	31.11	30.61
<hr/>		
News Commentary (1M words/language)		
SYSTRAN	24.41	18.09
PORTAGE	25.98	25.17
SYSTRAN+PORTAGE	28.80	26.79
<hr/>		

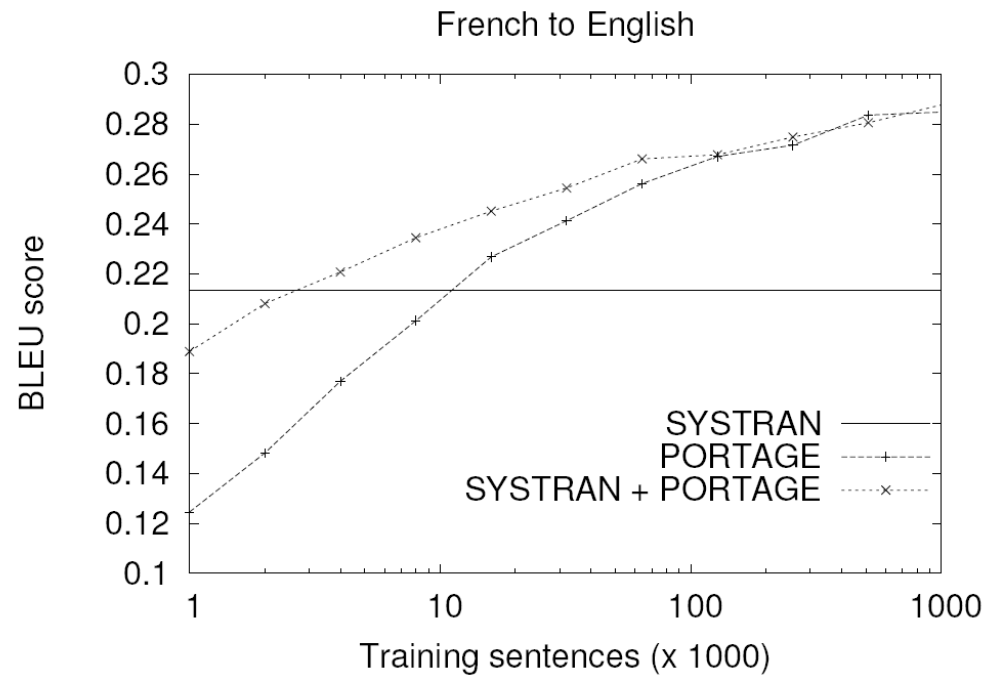
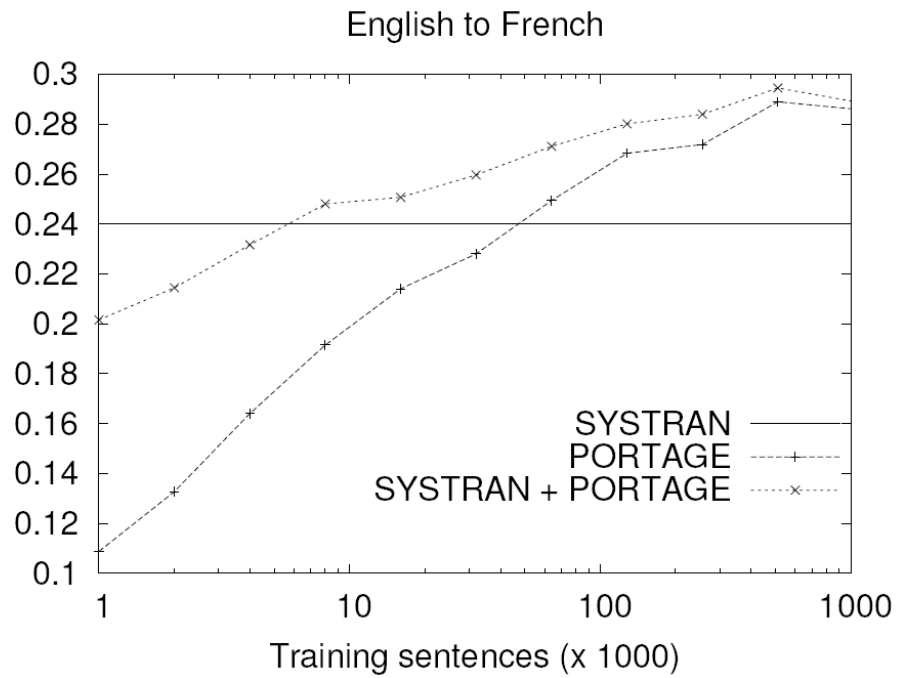
Observations

- large improvements on performance
 - over raw rule-based system
 - significant reduction of human post-editing effort
 - more dominant into English and for Europarl domain
 - over standalone statistical system
 - News Commentary domain
 - Fr>En: 1.5 BLEU points
 - En>Fr: ~3 BLEU points
 - Europarl domain
 - similar performances
 - Europarl corpus almost 30 times larger than NC
- APE better suited for domains with limited amounts of training data available

Serial Experiments

- training of series of APE and SMT systems on Europarl data
 - gradually increasing training data
- Results
 - improvements with more data for both APE and SMT
 - SMT improves more rapidly
 - more or less equivalent for more data
 - APE: little data required to improve upon rule-based system

Results



Intermediate Conclusions

- phrase-based post-editing significantly improves the output of a rule-based MT system
- with scarce training data → outperformance of direct phrase-based strategy
- no need for manually post-edited translations as training data
- most effective for scarce training data
- still competitive with SMT systems with larger amounts of data

Further investigations

- comparison to standard lexical customization of the rule-based system
- different configurations for APE and direct SMT systems
 - modification → better adapt to APE task
 - e.g. use source language text
- dive into the „black box“
 - which part of the rule-based processing can contribute further to the APE process

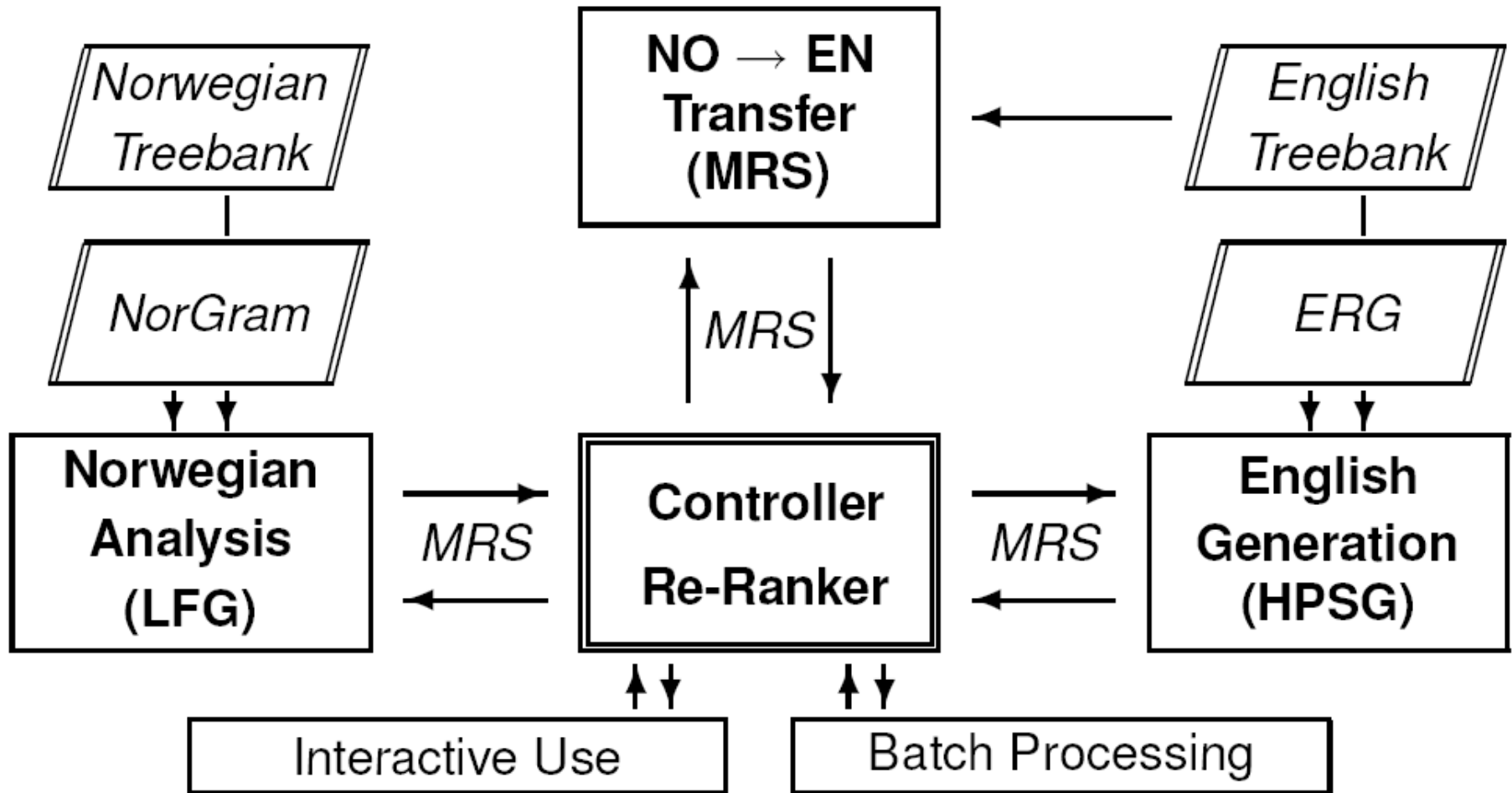
LOGON

- „Towards Hybrid Quality-Oriented Machine Translation“
- Oepen et al. (2007)
- Project of the Norwegian LOGON consortium
 - Universities of Oslo, Bergen and Trondheim
- hybrid MT architecture with semantic transfer backbone
- interplay of linguistic and stochastic processes

Hybrid deep MT

- deep core translation system
- LOGON pipeline
 - grammar-based parsing
 - NorGram: analysis grammar, based on LFG
 - transfer of underspecified MRS
 - unification-based, resource-sensitive rewriting of MRS terms
 - full tactical generation (realization)
 - ERG: generation grammar, based on HPSG

Architecture



Architecture

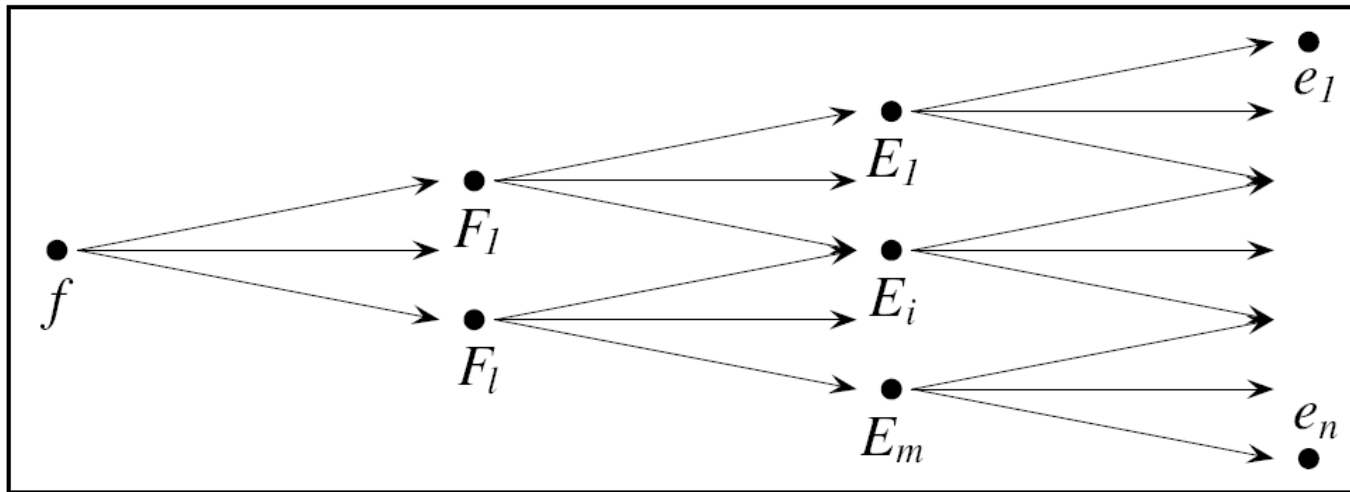
- rule-based components
 - produce grammatically and semantically coherent translations
- probabilistic task
 - ranking of competing hypotheses
 - selection of best candidate(s)

Resources

- parallel corpus of 50k words (tourism domain)
 - source text
 - up to three reference translations
 - 10% held out for evaluation
- model training and evaluation
 - manual treebanking

Fan-out

- at every intermediate step in the pipeline n-best hypotheses
- fan-out tree
- goal: find the path from input to translation



Where to apply statistics

- several possibilities for adding stochastic component
 - first translation strategy: rank components and choose best candidate sequentially
 - rank parsing results, choose best candidate →
 - rank transfer results, choose best candidate →
 - rank generation results, choose best candidate
 - most likely path through the fan-out tree
 - try to maximize:
 - most likely translation

- maximize:
$$\arg \max_{i,j,k} P(e_k|E_j)P(E_j|F_i)P(F_i|f)$$

$$\arg \max_e \sum_{F_i} \sum_{E_j} P(e_k|E_j)P(E_j|F_i)P(F_i|f)$$

Arguments for the strategies

- first translation
 - more likely to be the intended interpretation of the producer
 - corresponds better to human evaluation
 - most likely translation
 - favoured by BLEU score
- first translation preferred
- probabilities of later components in the pipeline may help to select earlier ones

First translation

- Parse Selection
 - treebanking of development corpus
 - application of statistical estimation algorithm → property weights to rank parses
- Ranking Transfer Output
 - smoothed trigram model over the reduction of MRSs into dependency triples
 - rank according to perplexity scores of the triples assigned by this „semantic“ model
- Realization Ranking
 - rank candidate surface forms
 - use of treebank to train model for conditional probability of a surface realization given an input MRS

End-to-End Re-Ranking

- most likely translation strategy
 - re-rank complete list of hypotheses when fan-out complete
 - component-internal probabilities are fallible
 - additional properties of source-target pair $\langle f, e \rangle$ besides level-internal information
 - re-ordering, length
- log-linear model to estimate posterior translation probability $P^\lambda(e|f)$
- set of features
 - learned weights for each feature

Features

- previous scores of
 - parse ranker
 - MRS ranker
 - realization ranker
- lexical probabilities
- string probabilities
 - according to trigram language model
- distortion
 - measure re-ordering among constituents
- string harmony
 - assumption of systematic correspondences between source and translation at string level
- transfer metrics
 - total number of transfer rules invoked
- semantic distance
 - MRS compatibility test

Results

set	#	chance	first	MMI	top	judge
JH_d	1391	34.18	40.95	44.10	49.89	–
JH_t	115	30.84	35.67	38.92	45.74	46.32

- actual translation of 64.8% (dev corpus) and 63.2% (test corpus) of the input
- first translation (first) shows improvement w.r.t. random choice of hypotheses
- re-ranking (MMI) even more improvement
- oracle scores: top, judge

Intermediate Conclusions

- hybrid MT based on computational grammars
- pipeline of hand-built linguistic resources
- coverage:
 - succeeds in translating ~60% of unseen input
 - bounded domain and limited vocabulary
- flexibility:
 - n-best beam search +
 - stepwise stochastic ranking or
 - end-to-end re-ranking
- high degree of quality:
 - BLEU scores
 - human inspection

Conclusions

- proposed hybrid approaches
 - parallel application of various RBMT and SMT systems
 - different ranking of hypotheses
 - enhanced SMT system
 - augment original SMT phrasetable with additional phrasetable constructed out of aligned RBMT input and output
 - post-edit RBMT output with SMT system
 - apply statistics at several steps of pipeline architecture
 - Parser, Transfer, Generation
- encouraging results
 - improvement w.r.t. raw RBMT and SMT performance
 - performance less dependent on larger amounts of data
- areas for further improvement remain

Outlook

- desirable further investigation
 - post-edit SMT output with RBMT system
 - deeper integration of rules and statistics
 - application of statistics
 - not only at the output level
 - on several levels

References

- Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie J. Dorr. 2007. Combining outputs from Multiple Machine Translation Systems. In *Proceedings of NAACL HLT 2007*. Rochester, NY, April 2007.
- Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus, and Silke Theison. 2007. Multi-Engine Machine Translation with an Open-Source SMT Decoder. To appear in *Proceedings of the Second ACL workshop on statistical machine translation*, June 23, 2007, Prague, Czech Republic.
- Simard Michel & al. 2007. Rule-based Translation With Statistical Phrase-based Post-editing. To appear in *Proceedings of the Second Workshop On Statistical Machine Translation*, Prague, Czech Republic.
- L. Dugast, J. Senellart, and P. Koehn. 2007. Statistical Post-Editon on SYSTRAN Rule-Based Translation System. To appear in *Proceedings of the Second Workshop On Statistical Machine Translation*, Prague, Czech Republic.
- M. Simard, C. Goutte, and P. Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508-515, Rochester, USA.
- F. Sadat, H. Johnson, A. Agbago, G. Foster, R. Kuhn, J. Martin, and A. Tikuisis. 2005. PORTAGE: A Phrase-Based Machine Translation System. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 129-132, Ann Arbor, USA.
- S. Oepen, E. Velldal, J. T. Lønning, P. Meurer, V. Rosén and D. Flickinger. 2007. Towards Hybrid Quality-Oriented Machine Translation – On Linguistics and Probabilities in MT.
<http://www.euromatrix.net/partners/saarland-university/logon-draft.pdf>