A Hybrid Morpheme-Word Representation for Machine Translation of Morphologically Rich Languages

Minh-Thang Luong, Preslav Nakov & Min-Yen Kan
EMNLP 2010, Massachusetts, USA

NUS
National University of Singapore

epäjärjestelmällisyydellistyttämättömyydellänsäkäänköhän

This is a Finnish word!!!

epä+ järjestelmä+ llisyy+ dellisty+ ttä+ mättö+ myy+ dellä+ nsä+ kään+ kö+ hän

epäjärjestelmällisyydellistyttämättömyydellänsäkäänköhän

↓

This is a Finnish word!!!

epä+ järjestelmä+ llisyy+ dellisty+ ttä+ mättö+ myy+ dellä+ nsä+ kään+ kö+ hän

↓

system

epäjärjestelmällisyydellistyttämättömyydellänsäkäänköhän

↓

This is a Finnish word!!!

epä+ järjestelmä+ llisyy+ dellisty+ ttä+ mättö+ myy+ dellä+ nsä+ kään+ kö+ hän

↓

unsystem

epäjärjestelmällisyydellistyttämättömyydellänsäkäänköhän

↓

This is a Finnish word!!!

epä+ järjestelmä+ llisyy+ dellisty+ ttä+ mättö+ myy+ dellä+ nsä+ kään+ kö+ hän

↓

unsystematic

epäjärjestelmällisyydellistyttämättömyydellänsäkäänköhän

↓

This is a Finnish word!!!

epä+ järjestelmä+ llisyy+ dellisty+ ttä+ mättö+ myy+ dellä+ nsä+ kään+ kö+ hän

↓

I wonder if it's not with his act of not having made something be seen as unsystematic

- **Morphologically rich languages (Arabic, Basque, Turkish, Russian, Hungarian, etc.)**
  - Extensive use of affixes

Morphologically rich languages are hard for MT
→ Analysis at the morpheme level is needed.

NUS National University of Singapore

Morphological Analysis Helps ?
– Translation into morphologically *poor* languages

- ***Morpheme*** **representation alleviates data sparseness**

  Arabic → English (Lee, 2004) *Czech → English* (Goldwater & McClosky, 2005)
  Finnish → English (Yang & Kirchhoff, 2006)

- **For large corpora, *word* representation captures context better**

  Arabic→ English (Sadat & Habash, 2006)
  Finnish→ English (de Gispert et al., 2009)

Our approach: the basic unit of translation is the ***morpheme***, but ***word*** boundaries are respected at all MT stages.

NUS National University of Singapore

Morphological Analysis Helps ?
– Translation into morphologically *rich* languages

- **A challenging translation direction**
- **Recent interest in:**
  - English → Arabic (Badr et al., 2008)
  - English → Turkish (Oflazer and El-Kahlout, 2007)

  → enhance the performance for small bi-texts only

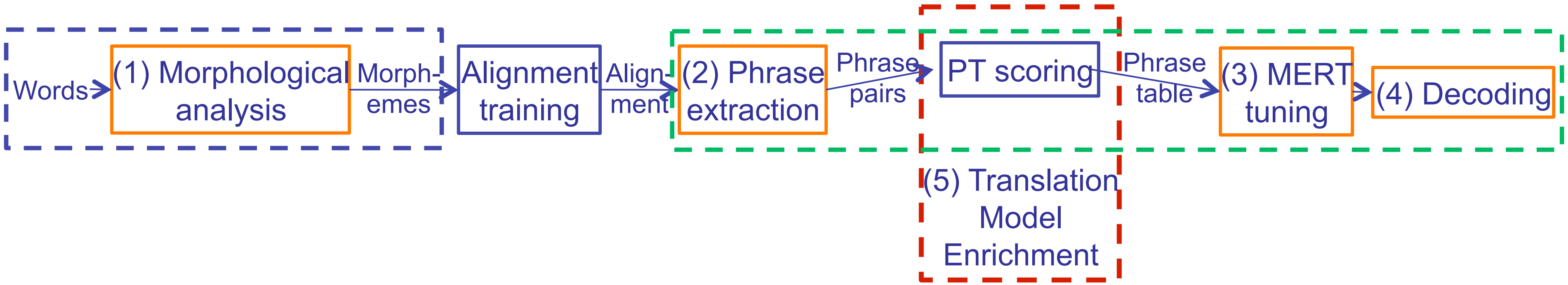
  - English → Greek (Avramidis and Koehn, 2008),
  - English → Russian (Toutanova et al., 2008)

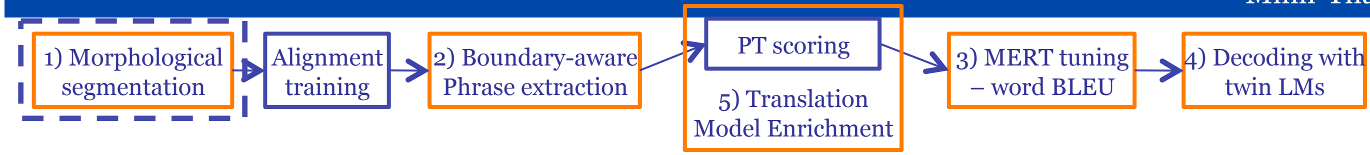
  →rely heavily on language-specific tools

We want an unsupervised approach
that works for large training bi-texts.

Methodology

- **Morphological Analysis** – *Unsupervised*
- **Morphological Enhancements** – *Respect word boundaries*
- **Translation Model Enrichment** – *Merge phrase tables (PT)*





Morphological Analysis

- Use *Morfessor* (Creutz and Lagus, 2007) - unsupervised morphological analyzer

- Segments words → morphemes (PRE, STM, SUF)

  un/*PRE+* care/*STM+* ful/*SUF+* ly/*SUF*

- "+" sign used to enforce word boundary constraints later

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Word Boundary-aware Phrase Extraction

SRC = the$_{STM}$ new$_{STM}$, un$_{PRE+}$ democratic$_{STM}$ immigration$_{STM}$ policy$_{STM}$

TGT = uusi$_{STM}$, epä$_{PRE+}$ demokraat$_{STM+}$ t$_{SUF+}$ i$_{SUF+}$ s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$ politiikan$_{STM}$

(*uusi*=new , *epädemokraattisen*=undemocratic *maahanmuuttopolitiikan*=immigration policy)

- **Typical SMT:** maximum phrase length n=7 words
- ***Problem***: morpheme phrases of length n
  - can span less than n words
  - may only partially span words

This problem is severe for morphologically rich languages.

- ***Solution***: morpheme phrases
  - span up to n words
  - fully span words

avoid proposing nonwords

TRAINING

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Morpheme MERT Optimizing Word BLEU

- **Why ?**
  - BLEU's brevity penalty is influenced by sentence length
  - The same # of words span different # of morphemes
    - a 6-morpheme Finnish word: *epä*$_{PRE+}$ *demokraat*$_{STM+}$ *t*$_{SUF+}$ *i*$_{SUF+}$ *s*$_{SUF+}$ *en*$_{SUF}$

  → Suboptimal weight for the SMT word penalty feature
- ***Solution***: optimize on word BLEU.
- Each MERT iteration:
  - Decode at the morpheme level
  - Convert morpheme translation → word sequence
  - Compute word BLEU
  - Convert back to morphemes

TUNING

Decoding with Twin Language Models

- **Morpheme language model (LM)**
  - *Pros*: alleviates data sparseness
  - *Cons*: phrases span fewer words
- **Introduce a second LM at the word level**
  - *Log-linear model*: add a separate feature
  - *Moses decoder*: add *word-level* “view” on the *morpheme-level* hypotheses

$\underbrace{\text{uusi}_{\text{STM}}\ ,\ \text{epä}_{\text{PRE+}}\ \text{demokraat}_{\text{STM+}}\ \text{t}_{\text{SUF+}}\ \text{i}_{\text{SUF+}}\ \text{s}_{\text{SUF+}}\ \text{en}_{\text{SUF}}}_{\textit{Previous hypotheses}}\ \underbrace{\text{maahanmuutto}_{\text{PRE+}}\ \text{politiikan}_{\text{STM}}}_{\textit{Current hypothesis}}$

- Morpheme LM:

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Decoding with Twin Language Models

- **Morpheme language model (LM)**
  - *Pros*: alleviates data sparseness
  - *Cons*: phrases span fewer words
- **Introduce a second LM at the word level**
  - *Log-linear model*: add a separate feature
  - *Moses decoder*: add *word-level* “view” on the *morpheme-level* hypotheses

Previous hypotheses *Current hypothesis*

uusi$_{STM}$, epä$_{PRE+}$ demokraat$_{STM+}$ t$_{SUF+}$ i$_{SUF+}$ s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$ politiikan$_{STM}$

- Morpheme LM: “s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$”

DECODING

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Decoding with Twin Language Models

- **Morpheme language model (LM)**
  - *Pros*: alleviates data sparseness
  - *Cons*: phrases span fewer words
- **Introduce a second LM at the word level**
  - *Log-linear model*: add a separate feature
  - *Moses decoder*: add *word-level* "view" on the *morpheme-level* hypotheses

Previous hypotheses — *Current hypothesis*

uusi$_{STM}$, epä$_{PRE+}$ demokraat$_{STM+}$ t$_{SUF+}$ i$_{SUF+}$ s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$ politiikan$_{STM}$

- Morpheme LM: "s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$" ; "en$_{SUF}$ maahanmuutto$_{PRE+}$ politiikan$_{STM}$"

DECODING

Decoding with Twin Language Models

- **Morpheme language model (LM)**
  - *Pros*: alleviates data sparseness
  - *Cons*: phrases span fewer words
- **Introduce a second LM at the word level**
  - *Log-linear model*: add a separate feature
  - *Moses decoder*: add *word-level* “view” on the *morpheme-level* hypotheses

$\underbrace{\text{uusi}_{\text{STM}}\ ,\ \text{epä}_{\text{PRE+}}\ \text{demokraat}_{\text{STM+}}\ \text{t}_{\text{SUF+}}\ \text{i}_{\text{SUF+}}\ \text{s}_{\text{SUF+}}\ \text{en}_{\text{SUF}}}_{\textit{Previous hypotheses}}\ \underbrace{\text{maahanmuutto}_{\text{PRE+}}\ \text{politiikan}_{\text{STM}}}_{\textit{Current hypothesis}}$

- Morpheme LM: “$\text{s}_{\text{SUF+}}$ en_{SUF} $\text{maahanmuutto}_{\text{PRE+}}$” ; “$\text{en}_{\text{SUF}}$ $\text{maahanmuutto}_{\text{PRE+}}$ $\text{politiikan}_{\text{STM}}$ ”
- Word LM: uusi , epädemokraattisen maahanmuuttopolitiikan

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Decoding with Twin Language Models

- **Morpheme language model (LM)**
  - *Pros*: alleviates data sparseness
  - *Cons*: phrases span fewer words
- **Introduce a second LM at the word level**
  - *Log-linear model*: add a separate feature
  - *Moses decoder*: add *word-level* "view" on the *morpheme-level* hypotheses

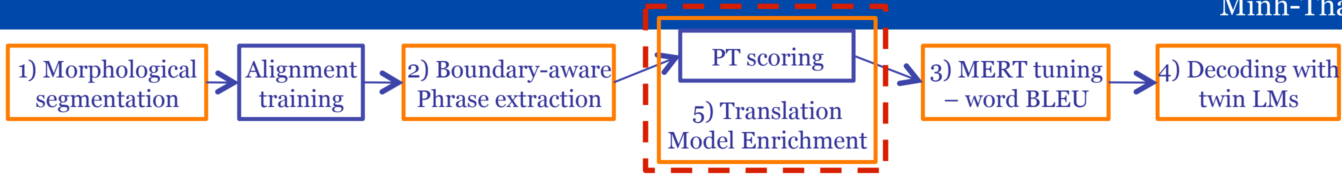
Previous hypotheses — *Current hypothesis*

uusi$_{STM}$, epä$_{PRE+}$ demokraat$_{STM+}$ t$_{SUF+}$ i$_{SUF+}$ s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$ politiikan$_{STM}$

- Morpheme LM: "s$_{SUF+}$ en$_{SUF}$ maahanmuutto$_{PRE+}$" ; "en$_{SUF}$ maahanmuutto$_{PRE+}$ politiikan$_{STM}$"

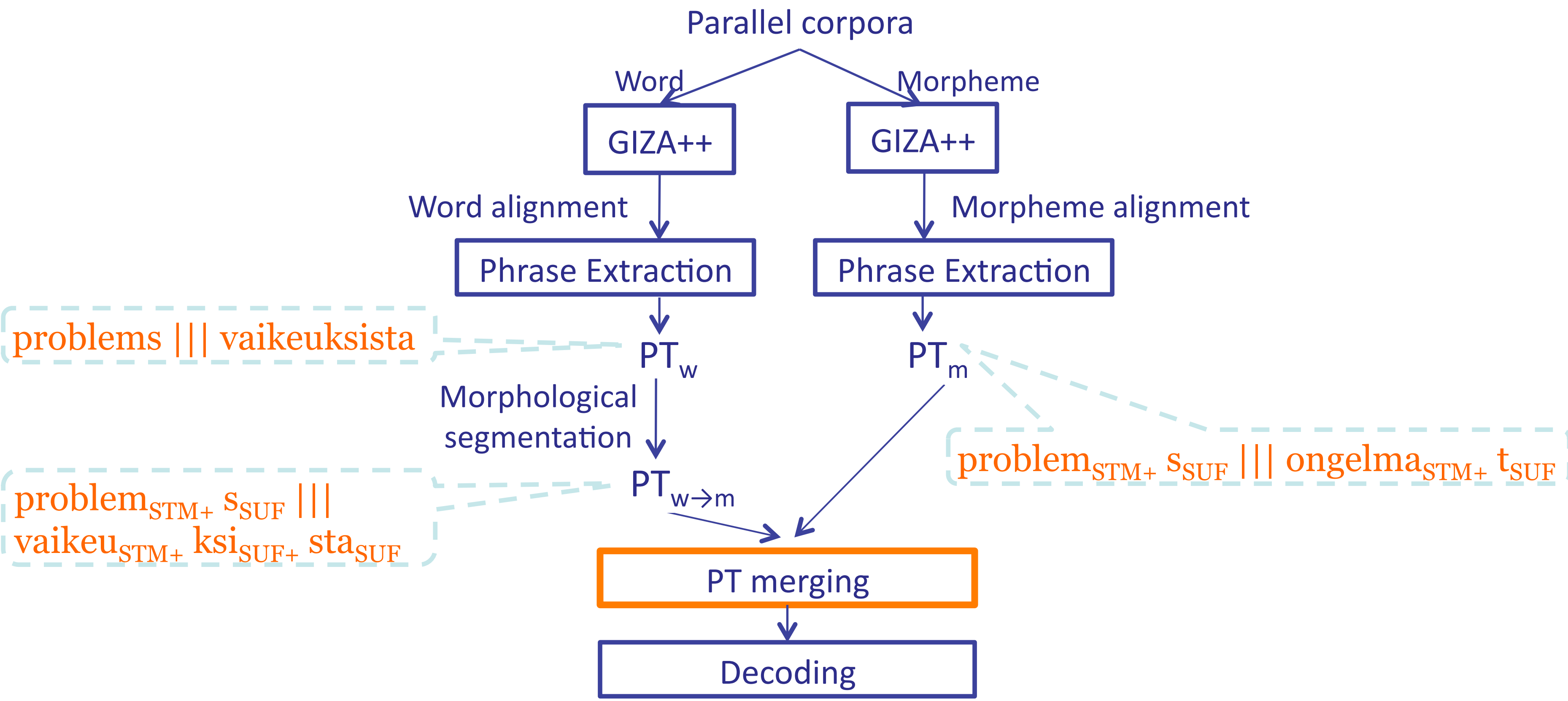
This is (1) different from scoring with two word-level LMs & (2) superior to n-best rescoring.

DECODING



NUS National University of Singapore

Building Twin Translation Models



From the same source, we generate two translation models.

ENRICHING

Phrase Table (PT) Merging

Phrase translation probabilities — *Lexicalized* translation probabilities

problem$_{STM+}$ s$_{SUF}$ ||| vaikeu$_{STM+}$ ksi$_{SUF}$+ sta$_{SUF}$ ||| 0.07 *0.11* 0.01 *0.01* 2.7
problem$_{STM+}$ s$_{SUF}$ ||| ongelma$_{STM+}$ t$_{SUF}$ ||| 0.37 *0.60* 0.11 *0.14* 2.7

Phrase penalty

- **Add-feature methods** e.g., (Chen et al., 2009)

in the 1st PT

problem$_{STM+}$ s$_{SUF}$ ||| vaikeu$_{STM+}$ ksi$_{SUF}$+ sta$_{SUF}$ ||| 0.07 *0.11* 0.01 *0.01* 2.7 **2.7** 1
problem$_{STM+}$ s$_{SUF}$ ||| ongelma$_{STM+}$ t$_{SUF}$ ||| 0.37 *0.60* 0.11 *0.14* 2.7 1 **2.7**

→ Heuristic-driven

in the 2nd PT

- **Interpolation-based methods** e.g., (Wu & Wang, 2007)
  - Linear interpolation of phrase and lexicalized translation probabilities

→ For two PTs originating from different sources

We take into account the fact that
our twin translation models are of equal quality.

ENRICHING

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

Our Method: Phrase Translation Probabilities

- **Preserve the normalized ML estimations (Koehn et al., 2003)**

$$\phi(\bar{f}|\bar{e}) = \frac{\#(\bar{f},\bar{e})}{\sum_{\bar{f}} \#(\bar{f},\bar{e})}$$

The number of times the pair $(\bar{f},\bar{e})$ was extracted from the training dataset

- **Use the raw counts of both models to compute**

$$\phi(\bar{f},\bar{e}) = \frac{\#_m(\bar{f},\bar{e}) + \#_{w\rightarrow m}(\bar{f},\bar{e})}{\sum_{\bar{f}} \#_m(\bar{f},\bar{e}) + \sum_{\bar{f}} \#_{w\rightarrow m}(\bar{f},\bar{e})}$$

PT_m $\quad$ $PT_{w\rightarrow m}$

ENRICHING

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Lexicalized Translation Probabilities

- **Use linear interpolation**
- **What if a phrase pair belongs to one PT only?**

problem$_{STM+}$ s$_{SUF}$ ||| vaikeu$_{STM+}$ ksi$_{SUF+}$ sta
problem$_{STM+}$ s$_{SUF}$ ||| ongelma$_{STM+}$ t$_{SUF}$

PT$_m$

problem$_{STM+}$ s$_{SUF}$ ||| vaikeu$_{STM+}$ ksi$_{SUF+}$ sta
problem$_{STM+}$ s$_{SUF}$ ||| ongelma$_{STM+}$ sta$_{SUF}$

PT$_{w\rightarrow m}$

- **Previous methods: interpolate with 0**
  → Might cause some good phrases to be penalized
- **Our method: induce all scores before interpolation**
  → Use the lexical model of one PT to score phrase pairs for the other one

$$\mathrm{lex}(\bar{f}|\bar{e}) = \alpha \times \mathrm{lex}_m(\bar{f}_m|\bar{e}_m) + (1-\alpha) \times \mathrm{lex}_w(\bar{f}_w|\bar{e}_w)$$

(problem$_{STM+}$ s$_{SUF}$ | ongelma$_{STM+}$ t$_{SUF}$) ...eme Le... Model (PT$_m$)

(problem$_{STM+}$ s$_{SUF}$ | ongelma$_{STM+}$ t$_{SUF}$)

Word Lexical M...el (PT$_w$)

(problems | ongelmat)

ENRICHING

Dataset & Settings

- **Dataset**
  - Past shared task WPT05 (en/fi)
  - 714K sentence pairs
  - Split into T1, T2, T3, and T4 of sizes 40K, 80K, 160K, and 320K

- **Standard phrase-based SMT settings:**
  - Moses
  - IBM Model 4
  - Case insensitive BLEU

SMT Baseline Systems

	w-system		m-system	
	BLEU	m-BLEU	BLEU	m-BLEU
T1	11.56	45.57	11.07	49.15
T2	12.95	48.63	12.68	53.78
T3	13.64	50.30	13.32	54.40
T4	14.20	50.85	13.57	54.70
Full	14.58	53.05	14.08	55.26

- **w-system:** word level
- **m-system:** morpheme level
- **m-BLEU:** morpheme version of BLEU

Either the *m-system* does not perform as well as the *w-system* *or* BLEU is not capable of measuring morpheme improvements.

1) Morphological segmentation → Alignment training → 2) Boundary-aware Phrase extraction → PT scoring / 5) Translation Model Enrichment → 3) MERT tuning – word BLEU → 4) Decoding with twin LMs

NUS National University of Singapore

Morphological Enhancements: Individual

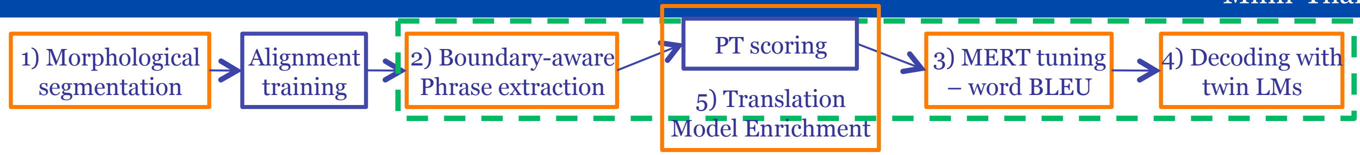
System	T1 (40K)	Full (714K)
w-system	11.56	14.58
m-system	11.07	14.08
m+phr	$11.44^{+0.37}$	$14.43^{+0.35}$
m+tune	$11.73^{+0.66}$	$14.55^{+0.47}$
m+lm	$11.58^{+0.51}$	$14.53^{+0.45}$

- **phr:** boundary-aware phrase extraction
- **tune:** MERT tuning for word BLEU
- **lm:** decoding with twin LMs

individual

The individual enhancements yield improvements for both small and large corpora.

EXPERIMENTS



Morphological Enhancements: Combined

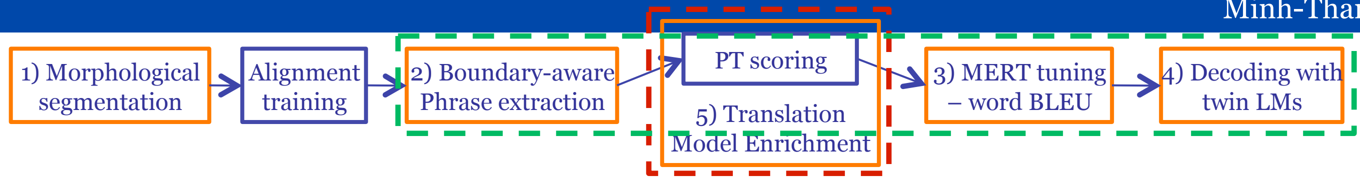
System	T1 (40K)	Full (714K)
w-system	11.56	14.58
m-system	11.07	14.08
m+phr+lm	$11.77^{+0.70}$	$\mathbf{14.58^{+0.50}}$
m+phr+lm+tune	$\mathbf{11.90^{+0.83}}$	$14.39^{+0.31}$

combined

- **phr:** boundary-aware phrase extraction
- **tune:** MERT tuning for word BLEU
- **lm:** decoding with twin LMs

The morphological enhancements are on par with the *w-system* and yield sizable improvements over the *m-system.*

EXPERIMENTS



Translation Model Enrichment

Merging methods	**Full (714K)**
m-system	14.08
w-system	14.58
add-1	$14.25^{+0.17}$
add-2	$13.89^{-0.19}$
interpolation	$14.63^{+0.55}$
ourMethod	$\mathbf{14.82^{+0.74}}$

- **add-1:** *one* extra feature
- **add-2:** *two* extra features
- **Interpolation:** linear interpolation
- **ourMethod**

Our method outperforms the *w-system* baseline.

Results Significance

	BLEU
m-system	14.08
w-system	14.58
ourSystem	**$14.82^{+0.74}$**

- **Absolute improvement of 0.74 BLEU over the *m-system*,**
  - non-trivial relative improvement of 5.6%
- **Outperformed the *w-system* by 0.24 points (1.56% relative)**
- **Statistically significant with $p < 0.01$ (Collins' sign test)**

Translation Proximity Match

- **Automatically extract phrase triples**

$\underbrace{\text{economic and social}}_{\text{src}}$, $\underbrace{\text{taloudellisia ja sosiaalisia}}_{\text{out}}$, $\underbrace{\text{taloudellisten ja sosiaalisten}}_{\text{ref}}$

- high character-level similarity between *out* & *ref*
- longest common subsequence ratio

- **16,262 triples:** 6,758 match the references exactly
- the remaining triples were close wordforms

Hypothesis: our approach yields translations close to the reference wordforms, but it is unjustly penalized by BLEU.

ANALYSIS

Human Evaluation

	our	vs. *m*	*our*	vs. *w*	*w*	vs. *m*
Judge 1	25	18	19	12	21	19
Judge 2	24	16	19	15	25	14
Judge 3	27†	12	17	11	27†	15
Judge 4	25	20	26†	12	22	22
Total	101‡	66	81‡	50	95†	70

- **4 native Finnish speakers**
  - 50 random test sentences
  - follow WMT'09 evaluation:
    - provided judges with the source sentence, its reference translation & outputs of (*m-system*, *w-system*, *ourSystem*) shown in random order
    - asked for three pairwise judgments

The judges consistently preferred: (1) *ourSystem* to the *m-system*, (2) *ourSystem* to the *w-system*, (3) *w-system* to the *m-system*.

Sample Translations

src: we were very constructive and we negotiated until the last minute of these talks in the hague .
ref: olimme erittain **rakentavia** ja neuvottelimme haagissa **viime hetkeen saakka** .
our: olemme olleet hyvin **rakentavia** ja olemme neuvotelleet **viime hetkeen saakka** naiden neuvottelujen haagissa .
w : ol Wrong case in **rakentavia** ja olemme neuvotelleet viime *tippaan niin* naiden neuvo Confusing meaning
m : olimme erittain *rakentavan* ja neuvottelimme **viime hetkeen saakka** naiden neuvotteluiden haagissa .
***Rank*: our > m ≥ w**

Match reference

src: it would be a very dangerous situation if the europeans were to become logistically reliant on russia .
ref: olisi **erittäin** vaarallinen tilanne , jos **eurooppalaiset** tulisivat **logistisesti** riippuvaisiksi venäjästä .
our: olisi **erittäin** vaarallinen tilanne , jos **eurooppalaiset** tulee **logistisesti** riippuvaisia venäjän .
w : *se* olisi **erittäin** vaaral Wrong case s *eurooppalaisten* tulisi *logistically* riippuvaisia OOV .
m : *se* olisi *hyvin* vaarallinen tilanne , jos **eurooppalaiset** *haluavat* tulla **logistisesti** riippuvaisia venäjän .
***Rank*: our > w ≥ m**

Match reference

Change the meaning

ourSystem consistently outperforms the *m-system* & *w-system*, and seems to blend well translations from both baselines.

NUS National University of Singapore

Conclusion

- **Our approach:**
  - The basic unit of translation is the ***morpheme***
  - But ***word*** boundaries are respected at all MT stages
  - Unsupervised method that works for large training bi-texts
- **Future work**:
  - Extend the morpheme-level framework
  - Incorporate morphological analysis directly into the translation process

Thank you!