Four methods for Morpheme-based Machine Translation

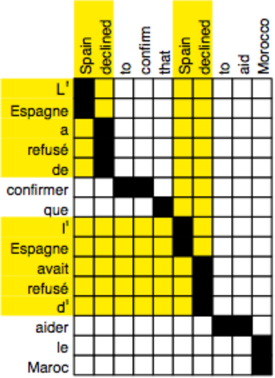
Anoop Sarkar
Joint work with Ann Clifton and Young-chan Kim
Natural Language Lab, Simon Fraser University
http://natlang.cs.sfu.ca

KAIST, July 2, 2012

Phrase-Based Machine Translation (The Basics)

Factored models are a subtype of phrase-based translation models.

Word alignments used to extract all possible phrases (Koehn et al, 2007):



L' Espagne	Spain
L' Espagne a refusé de	Spain declined
L' Espagne a refusé de confirmer	Spain declined to confirm
a refusé de	declined
a refusé de confirmer	declined to confirm
a refusé de confirmer que	declined to confirm that
confirmer	to confirm
confirmer que	to confirm that
que	that
que l' Espagne	that Spain
que l' Espagne avait refusé d'	that Spain declined
l' Espagne	Spain
l' Espagne avait refusé d'	Spain declined
l' Espagne avait refusé d' aider	Spain declined to aid
...	...

Phrase-Based Machine Translation (The Basics)

Phrase pairs along with their probabilities are used to score translation candidates.

Candidate translation $\bar{t}$ given $\bar{s}$ is scored using a log-linear model. $\bar{\lambda}$ are trained to minimize error on a tuning set (Och, 2003):

$$\log \Pr(\bar{t}|\bar{s}) \propto \lambda_1 \sum_{s,t \in \bar{s},\bar{t}} \log \Pr(s|t) + \lambda_2 \sum_{s,t \in \bar{s},\bar{t}} \log \Pr(t|s) + \lambda_3 \log \Pr(\bar{t})$$

Evaluation: BLEU (Papineni et al., 2002)

- Precision-based: counts candidate translation *n*-grams found in multiple reference translations.
- Higher scores are better; for some language pairs (e.g. Chinese-English) the best systems achieve $\sim 40\%$ BLEU

Challenges of Morphological Complexity for SMT

Human languages encode information in very diverse ways:

'ista+hta+isi+n+ko	suure+sa	talo+ssa	.'
sit+FREQ+COND+SG1+INTR	big+IN	house+IN	.

  'Should I sit down for a while in the big house ?'

Challenges of Morphological Complexity for SMT

Human languages encode information in very diverse ways:

- ‘ista+hta+isi+n+ko suure+sa talo+ssa .’
  sit+FREQ+COND+SG1+INTR big+IN house+IN .
  ‘Should I sit down for a while in the big house ?’

SMT should work for any language pair, but is implicitly worse for languages structurally dissimilar to English:

Challenges of Morphological Complexity for SMT

Human languages encode information in very diverse ways:

- ‘ista+hta+isi+n+ko suure+sa talo+ssa .’
  sit+FREQ+COND+SG1+INTR big+IN house+IN .
  ‘Should I sit down for a while in the big house ?’

SMT should work for any language pair, but is implicitly worse for languages structurally dissimilar to English:

- Data Sparsity

Challenges of Morphological Complexity for SMT

Human languages encode information in very diverse ways:

- ‘ista+hta+isi+n+ko suure+sa talo+ssa .’
  sit+FREQ+COND+SG1+INTR big+IN house+IN .
  ‘Should I sit down for a while in the big house ?’

SMT should work for any language pair, but is implicitly worse for languages structurally dissimilar to English:

- Data Sparsity
- Source-Target Asymmetry

Challenges of Morphological Complexity for SMT

Human languages encode information in very diverse ways:

- 'ista+hta+isi+n+ko suure+sa talo+ssa .'
  sit+FREQ+COND+SG1+INTR big+IN house+IN .
  'Should I sit down for a while in the big house ?'

SMT should work for any language pair, but is implicitly worse for languages structurally dissimilar to English:

- Data Sparsity
- Source-Target Asymmetry
- Lack of Resources

Challenges of Morphological Complexity for SMT

Human languages encode information in very diverse ways:

- 'ista+hta+isi+n+ko suure+sa talo+ssa .'
  sit+FREQ+COND+SG1+INTR big+IN house+IN .
  'Should I sit down for a while in the big house ?'

SMT should work for any language pair, but is implicitly worse for languages structurally dissimilar to English:

- Data Sparsity
- Source-Target Asymmetry
- Lack of Resources
- Automatic Evaluation

Four major approaches to Morphology in SMT

- Bilingual Morphology Induction using Parallel Data
  (Chung and Gildea, 2009; Naradowsky and Toutanova, 2011)
- Factored Phrase-Based Translation Models
  (Koehn & Hoang, 2007; Avramidis & Koehn, 2008; Yeniterzi & Oflazer, 2010)
- Segmented Translation
  (Brown et al, 1993; Martin et al, 2003; Goldwater & McClosky, 2005)
- Post-Processing Morphology Generation
  (Minkov, Toutanova & Suzuki, 2007; de Gispert & Mariño, 2008; Toutanova, Suzuki & Ruopp, 2008)

Four major approaches to Morphology in SMT

- Bilingual Morphology Induction using Parallel Data
  (Chung and Gildea, 2009; Naradowsky and Toutanova, 2011)
- Factored Phrase-Based Translation Models
  (Koehn & Hoang, 2007; Avramidis & Koehn, 2008; Yeniterzi & Oflazer, 2010)
- Segmented Translation
  (Brown et al, 1993; Martin et al, 2003; Goldwater & McClosky, 2005)
- Post-Processing Morphology Generation
  (Minkov, Toutanova & Suzuki, 2007; de Gispert & Mariño, 2008; Toutanova, Suzuki & Ruopp, 2008)

This work:

- Unsupervised morphological induction applied to SMT

Four major approaches to Morphology in SMT

- Bilingual Morphology Induction using Parallel Data
  (Chung and Gildea, 2009; Naradowsky and Toutanova, 2011)
- Factored Phrase-Based Translation Models
  (Koehn & Hoang, 2007; Avramidis & Koehn, 2008; Yeniterzi & Oflazer, 2010)
- Segmented Translation
  (Brown et al, 1993; Martin et al, 2003; Goldwater & McClosky, 2005)
- Post-Processing Morphology Generation
  (Minkov, Toutanova & Suzuki, 2007; de Gispert & Mariño, 2008; Toutanova, Suzuki & Ruopp, 2008)

This work:

- Unsupervised morphological induction applied to SMT
- Application to translation **into** languages with complex morphology.

Four major approaches to Morphology in SMT

- Bilingual Morphology Induction using Parallel Data
  (Chung and Gildea, 2009; Naradowsky and Toutanova, 2011)
- Factored Phrase-Based Translation Models
  (Koehn & Hoang, 2007; Avramidis & Koehn, 2008; Yeniterzi & Oflazer, 2010)
- Segmented Translation
  (Brown et al, 1993; Martin et al, 2003; Goldwater & McClosky, 2005)
- Post-Processing Morphology Generation
  (Minkov, Toutanova & Suzuki, 2007; de Gispert & Mariño, 2008; Toutanova, Suzuki & Ruopp, 2008)

This work:

- Unsupervised morphological induction applied to SMT
- Application to translation **into** languages with complex morphology.
- In this talk: experimental results on Korean and Finnish.

Our Contributions

We find that:

- Bidirectional bilingual segmentation outperforms bilingual segmentation in one direction. (Korean)

Our Contributions

We find that:

- Bidirectional bilingual segmentation outperforms bilingual segmentation in one direction. (Korean)
- A translation model trained on morphologically segmented data outperforms a word-based model.

Our Contributions

We find that:

- Bidirectional bilingual segmentation outperforms bilingual segmentation in one direction. (Korean)
- A translation model trained on morphologically segmented data outperforms a word-based model.
- Can be done using unsupervised methods of morphological segmentation, without reliance on expensive, rare supervised segmentation methods.

Our Contributions

We find that:

- Bidirectional bilingual segmentation outperforms bilingual segmentation in one direction. (Korean)
- A translation model trained on morphologically segmented data outperforms a word-based model.
- Can be done using unsupervised methods of morphological segmentation, without reliance on expensive, rare supervised segmentation methods.
- We can successfully apply these unsupervised morpheme segmentation methods to machine translation.

Our Contributions

We find that:

- Bidirectional bilingual segmentation outperforms bilingual segmentation in one direction. (Korean)
- A translation model trained on morphologically segmented data outperforms a word-based model.
- Can be done using unsupervised methods of morphological segmentation, without reliance on expensive, rare supervised segmentation methods.
- We can successfully apply these unsupervised morpheme segmentation methods to machine translation.
- Combining segmented translation with post-processing morphology generation (morphological awareness inside the MT model as well as in post-processing) improves translation morphological fluency. (Finnish)

Unsupervised Monolingual Morphology Induction

- **Input English**: but he failed
- **Output Korean**: 그러나 실패했다
- Consider monolingual segmentation of 실패했다

1. s(0001) 실패했다
2. s(0011) 실패했 다
3. s(0101) 실패 했다

   ⋮

16. s(1111) 실 패 했 다

Unsupervised Monolingual Morphology Induction

- **Input English**: but he failed
- **Output Korean**: 그러나 실패했다
- Consider monolingual segmentation of 실패했다

1. s(0001) 실패했다
2. s(0011) 실패했 다
3. s(0101) 실패 했다

⋮

16. s(1111) 실 패 했 다

1. P(실패했다)
2. P(실패했) P(다)
3. P(실패) P(했다)

⋮

16. P(실) P(패) P(했) P(다)

- Given a sequence of character $c_1, c_2, \ldots, c_n$ find the most likely segmentation s^*

Unsupervised Monolingual Morphology Induction

- **Input English**: but he failed
- **Output Korean**: 그러나 실패했다
- Consider monolingual segmentation of 실패했다

1. s(0001) 실패했다
2. s(0011) 실패했 다
3. s(0101) 실패 했다

⋮

16. s(1111) 실 패 했 다

1. P(실패했다)
2. P(실패했) P(다)
3. P(실패) P(했다)

⋮

16. P(실) P(패) P(했) P(다)

- Given a sequence of character $c_1, c_2, \ldots, c_n$ find the most likely segmentation s^*
- $s^* = \text{argmax}_{k_1, k_2, \ldots, k_\ell} P(k_1)P(k_2)\ldots P(k_\ell)$

Unsupervised Bilingual Morphology Induction

- source language sentence e, target language sentence k
- Start with IBM Model 1: $P(\bar{k} \mid \bar{e}) = P(\bar{k})P(\bar{e} \mid \bar{k})$

$$P(\bar{e} \mid \bar{k}) =$$

Unsupervised Bilingual Morphology Induction

- source language sentence e, target language sentence k
- Start with IBM Model 1: $P(\bar{k} \mid \bar{e}) = P(\bar{k})P(\bar{e} \mid \bar{k})$

$$P(\bar{e} \mid \bar{k}) = \sum_{a} P(\bar{a}, \bar{e} \mid \bar{k})$$

Unsupervised Bilingual Morphology Induction

- source language sentence e, target language sentence k
- Start with IBM Model 1: $P(\bar{k} \mid \bar{e}) = P(\bar{k})P(\bar{e} \mid \bar{k})$

$$P(\bar{e} \mid \bar{k}) =$$

$$\sum_a P(\bar{a}, \bar{e} \mid \bar{k})$$

$$\sum_a \prod_j \underbrace{P(e_j \mid k_{a_j})}$$

e_j : English word at position j

k_{a_j} : Korean word aligned to English word e_j

Unsupervised Bilingual Morphology Induction

- source language sentence e, target language sentence k
- Start with IBM Model 1: $P(\bar{k} \mid \bar{e}) = P(\bar{k})P(\bar{e} \mid \bar{k})$

$$P(\bar{e} \mid \bar{k}) = \sum_{a} P(\bar{a}, \bar{e} \mid \bar{k}) \quad \sum_{a} \prod_{j} \underbrace{P(e_j \mid k_{a_j})}$$

e_j : English word at position j

k_{a_j} : Korean word aligned to English word e_j

- For each e_j, k_i we can learn $P(e_j \mid k_i)$ using the EM algorithm, e.g.:

$$P(\text{failed} \mid \text{실패했다}) = \frac{ec(\text{failed}, \text{실패했다})}{\sum_e ec(e, \text{실패했다})}$$

Unsupervised Bilingual Morphology Induction

- Input English: but he failed
- Output Korean: 그러나 실패했다
- Bilingual segmentation of 실패했다

1. s(0001) 실패했다
2. s(0011) 실패했 다
3. s(0101) 실패 했다

   ⋮

16. s(1111) 실 패 했 다

s(0101) 실패 했다

$$P(\text{failed} \mid \text{실패}) = \frac{ec(\text{failed}, \text{실패})}{\sum_e ec(e, \text{실패})}.$$

- Once we have learned $P(e \mid k)$ using EM, we can obtain the segmentation based on the best alignment $\bar{a}^*$:

$$\bar{a}^* = \operatorname*{argmax}_a P(\bar{e}, \bar{a} \mid \bar{k})$$

Unsupervised Bilingual Morphology Induction

- This modified IBM Model 1 searches over alignments that align each English word with every possible substring of each Korean word.

Unsupervised Bilingual Morphology Induction

- This modified IBM Model 1 searches over alignments that align each English word with every possible substring of each Korean word.
- This naive exponential time method can be replaced with a dynamic programming algorithm.

Unsupervised Bilingual Morphology Induction

- This modified IBM Model 1 searches over alignments that align each English word with every possible substring of each Korean word.
- This naive exponential time method can be replaced with a dynamic programming algorithm.
- Extension of forward-backward algorithm for HMMs applied to IBM Model 1 training.

Unsupervised Bilingual Morphology Induction

- This modified IBM Model 1 searches over alignments that align each English word with every possible substring of each Korean word.
- This naive exponential time method can be replaced with a dynamic programming algorithm.
- Extension of forward-backward algorithm for HMMs applied to IBM Model 1 training.
- Like (Chung & Gildea, 2009) we also use Variational Bayes instead of EM, and use a length penalty.

Bidirectional Segmentation

- A problem with segmenting only the Korean side:

실패+ +했다

failed NULL

Bidirectional Segmentation

- A problem with segmenting only the Korean side:

실패+ +했다

failed NULL

- **Solution**: Segment the English side as well!

실패+ +했다

fail+ +ed

Bidirectional Segmentation

- A problem with segmenting only the Korean side:

실패+ +했다

failed NULL

- **Solution**: Segment the English side as well!

실패+ +했다

fail+ +ed

- To segment the English text, we need bilingual data, but at test time we only have English input!

Bidirectional Segmentation

- A problem with segmenting only the Korean side:

실패+ +했다

failed NULL

- **Solution**: Segment the English side as well!

실패+ +했다

fail+ +ed

- To segment the English text, we need bilingual data, but at test time we only have English input!
- **Solution**: get monolingual probabilities from the bilingual model:

$$P(e) = \sum_{k} \underbrace{P(k \mid e)}_{\text{Korean input with English output}} P(k)$$

English-Korean Experiments

- Datasets (sizes shown refer to the English side):
  1. For the morph segmenter: KAIST corpus, 590,000 words & 60,000 sentence pairs
  2. For training the SMT system: URochester corpus, 2,000,000 words & 60,000 sentence pairs

English-Korean Experiments

- Datasets (sizes shown refer to the English side):
  1. For the morph segmenter: KAIST corpus, 590,000 words & 60,000 sentence pairs
  2. For training the SMT system: URochester corpus, 2,000,000 words & 60,000 sentence pairs

Model	BLEU Score
Baseline	3.13
Monolingual(3)	3.25
Monolingual(5)	3.29
Monolingual(10)	3.20
Single-direction Bilingual	3.36
Bidirectional Bilingual (without VB)	3.22
Bidirectional Bilingual (with VB)	**3.46**

Table: BLEU Scores for English-Korean. Boldface indicates a result significantly better than the baseline.

Example Segmentations

English	there is no way that software can upgrade itself just by changing the hardware .
Korean	하드웨어를 바꾼다고 소프트웨어가 저절로 업그레이드될 리 없다 .

Example Segmentations

English	there is no way that software can upgrade itself just by changing the hardware .
Korean	하드웨어를 바꾼다고 소프트웨어가 저절로 업그레이드될 리 없다 .
Monolingual	하드웨+ +어를 바꾼+ +다고 소프트웨+ +어가 저+ +절로 업그레이+ +드될 리 없다 .

Example Segmentations

English	there is no way that software can upgrade itself just by changing the hardware .
Korean	하드웨어를 바꾼다고 소프트웨어가 저절로 업그레이드될 리 없다 .
Monolingual	하드웨+ +어를 바꾼+ +다고 소프트웨+ +어가 저+ +절로 업그레이+ +드될 리 없다 .
Bidirectional Bilingual	There is no way that software can up+ +grade itself just by changing the hardware .
Bidirectional Bilingual	하드웨어+ +를 바꾼다+ +고 소프트웨어+ +가 저절로 업그레이드될 리 없다 .

Four major approaches to Morphology in SMT

- ~~Bilingual Morphology Induction using Parallel Data~~
  ~~(Chung and Gildea, 2009; Naradowsky and Toutanova, 2011)~~
- Factored Phrase-Based Translation Models
  (Koehn & Hoang, 2007; Avramidis & Koehn, 2008; Yeniterzi & Oflazer, 2010)
- Segmented Translation
  (Brown et al, 1993; Martin et al, 2003; Goldwater & McClosky, 2005)
- Post-Processing Morphology Generation
  (Minkov, Toutanova & Suzuki, 2007; de Gispert & Mariño, 2008; Toutanova, Suzuki & Ruopp, 2008)

English-Finnish Experiments

Dataset: the European parliamentary proceedings corpus Europarl (1.2 million sentence parallel corpus), filtered for sentence length 40; 2,000 sentence dev and test sets

SMT system: We use the open-source phrase-based MT system Moses, (`http://statmt.org/moses`), with default settings

Evaluation: BLEU (Papineni et al., 2002)

- In our experiments, we only have one reference translation.

Why Finnish?

Why Finnish?

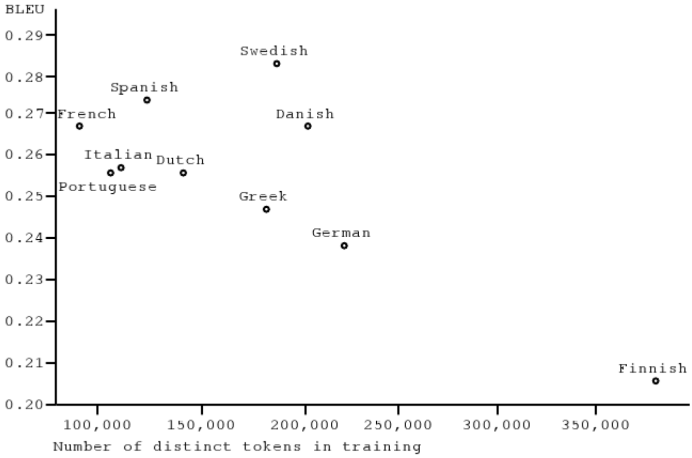


Figure: Translation scores (into English) v.s. No. of training tokens in the European Parliament corpus (Koehn, 2005)

Factored Models

Phrase pairs along with their probabilities are used to score translation candidates.

Candidate translation $\bar{t}$ given $\bar{s}$ is scored using a log-linear model. $\bar{\lambda}$ are trained to minimize error on a tuning set (Och, 2003):

$$\log \Pr(\bar{t}|\bar{s}) \propto \lambda_1 \sum_{s,t \in \bar{s},\bar{t}} \log \Pr(s|t) + \lambda_2 \sum_{s,t \in \bar{s},\bar{t}} \log \Pr(t|s) + \lambda_3 \log \Pr(\bar{t})$$

Factored Models

Phrase pairs along with their probabilities are used to score translation candidates.

Candidate translation $\bar{t}$ given $\bar{s}$ is scored using a log-linear model. $\bar{\lambda}$ are trained to minimize error on a tuning set (Och, 2003):

$$\log \Pr(\bar{t}|\bar{s}) \propto \lambda_1 \sum_{s,t \in \bar{s},\bar{t}} \log \Pr(s|t) + \lambda_2 \sum_{s,t \in \bar{s},\bar{t}} \log \Pr(t|s) + \lambda_3 \log \Pr(\bar{t})$$

In factored models,

- Words represented as (more general) factors, e.g., 'daughters|daughter|NN|PL'
- phrase translation probabilities are decomposed over the available factors for the words in the phrase.
- probabilistic generative model

Using Factored Models for Finnish

Words and Part-of-Speech (POS) tags for the source side factors,
Words, stems, and unsupervised morph segments for target factors:

'daughters⇒NNS'   'tyttäret⇒tyttär|et'

Factored models: prob. generative model for phrase pairs,
easily incorporated into phrase-based SMT decoders:

Using Factored Models for Finnish

Words and Part-of-Speech (POS) tags for the source side factors,
Words, stems, and unsupervised morph segments for target factors:
'daughters⇒NNS' 'tyttäret⇒tyttär|et'
Factored models: prob. generative model for phrase pairs,
easily incorporated into phrase-based SMT decoders:

1. translate source word to target stem
   daughters → tyttär

Using Factored Models for Finnish

Words and Part-of-Speech (POS) tags for the source side factors,
Words, stems, and unsupervised morph segments for target factors:
'daughters⇒NNS' 'tyttäret⇒tyttär|et'
Factored models: prob. generative model for phrase pairs,
easily incorporated into phrase-based SMT decoders:

1. translate source word to target stem
   daughters → tyttär
2. generate set of possible suffixes for the target stem
   tyttär|et (plural) tytär|ϕ (singular)
   tyttär|eni (my daughters)

Using Factored Models for Finnish

Words and Part-of-Speech (POS) tags for the source side factors,
Words, stems, and unsupervised morph segments for target factors:

'daughters$\Rightarrow$NNS' 'tyttäret$\Rightarrow$tyttär|et'

Factored models: prob. generative model for phrase pairs,
easily incorporated into phrase-based SMT decoders:

1. translate source word to target stem
   daughters $\rightarrow$ tyttär
2. generate set of possible suffixes for the target stem
   tyttär|et (plural) tytär|ϕ (singular)
   tyttär|eni (my daughters)
3. translate source POS tag to target suffix
   NNS $\rightarrow$ et NNS $\rightarrow$ ϕ
   NNS $\rightarrow$ eni

Using Factored Models for Finnish

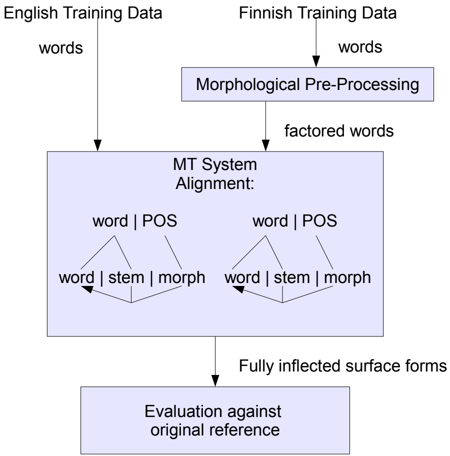
Words and Part-of-Speech (POS) tags for the source side factors,
Words, stems, and unsupervised morph segments for target factors:

‘daughters⇒NNS’ ‘tyttäret⇒tyttär|et’

Factored models: prob. generative model for phrase pairs, easily incorporated into phrase-based SMT decoders:

1. translate source word to target stem
   daughters $\rightarrow$ tyttär
2. generate set of possible suffixes for the target stem
   tyttär|et (plural) tytär|ϕ (singular)
   tyttär|eni (my daughters)
3. translate source POS tag to target suffix
   NNS $\rightarrow$ et NNS $\rightarrow$ ϕ
   NNS $\rightarrow$ eni
4. generate most likely surface form from stem and suffix
   tyttär|et $\rightarrow$ tyttäret

Model Comparison: Factored Translation



Better Unsupervised Monolingual Segmentation

⇒ Morfessor segmentation model (Creutz and Lagus, 2005):

(1) Input sentence: monet meistä haluavat kansallisten jäsenvaltioiden muodostamaa liittovaltiota .

(2) Segmented output: monet me+ +istä haluavat kansallisten jäsen+ +valtioiden muodo+ +sta+ +ma+ +a liittovaltiota .

Better Unsupervised Monolingual Segmentation

$\Rightarrow$ Morfessor segmentation model (Creutz and Lagus, 2005):

(3) Input sentence: monet meistä haluavat kansallisten jäsenvaltioiden muodostamaa liittovaltiota .

(4) Segmented output: monet me+ +istä haluavat kansallisten jäsen+ +valtioiden muodo+ +sta+ +ma+ +a liittovaltiota .

F-score of $\sim$ 70% trained on a 16 million word corpus (1.4M types) against human-annotated gold-standard.

Better Unsupervised Monolingual Segmentation

$\Rightarrow$ Morfessor segmentation model (Creutz and Lagus, 2005):

(5) Input sentence: monet meistä haluavat kansallisten jäsenvaltioiden muodostamaa liittovaltiota .

(6) Segmented output: monet me+ +istä haluavat kansallisten jäsen+ +valtioiden muodo+ +sta+ +ma+ +a liittovaltiota .

F-score of $\sim$ 70% trained on a 16 million word corpus (1.4M types) against human-annotated gold-standard.

We varied this model over granularity and coverage parameters to get the best segmentation for our MT system.

Factored Models: Results and Analysis

Model	BLEU Score
Baseline	14.39
Factored	13.98

Table: Factored Model BLEU Scores

Factored model analysis:

- Difficult representational form for language with this degree of morphological complexity

Factored Models: Results and Analysis

Model	BLEU Score
Baseline	14.39
Factored	13.98

Table: Factored Model BLEU Scores

Factored model analysis:

- Difficult representational form for language with this degree of morphological complexity
- Generates candidate translation phrase pairs on the fly;
  $\Rightarrow$ multiple generation steps and large suffix set cause combinatorial explosion

Factored Models: Results and Analysis

Model	BLEU Score
Baseline	14.39
Factored	13.98

Table: Factored Model BLEU Scores

Factored model analysis:

- Difficult representational form for language with this degree of morphological complexity
- Generates candidate translation phrase pairs on the fly; $\Rightarrow$ multiple generation steps and large suffix set cause combinatorial explosion
- Productive morphology limited to phrase pairs

Factored Models: Results and Analysis

Model	BLEU Score
Baseline	14.39
Factored	13.98

Table: Factored Model BLEU Scores

Factored model analysis:

- Difficult representational form for language with this degree of morphological complexity
- Generates candidate translation phrase pairs on the fly;
  $\Rightarrow$ multiple generation steps and large suffix set cause combinatorial explosion
- Productive morphology limited to phrase pairs
- No long-distance dependencies between morphemes

Segmented Translation: The Goal

- Segment and translate target morphology with lexical equivalents in source, while generalizing over morphological information that we can't translate.
- By treating segmented forms the same as stems, allow productive combination of morphs and stems across phrase boundaries

Example:

(7) ‘ista+ $\boxed{\text{hta}}$ + $\boxed{\text{isi}}$ + $\boxed{\text{n}}$ + $\boxed{\text{ko}}$ talo+ $\boxed{\text{ssa}}$.’
sit+ $\boxed{\text{FREQ}}$ + $\boxed{\text{COND}}$ + $\boxed{\text{SG1}}$ + $\boxed{\text{INTER}}$ house+ $\boxed{\text{IN}}$.
‘ $\boxed{\text{Should}}$ $\boxed{\text{I}}$ sit down $\boxed{\text{for a while}}$ $\boxed{\text{in}}$ the house $\boxed{\text{?}}$ ’

Conditional, interrogative, and case markers all have lexical reflexes in the English parallel sentence.

Segmented Translation: The Approach

Segment corpus before training translation model, treating morphs like words. (Oflazer and El-Kahlout, 2007; Virpioja et al., 2007)

Supervised or unsupervised means of segmentation:

- 'toimivaltaan'
  - Supervised: Omorfi (Pirinen and Listenmaa, 2007)
    toimia|VERB/ACT/INF/SG/ABL/POSS
    toimiv+ +altaan
  - Unsupervised (Morfessor)
    toimi+ +valtaa+ +n

Segmented Translation: The Approach

Segment corpus before training translation model, treating morphs like words. (Oflazer and El-Kahlout, 2007; Virpioja et al., 2007)

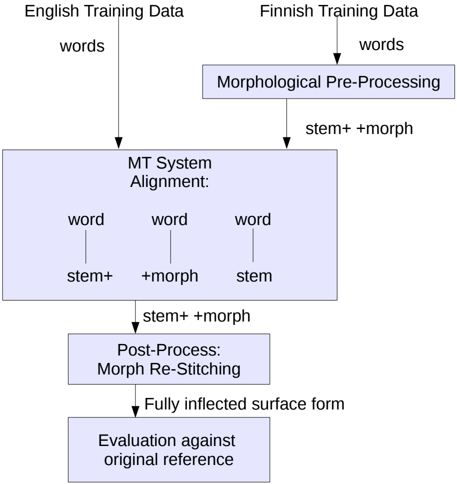
Supervised or unsupervised means of segmentation:

- 'toimivaltaan'
  - Supervised: Omorfi (Pirinen and Listenmaa, 2007)
    toimia|VERB/ACT/INF/SG/ABL/POSS
    toimiv+ +altaan
  - Unsupervised (Morfessor)
    toimi+ +valtaa+ +n

The MT system trained on the segmented model ('Unsup') generates a phrase table with $\sim$ 17% of phrases bounded by a productive word-internal morph ('Hanging Morph').

	Unsup
Total	64,106,047
Morph	30,837,615
Hanging Morph	10,906,406

Model Comparison: Segmented Translation



Segmented Translation: Experiments

- Frequently seen complex words may be left unsegmented, but the segmentation model remembers their hierarchical substructure
- Because common (complex) words have their own entries in the segmentation lexicon, Morfessor tends to undersegment
- To make it segment more aggressively, we devised a longest-suffix-substring-match heuristic ('L-match') to do additional word splitting, e.g.,

regular segmentation: '**euroopan**'
substring matching: unsegmented '**euroopan**' matches suffix '**-an**'
$\Rightarrow$ additionally split segmentation: '**euroop+ +an**'

Segmented Translation: Results (1)

Segmentation	BLEU Score
Baseline	14.39
Factored	13.98
Sup	14.58
Luong et al (2010)	14.82
Unsup	**14.94**
Unsup L-match	**15.09**

- Boldface indicates a result significantly better than the baseline.
- We also outperform other systems on WER and TER.

Segmented Translation: Results (2)

Segmentation	*m*-BLEU Score	
Baseline	14.84	
Sup	18.41	
Unsup	16.07	
Unsup L-match	20.74	
Luong at al (2010)	**55.64**	

- *m*-**BLEU**: output was evaluated in segmented form before morph re-stitching against a segmented version of the baseline *and* reference.

Segmented Translation: Results (2)

Segmentation	*m*-BLEU Score	No Unigram
Baseline	14.84	9.89
Sup	18.41	13.49
Unsup	16.07	10.46
Unsup L-match	20.74	**15.89**
Luong at al (2010)	**55.64**	-

- *m*-**BLEU**: output was evaluated in segmented form before morph re-stitching against a segmented version of the baseline *and* reference.
- **No Unigram**: *m*-BLEU scores without unigrams.

Segmented Translation: Results (2)

Segmentation	*m*-BLEU Score	No Unigram
Baseline	14.84	9.89
Sup	18.41	13.49
Unsup	16.07	10.46
Unsup L-match	20.74	**15.89**
Luong at al (2010)	**55.64**	-

- *m*-**BLEU**: output was evaluated in segmented form before morph re-stitching against a segmented version of the baseline *and* reference.
- **No Unigram**: *m*-BLEU scores without unigrams.
- 39.75% percent improvement over baseline *m*-BLEU compared to Luong et al (2010) with 0.6% percent improvement.

Segmented Translation: Results (2)

Segmentation	*m*-BLEU Score	No Unigram
Baseline	14.84	9.89
Sup	18.41	13.49
Unsup	16.07	10.46
Unsup L-match	20.74	**15.89**
Luong at al (2010)	**55.64**	-

- *m*-**BLEU**: output was evaluated in segmented form before morph re-stitching against a segmented version of the baseline *and* reference.
- **No Unigram**: *m*-BLEU scores without unigrams.
- 39.75% percent improvement over baseline *m*-BLEU compared to Luong et al (2010) with 0.6% percent improvement.
- *m*-BLEU is not correlated with human judgements.

Unsupervised Segmented Translation: Analysis (1)

(8) a. Input: '..summarises and makes visible **the fundamental rights which the public are entitled to**'

b. Reference: 'kansalaisten/GEN **perusoikeudet/ACC**

(9) a. Regular Unsup: 'perusoikeuksia/PAR ja näkyvyyttä/PAR , johon/ILL kansalaisilla/ADE on **oikeus/NOM**'

b. Back-translation: '**the fundamental rights and visibility, which the citizens have the right**'

(10) a. Unsup L-match: 'perusoikeudet/ACC, jotka kansalaiset/ACC ovat **oikeutettuja/PAR**'

b. Back-translation: '**basic rights that citizens are entitled to**'

Segmentation model remembers substructure of frequent complex words, but does not pass it on to the translation model.
Extra-split segmentation model propagates substructure awareness forward to the translation model,
⇒ more frequent correct case-marking in system output.

Unsupervised Segmented Translation: Analysis (2)

Benefits and drawbacks of segmented translation:

- Model benefits from morphology it can translate, but not morphology without lexical reflexes (e.g., agreement);
- Only able to utilize a coarse-grained segmentation; finer-grained segmentation causes overfitting.
- Model is unaware of the relationship between stems and morphs, and is unable to distinguish between them.

CRFs for Post-Processing Morphology Generation

- In this model, translation is performed on stemmed words, then morphology is predicted on MT output, using a Conditional Random Field (CRF).

Word Stem	$s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$
Morph Prediction	y_{t-2}, y_{t-1}, y_t

CRFs for Post-Processing Morphology Generation

- In this model, translation is performed on stemmed words, then morphology is predicted on MT output, using a Conditional Random Field (CRF).

Word Stem	$s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$
Morph Prediction	y_{t-2}, y_{t-1}, y_t

- Using a post-processing morphology generation model allows us to model stems and morphs differently (unlike the segmented models), without relegating morphology generation to phrase pairs only (unlike the factored model).

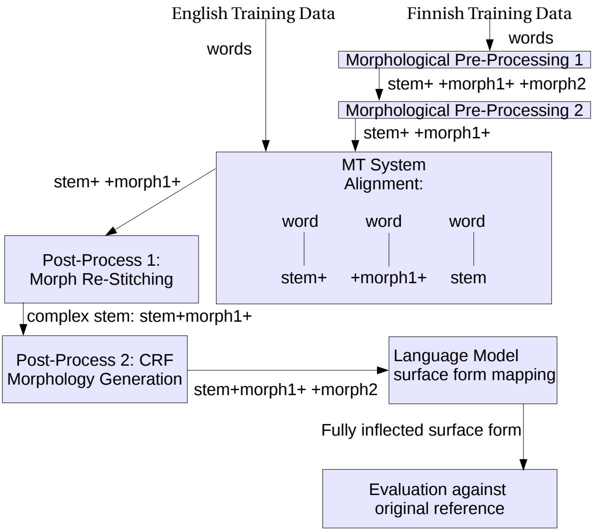
CRFs for Post-Processing Morphology Generation

- In this model, translation is performed on stemmed words, then morphology is predicted on MT output, using a Conditional Random Field (CRF).

Word Stem	$s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$
Morph Prediction	y_{t-2}, y_{t-1}, y_t

- Using a post-processing morphology generation model allows us to model stems and morphs differently (unlike the segmented models), without relegating morphology generation to phrase pairs only (unlike the factored model).
- We preprocess the data in two steps to keep one layer of morphology in the translation model and another for post-processing to capture lexical equivalents as well as target-side morphological dependencies.

Model Comparison: Segmented Translation with Post-Processing Morphology Generation



CRFs for Morphology Prediction

- CRF trained on a $\sim$210K Finnish sentences, consisting of $\sim$1.5 million tokens

CRFs for Morphology Prediction

- CRF trained on a ∼210K Finnish sentences, consisting of ∼1.5 million tokens
- Input symbols: word stems generated by segmentation model: $s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$

CRFs for Morphology Prediction

- CRF trained on a ~210K Finnish sentences, consisting of ~1.5 million tokens
- Input symbols: word stems generated by segmentation model: $s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$
- Output Labels: the morphology set from which a morph is chosen to inflect that stem in its given context: y_{t-2}, y_{t-1}, y_t

CRFs for Morphology Prediction

- CRF trained on a ∼210K Finnish sentences, consisting of ∼1.5 million tokens
- Input symbols: word stems generated by segmentation model: $s_{t-n}, .., s_t, .., s_{t+n} (n = 4)$
- Output Labels: the morphology set from which a morph is chosen to inflect that stem in its given context: y_{t-2}, y_{t-1}, y_t
- Intermediate representation of the morphs to capture more general patterns of morphological distribution.

CRFs for Morphology Prediction

- Intermediate representation of the morphs to capture more general patterns of morphological distribution:

CRFs for Morphology Prediction

- Intermediate representation of the morphs to capture more general patterns of morphological distribution:
  - Supervised model: morphology represented as morph-category-value-tags, e.g.,
    'neljänn**eksi**' $\rightarrow$ 'neljäs +**TRA/SG**'
    'levittämis**eksi** $\rightarrow$ 'levittää +**TRA/SG**'

CRFs for Morphology Prediction

- Intermediate representation of the morphs to capture more general patterns of morphological distribution:
  - Supervised model: morphology represented as morph-category-value-tags, e.g.,
    'neljänn**eksi**' $\rightarrow$ 'neljäs +**TRA/SG**'
    'levittämis**eksi** $\rightarrow$ 'levittää +**TRA/SG**'
  - Unsupervised model: suffixes with vowels collapsed into equivalence classes for vowel harmony, e.g.,
    mietintö**ssä** $\rightarrow$ mietintö+ +**ssA**
    vaihee**ssa** $\rightarrow$ vaihee+ +**ssA**

CRFs for Morphology Prediction

- Intermediate representation of the morphs to capture more general patterns of morphological distribution:
  - Supervised model: morphology represented as morph-category-value-tags, e.g.,
    'neljänn**eksi**' → 'neljäs +**TRA/SG**'
    'levittämis**eksi** → 'levittää +**TRA/SG**'
  - Unsupervised model: suffixes with vowels collapsed into equivalence classes for vowel harmony, e.g.,
    mietintö**ssä** → mietintö+ +**ssA**
    vaihee**ssa** → vaihee+ +**ssA**
- Language model recovers fully inflected surface forms from ambiguous stem+morphology pairs, e.g.,
  'alus+PAR/PL': candidate surface forms 'aluksia' 'aluksiaan', 'aluksiamme';
  language model selects appropriate surface form in the sentence context.

Training with post-processing morphology prediction

original training data:
koskevaa mietintöä käsitellään

Training with post-processing morphology prediction

original training data:
koskevaa mietintöä käsitellään
segmentation:
koske+ +va+ +a mietintö+ +ä käsi+ +te+ +llä+ +ä+ +n

Training with post-processing morphology prediction

original training data:
koskevaa mietintöä käsitellään
segmentation:
koske+ +va+ +a mietintö+ +ä käsi+ +te+ +llä+ +ä+ +n
(train bigram language model on complex stems with suffix)
map final suffix to abstract tag-set (e.g. mapping $A = \{ a, ä \}$):
koskeva+ +A mietintö+ +A käsitelllää+ +n

Training with post-processing morphology prediction

original training data:
koskevaa mietintöä käsitellään
segmentation:
koske+ +va+ +a mietintö+ +ä käsi+ +te+ +llä+ +ä+ +n
(train bigram language model on complex stems with suffix)
map final suffix to abstract tag-set (e.g. mapping A = { a, ä }):
koskeva+ +A mietintö+ +A käsitellää+ +n
(train CRF model to predict the final suffix from full segmentation)
peeling of final suffix:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+

Training with post-processing morphology prediction

original training data:
koskevaa mietintöä käsitellään
segmentation:
koske+ +va+ +a mietintö+ +ä käsi+ +te+ +llä+ +ä+ +n
(train bigram language model on complex stems with suffix)
map final suffix to abstract tag-set (e.g. mapping $A = \{ a, ä \}$):
koskeva+ +A mietintö+ +A käsitellää+ +n
(train CRF model to predict the final suffix from full segmentation)
peeling of final suffix:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
(train SMT model on this transformation of training data)

Decoding with post-processing morphology prediction

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+

Decoding with post-processing morphology prediction

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
decoder output stitched up:
koskeva+ mietintö+ käsitellää+

Decoding with post-processing morphology prediction

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
decoder output stitched up:
koskeva+ mietintö+ käsitellää+
CRF model prediction:
koskeva+ +A mietintö+ +A käsitellää+ +n

Decoding with post-processing morphology prediction

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
decoder output stitched up:
koskeva+ mietintö+ käsitellää+
CRF model prediction:
koskeva+ +A mietintö+ +A käsitellää+ +n
language model disambiguation:
koskeva+ +a mietintö+ +ä käsitellää+ +n

Decoding with post-processing morphology prediction

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
decoder output stitched up:
koskeva+ mietintö+ käsitellää+
CRF model prediction:
koskeva+ +A mietintö+ +A käsitellää+ +n
language model disambiguation:
koskeva+ +a mietintö+ +ä käsitellää+ +n
final stitching:
koskevaa mietintöä käsitellään

Decoding with post-processing morphology prediction

decoder output:
koske+ +va+ mietintö+ käsi+ +te+ +llä+ +ä+
decoder output stitched up:
koskeva+ mietintö+ käsitellää+
CRF model prediction:
koskeva+ +A mietintö+ +A käsitellää+ +n
language model disambiguation:
koskeva+ +a mietintö+ +ä käsitellää+ +n
final stitching:
koskevaa mietintöä käsitellään
(the output is then compared to the reference translation)

Supervised Morphology Generation: Results

Intrinsic evaluation on reference translation:

- 77.18% accuracy predicting the morphology tag sequences.
- comparable to a similar task performed using log-linear models on Russian and Arabic (Minkov, Toutanova, and Suzuki, 2007), languages with simpler morphological systems than Finnish

Extrinsic evaluation after the application of the language model to restore surface form:

Model	BLEU Score
Baseline	14.39
Factored	13.98
Unsup L-match	**15.09**
CRF-Sup	10.09

Table: Supervised Prediction Model BLEU Scores

Unsupervised Morphology Generation: Results

Intrinsic evaluation on the reference translation:

Model	All	Predictions Only
CRF-Unsup	95.61	77.57

Table: Model Accuracy

Unsupervised Morphology Generation: Results

Intrinsic evaluation on the reference translation:

Model	All	Predictions Only
CRF-Unsup	95.61	77.57

Table: Model Accuracy

Extrinsic evaluation on MT system output:

Model	BLEU Score
Baseline	14.39
Factored	13.98
Unsup L-match	**15.09**
CRF-Sup	10.09
CRF-Unsup	**14.55**

Table: Unsupervised Prediction Model BLEU Scores

Unsupervised Morphology Generation: Further Results and Morphological Fluency Analysis (1)

Translation fluency evaluation on the sub-word level:

Construction	Baseline		Unsup L-match		CRF-LM	
	P	R	P	R	P	R
Noun Marking	51.74	78.48	53.11	**83.63**	**54.99**	80.21
Trans Obj	32.35	27.50	33.47	29.64	**35.83**	**30.71**
Noun-Adj Agr	72.75	67.16	69.62	**71.00**	**73.29**	62.58
Subj-Verb Agr	56.61	40.67	55.90	**48.17**	**57.79**	40.17
Postpositions	43.31	29.89	39.31	**36.96**	**47.16**	31.52
Possession	66.67	**70.00**	75.68	**70.00**	**78.79**	60.00

Table: Model Accuracy: Morphological Constructions.

Unsupervised Morphology Generation: Further Results and Morphological Fluency Analysis (2)

Example: postposition ‘kanssa’ requires genitive preceding noun.

Input: ‘with the basque nationalists’:

Reference: ‘baskimaan kansallismielist**en** kanssa’
basque-SG/NOM+land-SG/GEN,ACC
nationalists-PL/**GEN** with-POST

Baseline: ∗‘baskimaan kansallismielis**et** kanssa’
basque-SG/NOM-+land-SG/GEN,ACC
kansallismielinen-PL/**NOM,ACC**-nationalists POST-with

CRF: ‘kansallismielisten baski**en** kanssa’
nationalists-PL/GEN basques-PL/**GEN** with-POST

CRF’s grammatically correct translation scored as incorrect under BLEU because doesn’t match reference.

Conclusion and Future Work

In this work:

- Bilingual unsupervised morpheme segmentation shows promise for Korean MT $\boxed{3.13 \rightarrow 3.46}$
- Morphologically segmented MT model outperforms word-based model for Finnish MT $\boxed{14.39 \rightarrow 15.09}$
- Linguistic analysis of CRF post-processing model output shows improved translation fluency across different morphological constructions

Conclusion and Future Work

In this work:

- Bilingual unsupervised morpheme segmentation shows promise for Korean MT $\boxed{3.13 \rightarrow 3.46}$
- Morphologically segmented MT model outperforms word-based model for Finnish MT $\boxed{14.39 \rightarrow 15.09}$
- Linguistic analysis of CRF post-processing model output shows improved translation fluency across different morphological constructions

In future work:

- Evaluation metrics for incorporating subword information (modified character-level BLEU?)
- Totally integrated model encompassing morphological segmentation, MT system training, and morphology generation

Thank you

감사+ +합니다

Thank you (?)

References I

Avramidis, Eleftherios and Philipp Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, page 763?770, Columbus, Ohio, USA. Association for Computational Linguistics.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chang, Pi-Chuan, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio, June. Association for Computational Linguistics.

References II

Collins, Michael, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics.

Creutz, Mathias and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*, pages 106–113, Espoo, Finland.

Creutz, Mathias and Krista Lagus. 2006. Morfessor in the morpho challenge. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*.

de Gispert, Adriá and José Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*, 50(11-12).

References III

Goldwater, Sharon and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, B.C., Canada. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand. Association for Computational Linguistics.

Koehn, Philipp and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.

References IV

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–108, Prague, Czech Republic. Association for Computational Linguistics.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Francisco, California, USA. Association for Computing Machinery.

References V

Luong, Minh-Thang, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 148–157, Cambridge, Massachusetts. Association for Computational Linguistics.

Ma, Yanjun, Nicolas Stroppa, and Andy Way. 2007. Bootstrapping word alignment via word packing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 304—-311, Prague, Czech Republic. Association for Computational Linguistics.

Minkov, Einat, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL07)*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

References VI

Monson, Christian. 2008. Paramor and morpho challenge 2008. In *Lecture Notes in Computer Science: Workshop of the Cross-Language Evaluation Forum (CLEF 2008), Revised Selected Papers*.

Monson, Christian, Jaime Carbonell, Alon Lavie, and Lori Levin. 2007. Paramor: Minimally supervised induction of paradigm structure and morphological analysis. In *Proceedings of SIGMORPHON*.

Nizar, Habash. 2007. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, Columbus, Ohio. Association for Computational Linguistics.

References VII

Oflazer, Kemal and Ilknur El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 25–32, Prague, Czech Republic, June. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pirinen, Tommi and Inari Listenmaa. 2007. Omorfi morphological analzer. http://gna.org/projects/omorfi.

References VIII

Popoviç, Maja and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1585–1588, Lisbon, Portugal. European Language Resources Association (ELRA).

Ramanathan, Ananthakrishnan, Hansraj Choudhary, Avishek Ghosh, and Pushpak Bhattacharyya. 2009. Case markers and morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 800–808, Suntec, Singapore. Association for Computational Linguistics.

Stolcke, Andreas. 2002. Srilm – an extensible language modeling toolkit. *7th International Conference on Spoken Language Processing*, 3:901–904.

References IX

Talbot, David and Miles Osborne. 2006. Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 969–976, Sydney, Australia, July. Association for Computational Linguistics.

Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 514–522, Columbus, Ohio, USA. Association for Computational Linguistics.

Virpioja, Sami, Jaakko J. Vyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI*.

References X

Yang, Mei and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 41–48, Trento, Italy. Association for Computational Linguistics.