Multi-Source Translation Methods

Lane Schwartz
lane@cs.umn.edu

University of Minnesota

23 Oct 2008

Motivation

Principle of Translational Promiscuity:

If a document is translated into more than 1 language,
it will likely be translated into many more languages.

- Translate into first *n* target languages by hand
- Translate into remaining target languages using MT

Motivation

Principle of Translational Promiscuity:

> If a document is translated into more than 1 language,
> it will likely be translated into many more languages.

- Translate into first n target languages by hand
- Translate into remaining target languages using MT

Motivation

Principle of Translational Promiscuity:

If a document is translated into more than 1 language,
it will likely be translated into many more languages.

- Translate into first n target languages by hand
- Translate into remaining target languages using MT

Related Work

Can using multiple sources of information improve translation?

- Lattice inputs
- Consensus decoding
- Hypothesis ranking

Related Work

Can using multiple sources of information improve translation?

- Lattice inputs
- Consensus decoding
- Hypothesis ranking

Related Work

Can using multiple sources of information improve translation?

- Lattice inputs
- Consensus decoding
- Hypothesis ranking

Related Work

Can using multiple sources of information improve translation?

- Lattice inputs
- Consensus decoding
- Hypothesis ranking

Related Work

Can using multiple sources of information improve translation?

- Lattice inputs
- Consensus decoding
- Hypothesis ranking

Related Work — Lattice Input

- Begin with alternate representations of a source sentence
  Chinese word segmentations
  Arabic morphological analyses
- Align alternate representations into a word lattice
- Use standard decoding algorithms, modified to accept lattice input (Dyer et al., 2008)
- Can be extended to accept multilingual inputs
  (ongoing work by Josh Schroeder)

Related Work — Lattice Input

- Begin with alternate representations of a source sentence
  Chinese word segmentations
  Arabic morphological analyses
- Align alternate representations into a word lattice
- Use standard decoding algorithms, modified to accept lattice input (Dyer et al., 2008)
- Can be extended to accept multilingual inputs
  (ongoing work by Josh Schroeder)

Related Work — Lattice Input

- Begin with alternate representations of a source sentence
  Chinese word segmentations
  Arabic morphological analyses
- Align alternate representations into a word lattice
- Use standard decoding algorithms, modified to accept lattice input (Dyer et al., 2008)
- Can be extended to accept multilingual inputs
  (ongoing work by Josh Schroeder)

Related Work — Lattice Input

- Begin with alternate representations of a source sentence
  Chinese word segmentations
  Arabic morphological analyses
- Align alternate representations into a word lattice
- Use standard decoding algorithms, modified to accept lattice input (Dyer et al., 2008)
- Can be extended to accept multilingual inputs
  (ongoing work by Josh Schroeder)

Related Work — Lattice Input

- Begin with alternate representations of a source sentence
  Chinese word segmentations
  Arabic morphological analyses
- Align alternate representations into a word lattice
- Use standard decoding algorithms, modified to accept lattice input (Dyer et al., 2008)
- Can be extended to accept multilingual inputs
  (ongoing work by Josh Schroeder)

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding
  (Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding (Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding
(Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
    - Translate source text using *n* different systems
    - Align the *n* output hypotheses into a weighted word lattice
    - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
    - Matusov et al. (2006)
    - Japanese and Chinese into English
    - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding
(Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding
(Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding
(Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation

- Bilingual consensus decoding
(Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
    - Translate source text using *n* different systems
    - Align the *n* output hypotheses into a weighted word lattice
    - Intersect word lattice with n-gram language model

- Multilingual consensus decoding
    - Matusov et al. (2006)
    - Japanese and Chinese into English
    - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding (Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Consensus Decoding

- Given a set of translations, find a *consensus* translation
- Bilingual consensus decoding
(Frederking and Nirenburg, 1994; Bangalore et al., 2001; Jayaraman and Lavie, 2005; Rosti et al., 2007)
  - Translate source text using *n* different systems
  - Align the *n* output hypotheses into a weighted word lattice
  - Intersect word lattice with n-gram language model
- Multilingual consensus decoding
  - Matusov et al. (2006)
  - Japanese and Chinese into English
  - 4.8 BLEU improvement

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining [illegible]
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining [illegible]

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining up to 3 languages
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining up to 6 languages

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining up to 3 languages
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining up to 6 languages

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining up to 3 languages
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining up to 6 languages

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining up to 3 languages
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining up to 6 languages

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining up to 3 languages
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining up to 6 languages

Related Work — Hypothesis Ranking

- Given a set of translations, find the *best* translation in the set
- Bilingual language model ranking (Kaki et al., 1999; Callison-Burch and Flourney, 2001)
- Multilingual translation model ranking (Och and Ney, 2001)
  - MAX
    $\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$
    Positive results reported combining up to 3 languages
  - PROD
    $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
    Positive results reported combining up to 6 languages

Experiment — Hypothesis Ranking using an Oracle

- What are the maximum gains possible from hypothesis ranking?
- Oracle experiment — choose hypothesis based on WER distance to the reference.

Experiment — Hypothesis Ranking using an Oracle

- What are the maximum gains possible from hypothesis ranking?
- Oracle experiment — choose hypothesis based on WER distance to the reference.

Phrase-based bilingual systems

languages	BLEU	TER	METEOR
da-en	28.4	57.5	52.9
de-en	27.3	58.9	52.4
el-en	29.3	56.4	53.6
es-en	**32.5**	**52.8**	**56.3**
fi-en	24.6	62.1	50.4
fr-en	31.9	53.1	55.8
it-en	29.2	57.1	53.7
nl-en	25.7	62.7	50.4
pt-en	31.8	53.7	56.0
sv-en	**32.7**	**52.3**	**56.6**

Results of ten bilingual phrase based decoders into English. All systems were trained on Europarl v3. Test set is Europarl test05. Best results are bold.

Oracle BLEU scores

	da	de	el	es	fi	fr	it	nl	pt	sv
da	—	3.2	**3.7**	2.4	1.9	2.6	**4.0**	2.4	2.4	1.7
de		—	2.7	2.0	1.9	2.0	3.3	2.7	2.1	1.6
el			—	2.1	1.8	2.3	**3.7**	2.6	2.5	2.5
es				—	1.2	3.1	2.4	1.7	3.1	**3.7**
fi					—	1.0	1.9	2.7	1.1	0.6
fr						—	2.4	1.6	3.5	**3.7**
it							—	2.4	2.5	2.7
nl								—	1.8	1.3
pt									—	3.5
sv										—

Absolute change in BLEU after combining two languages using oracle compared with the best BLEU of either language individually.

Oracle BLEU scores

- Oracle improvements ranged from 0.6 to 4.0 BLEU for two languages
- Much greater gains are seen when combining more languages

languages	BLEU	TER	METEOR
oracle-all	40.8	40.5	62.5

Combining ten systems results in **8.0 BLEU** improvement over best bilingual system.

Oracle BLEU scores

- Oracle improvements ranged from 0.6 to 4.0 BLEU for two languages
- Much greater gains are seen when combining more languages

languages	BLEU	TER	METEOR
oracle-all	40.8	40.5	62.5

Combining ten systems results in **8.0 BLEU** improvement over best bilingual system.

Oracle All Systems

system	% selected
da-en	14.1
de-en	9.6
el-en	10.3
es-en	14.0
fi-en	4.0
fr-en	12.9
it-en	7.2
nl-en	5.5
pt-en	9.8
sv-en	12.9

Percentage of time that sentences from each system were selected in an All-English oracle WER experiment. Score for overall oracle output was 43.8 WER and 40.8 BLEU.

Revisiting Max in Och and Ney (2001)

Max

$\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$

- Positive results reported combining up to 3 languages
- All reported combinations using Max had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Conducted new experiment using larger Europarl corpus.

- Experiment using Europarl (10 source languages)
- Include longer sentences (average 29 words)

Revisiting Max in Och and Ney (2001)

Max
$\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$

- Positive results reported combining up to 3 languages
- All reported combinations using Max had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Conducted new experiment using larger Europarl corpus.

- Experiment using Europarl (10 source languages)
- Include longer sentences (average 29 words)

Revisiting MAX in Och and Ney (2001)

MAX

$\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$

- Positive results reported combining up to 3 languages
- All reported combinations using MAX had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Conducted new experiment using larger Europarl corpus.

- Experiment using Europarl (10 source languages)
- Include longer sentences (average 29 words)

Revisiting Max in Och and Ney (2001)

Max

$\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$

- Positive results reported combining up to 3 languages
- All reported combinations using Max had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Conducted new experiment using larger Europarl corpus.

- Experiment using Europarl (10 source languages)
- Include longer sentences (average 29 words)

Revisiting MAX in Och and Ney (2001)

MAX

$\hat{e} = \arg\max_e \{p(e) \cdot \max_n p(f_n|e)\}$

- Positive results reported combining up to 3 languages
- All reported combinations using MAX had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Conducted new experiment using larger Europarl corpus.

- Experiment using Europarl (10 source languages)
- Include longer sentences (average 29 words)

Experiment — MAX

	da	de	el	es	fi	fr	it	nl	pt	sv
da	—	**0.4**	**0.1**	-0.8	-1.3	-1.3	**0.3**	-1.4	-0.7	-1.6
de		—	-0.2	-0.6	-0.8	-2.0	-0.1	-0.8	-0.8	-1.1
el			—	-0.2	-1.8	-1.0	**0.6**	-1.9	-0.3	-0.5
es				—	-1.5	**0.5**	-0.9	-2.6	**0.1**	**0.3**
fi					—	-2.9	-1.3	-0.3	-1.9	-2.3
fr						—	-1.6	-3.7	**0.2**	**0.2**
it							—	-1.5	-1.0	-1.0
nl								—	-2.4	-2.9
pt									—	-0.1
sv										—

Absolute change in BLEU after combining two languages using MAX ranking method compared with the best BLEU of either language individually. Only 20% of MAX pairwise combinations led to an improvement in BLEU.

Revisiting PROD in Och and Ney (2001)

PROD

$\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$

- Positive results reported combining up to 6 languages
- All but 2 reported combinations using PROD had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Attempted new experiment using larger Europarl corpus.

Revisiting PROD in Och and Ney (2001)

PROD
$\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$

- Positive results reported combining up to 6 languages
- All but 2 reported combinations using PROD had positive results
- No results reported for German-English or Finnish-English
- Test sentences were short (10-14 words)

Attempted new experiment using larger Europarl corpus.

Constraint Decoding

PROD requires each system to calculate a translation model probability for the output hypotheses of every system.

- n systems each produce one target hypothesis
- $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
- Each system must calculate $p(f_n|e)$ for all n target hypotheses.

	da-en	de-en	es-en	fr-en
% reachable	10.5	9.8	11.5	10.6

Percentage of sentences reachable by the Swedish-English system when constrained by the output of the listed systems.

Constraint Decoding

PROD requires each system to calculate a translation model probability for the output hypotheses of every system.

- n systems each produce one target hypothesis
- $\hat{e} = \arg\max_e \{p(e) \cdot \prod_{n=1}^{N} p(f_n|e)\}$
- Each system must calculate $p(f_n|e)$ for all n target hypotheses.

	da-en	de-en	es-en	fr-en
% reachable	10.5	9.8	11.5	10.6

Percentage of sentences reachable by the Swedish-English system when constrained by the output of the listed systems.

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Conclusions

- Hypothesis ranking has the potential to produce large improvements in translation quality
- MAX does not consistently produce positive results
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - New results show only 20% of MAX pairwise combinations led to an improvement in BLEU.
- Unable to replicate PROD
  - Och and Ney (2001) reported consistent positive results for up to 3 source languages
  - Vast majority of hypothesis unreachable during constraint decoding

Future Work

- Approximate translation model probabilities for PROD
- Incorporate system weighting
- Multilingual consensus decoding
- Multilingual lattice inputs
- Multi-synchronous decoding algorithms

Future Work

- Approximate translation model probabilities for PROD
- Incorporate system weighting
- Multilingual consensus decoding
- Multilingual lattice inputs
- Multi-synchronous decoding algorithms

Future Work

- Approximate translation model probabilities for PROD
- Incorporate system weighting
- Multilingual consensus decoding
- Multilingual lattice inputs
- Multi-synchronous decoding algorithms

Future Work

- Approximate translation model probabilities for PROD
- Incorporate system weighting
- Multilingual consensus decoding
- Multilingual lattice inputs
- Multi-synchronous decoding algorithms

Future Work

- Approximate translation model probabilities for PROD
- Incorporate system weighting
- Multilingual consensus decoding
- Multilingual lattice inputs
- Multi-synchronous decoding algorithms

Thank you!

Thank you!

Mahalo!