Context and Machine Translation

Aaron B. Phillips

Advanced MT Seminar

27 February 2008

con·text (n)

1. The part of a text or statement that surrounds a particular word or passage and determines its meaning.

2. The circumstances in which an event occurs; a setting.

In the beginning…

- Machine translation used to be performed word by word
- Eventually phrases were incorporated into the translation process
  - Encapsulates some *internal* local context

Phrase Tables

- Most phrase tables discard all *external* context of a translation
  - Partially a result of using Maximum Likelihood Estimation (MLE)
- Context is only handled by matching longer phrases
  - Unfortunately, the ‘correct’ long phrases frequently do not exist in our training data

Phrase Tables

- Context information usually not present in phrase tables:
  - POS, Lemma, or Chunk tags
  - Nearby words
  - Sentence information
  - Document information
  - Genre information

Three Different Approaches

- Modify the translation model (either ‘hard’ changes or incorporating new features)
  - Context as word sense disambiguation
- Make translation model specific to the task (adaptation)
  - Context as information retrieval
- A fluent target will result in proper selection and disambiguation
  - Context as language modeling

Context as Word Sense Disambiguation

Context is not helpful :(

Marine Carpuat and Dekai Wu. “Word sense disambiguation vs. statistical machine translation.” In Proceedings of the 43rd Annual Meeting on Association For Computational Linguistics. Ann Arbor, Michigan, June 2005.

- People assume context is helpful, is it really?
- SMT does some disambiguation based on local context, but using a SMT system for a WSD task resulted in significantly lower performance
  - We can hope for improvement from stronger WSD in SMT
- Yet, in real-world evaluation it doesn’t help

Context is not helpful :(

- Used WSD module built from Senseval-3 data (only 20 Chinese words)
- Evaluation on selection of MT04 (Chinese-English) with the known ambiguous Chinese words
  - Baseline SMT system: 0.1310 BLEU
  - Constrained phrase table based on WSD module output: 0.1239 BLEU
  - Post-processed translations: 0.1253 BLEU
    - Replaced ambiguous target word with highest scoring WSD module prediction to over-come language model effects
  - Extended WSD modules’ possible translations with SMT dictionary: 0.1232 BLEU
    - WSD module’s proposed translations were originally glosses in HowNet dictionary

Context is not helpful :(

- Evaluation showed that for all but two of the target words, the WSD module either hurt the BLEU score or did not help it
  - WSD constraints hurt neighboring words due to the n-gram language model
    - For example, “impact” is a better translation than “shock” for a particular Chinese word, but the SMT model did not know how to use “impact” better in a sentence
    - Thus, the SMT model cannot make better use of the WSD predictions
  - Post-processing avoids this, but then the output is not always as fluent
  - Even BLEU-1 had lower scores when using WSD, so it is not just a problem of BLEU favoring long n-grams

Phrase Translation as Classification

David Vickrey et al. “Word-sense disambiguation for machine translation.” In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, October 2005.

- Preliminary work in *successfully* integrating WSD and SMT
- Apply standard WSD techniques to build classifiers from source words to target words
  - Each unique target word represents a different ‘sense’ of the source word

Phrase Translation as Classification

- Learn a logistic regression model using all examples in the corpus with simple features
  - POS tag for the source word
  - A binary ‘occurs’ variable for each word within a specific range of the source word
- Single-word accuracy
  - Baseline (most frequent): 0.526
  - Simple logistic model: 0.605

Phrase Translation as Classification

- Gap-filling Experiment
  - Leave target sentence in-tact except for a single word that needs to be translated
  - Combine scores from translations to fill the gap along with a target language model (SMT-like architecture)
    - Should perform similarly to an SMT system, while holding the outside context of the translation constant
  - Accuracy
    - Baseline: 0.833
      - Language model + P(s|t) and P(t|s) from alignments
    - Baseline + WSD: 0.846

- Results are an under-estimate because multiple possible translations are usually valid and this only considered the one target word (present in the corpus) as correct

Discriminative Phrase Translation

Jesus Gimenez and Lluis Marques. “Context-aware Discriminative Phrase Selection for Statistical Machine Translation.” In Proceedings of the Second Workshop on Statistical Machine Translation. Prague, Czech Republic, June 2007.

- Extends previous work to handle *phrasal* translation
- Moves from *blank-filling* task to *full translation* task

Discriminative Phrase Translation

- Every phrase pair (f, e) is transformed into a multi-class classification problem where every possible e is a class
- Train a set of local linear SVMs (one-vs-all classification) with positive and negative examples

Discriminative Phrase Translation

- SVM features are extracted from each source sentence
  - Employs WSD methods from (Yarowsky et al. 2001)
  - Local context: Within a window of five tokens to the left and right of the phrase calculate n-grams (1,2,3) of
    - Words
    - Part-of-Speech
    - Lemmas
    - Base-Phrase chunk labels
  - Global context: Topical information collected by treating source sentence as a bag of lemmas

Discriminative Phrase Translation

- Convert SVM score into P(e|f) using softmax function (Bishop 1995)
- Generate phrase table such that each distinct occurrence of a source phrase has a separate list of possible phrase translation candidates with their corresponding scores
  - Input document is transformed into sequence of identifiers
- This strategy is compatible with a standard decoder, but it cannot model features related to the target sentence under consideration or alignment information

Discriminative Phrase Translation

- Used DPT scoring for 41 frequent phrases, the remainder scored using MLE
  - DPT yields higher accuracy when there exists a sufficient number of examples of the phrase pair (over 10,000).

	BLEU	METEOR	ROUGE
MLE(f|e)	0.59	0.77	0.42
MLE(e|f)	0.62	0.77	0.43
DPT(e|f)	0.62	0.78	0.44

*Monotone decoding

Discriminative Phrase Translation

- In human evaluation DPT improves adequacy, but not fluency
  - Further investigation of the integration of DPT probabilities into the statistical framework is warranted

	Adequacy	Fluency	Overall
MLE > DPT	39	**84**	83
MLE = DPT	100	76	46
MLE < DPT	**89**	68	**99**

WSD in Hiero

Yee Seng Chan, Hwee Tou Ng, and David Chiang. "Word Sense Disambiguation Improves Statistical Machine Translation" In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Prague, Czech Republic, 2007.

- Added two features to every grammar rule (dynamically during decoding)
  - Contextual probability of WSD choosing t as a translation for s, P(t|s)
    - This does not apply to all rules
  - A negative weight that rewards rules that use translations suggested by the WSD module

WSD in Hiero

- Evaluation on MT03 (Chinese-English)
  - Heiro: 29.73 BLEU
    - Much stronger baseline than found in other work
  - Hiero + WSD: 30.30 BLEU

Carpuat Revisited

Marine Carpuat and Dekai Wu. “Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation.” Machine Translation Summit XI. Copenhagen, September 2007.

Marine Carpuat and Dekai Wu. “How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation.” 11th International Conference on Theoretical and Methodological Issues in Machine Translation. Skovde, September 2007.

Phrase Sense Disambiguation

- Use features found to work well in Senseval WSD tasks
  - Bag-of-words context
  - Local collocations
  - Position sensitive local POS tags
  - Basic dependency features

Phrase Sense Disambiguation

- Use state-of-the-art WSD model to provide a context-dependent probability distribution over the possible translation candidates for a given phrasal lexicon entry
  - Best performing WSD model in Senseval-3 (Carpuat et al, 2004)
- Adds new features to the phrase table

Phrase Sense Disambiguation

- Model disambiguates phrases, not just words as in Senseval tasks
- The sense candidates are defined by the entries in the baseline phrase table (learned from data)
- Training is performed using the original corpus for every example with a consistent phrasal alignment

Phrase Sense Disambiguation

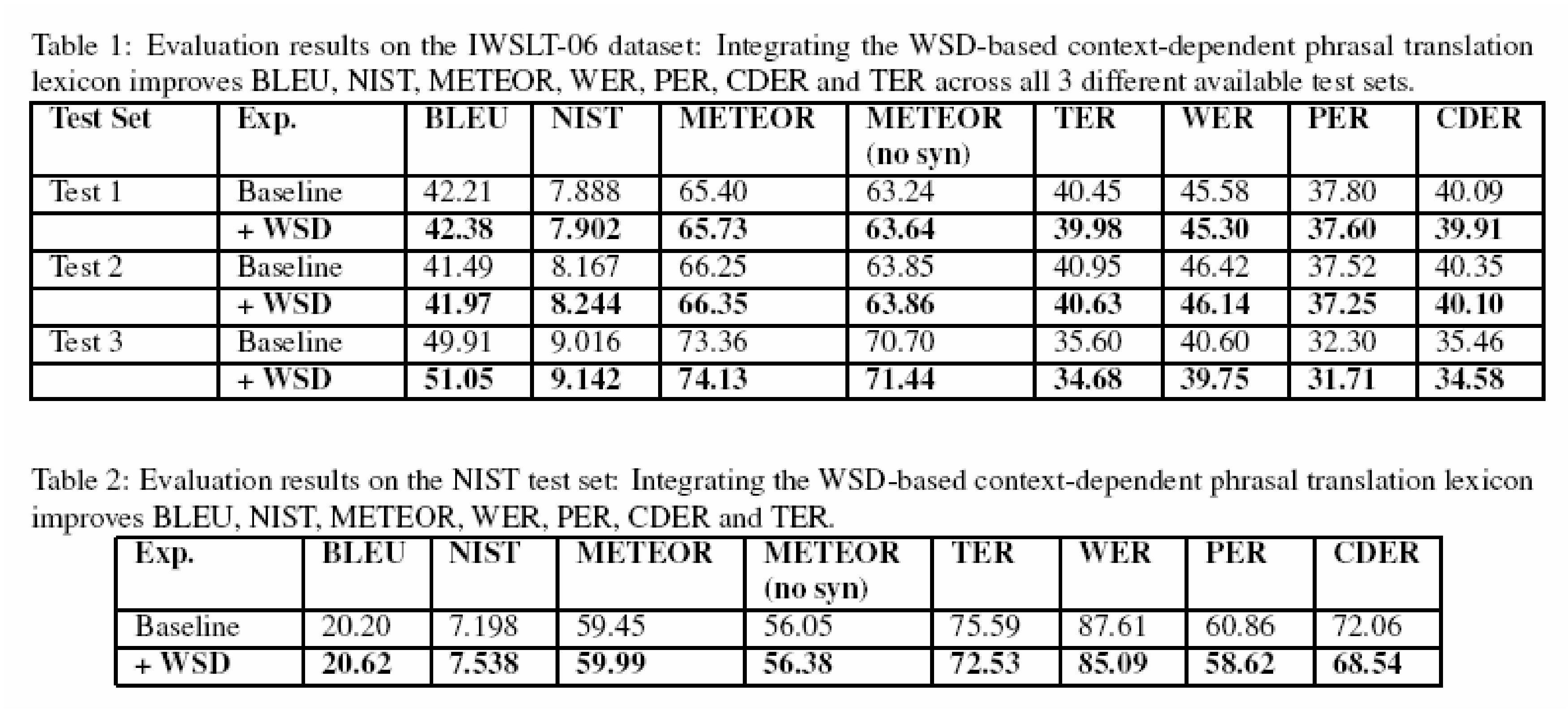


Table 1: Evaluation results on the IWSLT-06 dataset: Integrating the WSD-based context-dependent phrasal translation lexicon improves BLEU, NIST, METEOR, WER, PER, CDER and TER across all 3 different available test sets.

Test Set	Exp.	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
Test 1	Baseline	42.21	7.888	65.40	63.24	40.45	45.58	37.80	40.09
	+ WSD	**42.38**	**7.902**	**65.73**	**63.64**	**39.98**	**45.30**	**37.60**	**39.91**
Test 2	Baseline	41.49	8.167	66.25	63.85	40.95	46.42	37.52	40.35
	+ WSD	**41.97**	**8.244**	**66.35**	**63.86**	**40.63**	**46.14**	**37.25**	**40.10**
Test 3	Baseline	49.91	9.016	73.36	70.70	35.60	40.60	32.30	35.46
	+ WSD	**51.05**	**9.142**	**74.13**	**71.44**	**34.68**	**39.75**	**31.71**	**34.58**

Table 2: Evaluation results on the NIST test set: Integrating the WSD-based context-dependent phrasal translation lexicon improves BLEU, NIST, METEOR, WER, PER, CDER and TER.

Exp.	BLEU	NIST	METEOR	METEOR (no syn)	TER	WER	PER	CDER
Baseline	20.20	7.198	59.45	56.05	75.59	87.61	60.86	72.06
+ WSD	**20.62**	**7.538**	**59.99**	**56.38**	**72.53**	**85.09**	**58.62**	**68.54**

Phrase Sense Disambiguation

- Most useful features
  - POS tag preceding phrase and POS tag following phrase
  - Bag-of-words full sentence context

Phrase Sense Disambiguation

- Limiting WSD predictions to single words does NOT reliably improve translation quality

Summary: Context as Word Sense Disambiguation

- Context is beneficial when *properly* integrated into the translation model
  - Hard filtering of the phrase table
    - BAD
  - Use WSD to generate P(e|f)
    - SO-SO
  - Use WSD as a feature in phrase table
    - GOOD

Context as Information Retrieval

Context as Information Retrieval

- Context as Word Sense Disambiguation
  - Local changes
- Context as information retrieval
  - Global changes

Filtering Training Data

Almut Silja Hildebrand, et al. “Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval”, Proceedings of EAMT 2005. Budapest, Hungary, May 2005.

- Select n most similar sentences from the training data
- Build an adapted translation model using the filtered training data
- Translate document with the adapted translation model

Filtering Training Data

- Use TF-IDF as similarity metric between input sentences and training documents
- Instead of training separate models, train one model on the set of all similar sentences
  - A translation model trained from only a few hundred sentences is unlikely to give robust probabilities
  - Unlikely that a test document changes genre very quickly

Filtering Training Data

- Results (Chinese-English Tourism Dialogues)
  - ID: 0.4621 BLEU
  - ID + 15k Random OOD: 0.4850 BLEU
  - ID + 75k Random OOD: 0.4501 BLEU
  - Best-case Adapted: 0.4924 BLEU

Mixture Modeling

George Foster and Roland Kuhn. "Mixture-model adaptation for SMT." In Proceedings of the Second Workshop on Statistical Machine Translation. Prague, Czech Republic, June 2007.

- Split the training data into multiple corpora
  - Currently use genre information
  - Could perform automatic topic identification and clustering
- Train a separate models for each corpus
- Weight models and combine into a single global model

Mixture Modeling

- “Static” Context Matching
  - Appropriate if the development set is very similar to the test set
  - Weights can be set by tuning

Mixture Modeling

- "Dynamic" Context Matching
  - Set mixture weights according to each corpora's fit with the test data
  - Define a distance metric as a function of these features
    - TF-IDF
    - Latent Semantic Analysis
    - Perplexity
    - Weights for each corpus that maximize P(t|corpus) when treated as a mixture model

Mixture Modeling

- "Static" Context Matching
  - Translation model adaptation is effective (~1 BLEU point)
  - Language model adaptation works better, and essential covers everything from translation model adaptation
- "Dynamic" Context Matching
  - Translation model adaptation does not show significant improvement over the baseline
  - Language model adaptation does continue to improve over the baseline by ~1 BLEU point

Online Adaptation

Ralf D. Brown. “Context-Sensitive Retrieval for Example-Based Translation.” In Proceedings of the Tenth Machine Translation Summit. Phuket, September 2005.

- Give a bonus to examples with contextual matches

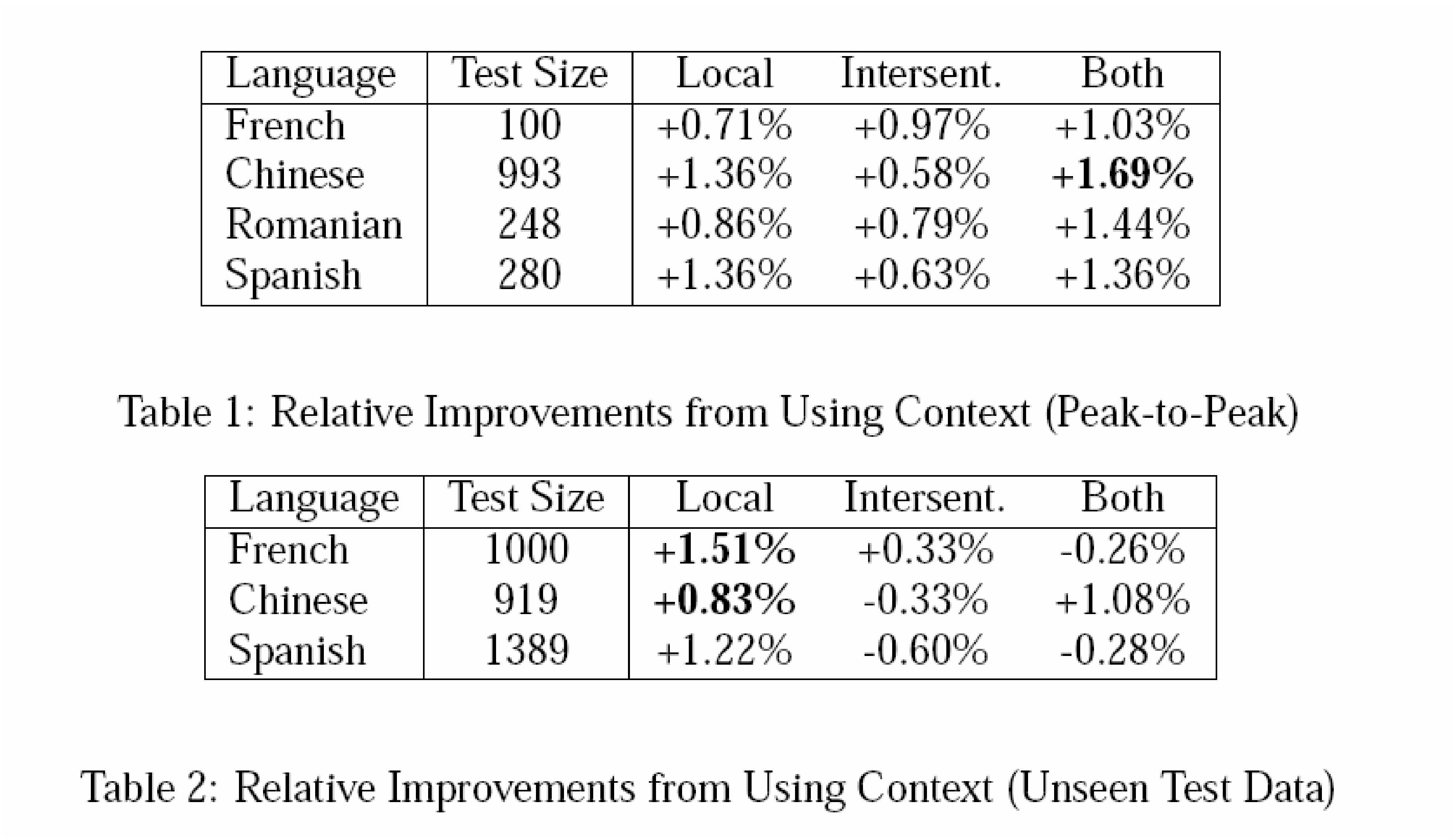
Online Adaptation

- Local Context
  - If a training sentence is largely the same as the sentence to be translated, multiple examples will be retrieved from the training sentence (1-gram, 2-gram, 3-gram, etc.)
  - For each input sentence, give a bonus when multiple examples are retrieved from the same training sentence

Online Adaptation

- Inter-Sentential Context
  - Record usage counts of each example across all sentences
  - Give a bonus to examples that have been used frequently or are *near* examples that are used frequently

Online Adaptation



Language	Test Size	Local	Intersent.	Both
French	100	+0.71%	+0.97%	+1.03%
Chinese	993	+1.36%	+0.58%	**+1.69%**
Romanian	248	+0.86%	+0.79%	+1.44%
Spanish	280	+1.36%	+0.63%	+1.36%

Table 1: Relative Improvements from Using Context (Peak-to-Peak)

Language	Test Size	Local	Intersent.	Both
French	1000	**+1.51%**	+0.33%	-0.26%
Chinese	919	**+0.83%**	-0.33%	+1.08%
Spanish	1389	+1.22%	-0.60%	-0.28%

Table 2: Relative Improvements from Using Context (Unseen Test Data)

Summary: Context as Information Retrieval

- Filters or applies weights to training data *before* building a phrase table
- *Globally* skews the translation candidates and their probabilities
- Potential for improvement, but no clear winning technique

Context as Language Modeling

Context as Language Modeling

- What we already do…
- Meaningful Machines CBMT

Context as Language Modeling

- Context as Language Modeling
  - Does not model any correspondence between the source and target
  - Only models the fluency of the target

To the Future and Beyond…

- Statistically significant, but not "significant" increases in evaluation metrics
- Partially a problem with the sensitivity of evaluation metrics

To the Future and Beyond…

- Rather than expecting dramatic improvements, contextual information should better inform our system regarding subtleties of the text
- Necessary for real human uses of machine translation
- Perhaps we aren’t ready for it yet!

To the Future and Beyond…

- Context as Word Sense Disambiguation
  - Appropriate for word or sentence level changes
- Context as information retrieval
  - Appropriate for genre or document-level changes

To the Future and Beyond…

- Context as Word Sense Disambiguation is more mature and better integrated than Context as Information Retrieval
- Neither method works real well with static phrase tables and push us toward generating dynamic, online, phrase tables

Source Channel Paradigm and Context

Maximize P(f|e) * P(e)

- P(f|e): Translation Model
  - Currently models are weak
- P(e): Language Model
  - Currently models are strong and this compensates for the weak translation model

Source Channel Paradigm and Context

- Small gains in P(f|e) won’t be as noticeable if we rely strongly on P(e) to perform discrimination

- P(e) is more heavily relied on, so adoption/changes to the language model will *currently* make a bigger difference

Source Channel Paradigm and Context

- All the context information is on the source side (f) so P(f|e) is much more difficult to model than P(e|f)

- Perhaps to make full use of contextual information we need a more direct translation approach

To the Future and Beyond…

- And maybe something else…