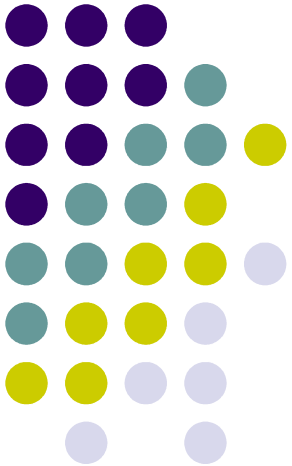
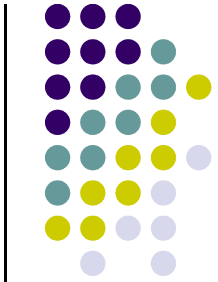
The University of Washington Machine Translation System for the IWSLT 2007 Competition

Katrin Kirchhoff & Mei Yang

Department of Electrical Engineering
University of Washington, Seattle, USA



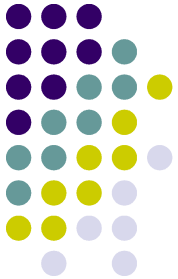
Overview



- Two systems, I-EN and AR-EN
- Main focus on exploring use of out-of-domain data and Arabic word segmentation

- Basic MT System
- Italian-English system
  - Out-of-domain data experiments
- Arabic-English system
  - Semi-supervised Arabic segmentation

Basic MT System



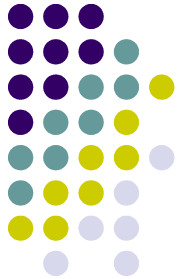
- Phrase-based SMT system
- Translation model: log-linear model

$$e^* = \arg\max_e p(e|f) = \arg\max_e \{\sum_{k=1}^{K} \lambda_k \phi_k(e,f)\}$$

- Feature functions:

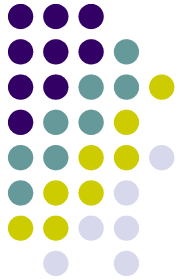
- **2 phrasal translation probabilities**
- **2 lexical translation scores**
- **Word count penalty**
- **Phrase count penalty**
- **LM score**
- **Distortion score**
- **Data source feature**

Basic MT System



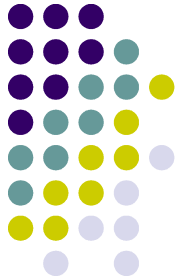
- Word alignment
  - HMM-based word-to-phrase alignment [Deng & Byrne 2005] as implemented in MTTK
  - Comparable in performance to GIZA++
- Phrase extraction:
  - Method by [Och & Ney 2003]

Basic MT System



- Decoding/Rescoring:
  - Minimum-error rate training to optimize weights for first pass (optimization criterion: BLEU)
  - MOSES decoder
  - First pass: up to 2000 hypotheses/sentence
  - additional features for rescoring
    - POS-based language model
    - rank-based feature [Kirchhoff et al, IWSLT06]
  - → Final 1-best hypothesis

Basic MT System



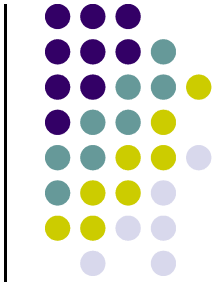
- Postprocessing:
  - Hidden-ngram model [Stolcke & Shriberg 1996] for punctuation restoration:

$$P(e_1,\ldots,e_T) \approx \prod_{t=1}^{T} P(e_t \mid e_{t-1},\ldots,e_{t-n+1})$$

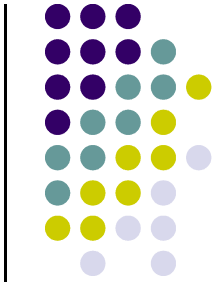
Event set E: words and punctuation signs

  - Noisy-channel model for truecasing
    - SRILM *disambig* tool

Basic MT System

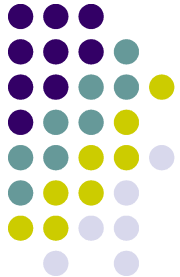


- Spoken-language input:
  - IWSLT 06: mixed results using confusion network input for spoken language
  - No specific provisions for ASR output this year!



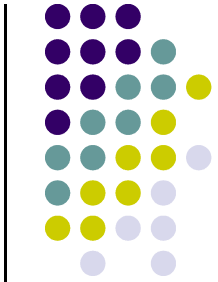
Italian-English System

Data Resources



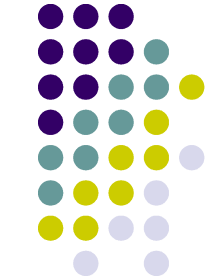
- Training:
  - BTEC (all data sets from previous years, 160k words)
  - Europarl Corpus (parliamentary proceedings, 17M words)
- Development:
  - IWSLT 2007 (SITAL) dev set (development only, no training on this set)
  - Split into dev set (500 sentences) and held-out set (496 sentences)
- BTEC name list

Preprocessing



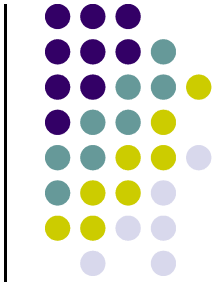
- Sentence segmentation into smaller chunks based on punctuation marks
- Punctuation removal and lowercasing
- Automatic re-tokenization on English and Italian side:
  - For N most frequent many-to-one alignments (N = 20), merge multiply aligned words into one
    - E.g. *per piacere – please* → *per_piacere*
  - Improves noisy word alignments

Translation of names/numbers



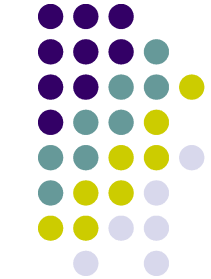
- Names in the BTEC name list were translated according to list, not statistically
- Small set of hand-coded rules to translate dates and times

Out-of-vocabulary words



- To translate OOV words, map OOV forms to all words in training data that do not differ in length by more than 2 characters
- Mapping done by string alignment
- For all training words with edit distance < 2, reduplicate phrase table entries with word replaced by OOV form
- Best-matching entry chosen during decoding
- Captures misspelled or spoken-language specific forms: *senz', undic', quant', etc.*

Using out-of-domain data



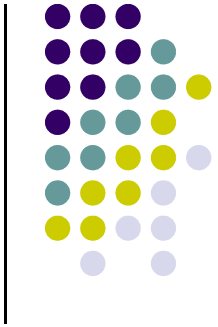
- Experience in IWSLT 2006:
  - additional English data for language model did not help
  - out-of-domain data for translation model (IE) was useful for text but not for ASR output
- Adding data:
  - second phrase table trained from Europarl corpus
  - Use phrase tables from different sources jointly
  - each entry has *data source feature*: binary feature indicating corpus it came from

Using out-of-domain data

- Phrase coverage (%) and translation results on IE 2007 dev set

	1	2	3	4	5	BLEU/PER
BTEC	78.3	29.6	6.7	1.3	0.2	18.9/55.1
Europarl	83.9	37.0	6.4	0.7	0.1	18.5/55.4
combined	85.8	39.9	9.4	1.7	0.2	20.7/53.5

Using out-of-domain data



Diagnostic experiments: importance of matched data vs. data size

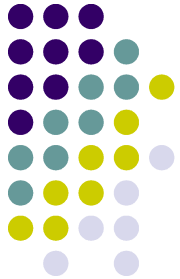
Training set	
500 sentences from SITAL	Small amount of matched data
BTEC training corpus	Moderate amount of domain-related by stylistically different data
Europarl	Large amount of mismatched data
BTEC & Europarl	Combination of sources
BTEC & Europarl & SITAL	

Using out-of-domain data

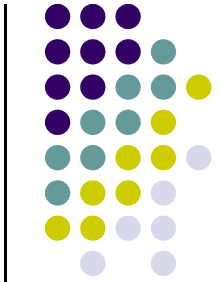
Performance on held-out SITAL data (%BLEU/PER)

Training set	Text	ASR output
500 sentences from SITAL	28.0/46.8	26.1/49.0
BTEC training corpus	18.9/55.1	16.6/57.1
Europarl	18.5/55.4	17.3/56.4
BTEC & Europarl	20.7/53.3	18.6/55.3
BTEC & Europarl & SITAL	30.1/41.9	27.7/44.8

Effect of different techniques

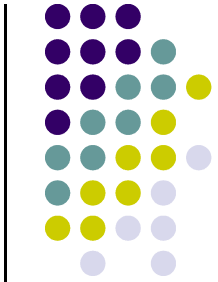


Technique	BLEU(%)/PER
Baseline	18.9/55.1
OOD data	20.7/53.4
Rescoring	22.0/52.7
Rule-based number trans.	22.6/52.2
Postprocessing	21.2/50.4



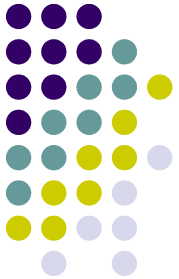
Arabic-English System

Data Resources



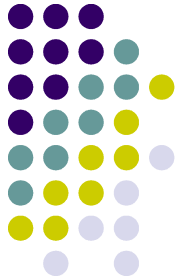
- Training
  - BTEC data, without dev4 and dev5 sets (160k words)
  - LDC parallel text (Newswire, MTA, ISI) (5.5M words)
- Development
  - BTEC dev4 + dev5 sets
- Buckwalter stemmer

Preprocessing



- Chunking based on punctuation marks
- Conversion to Buckwalter format
- Tokenization
  - Linguistic tokenization
  - Semi-supervised tokenization

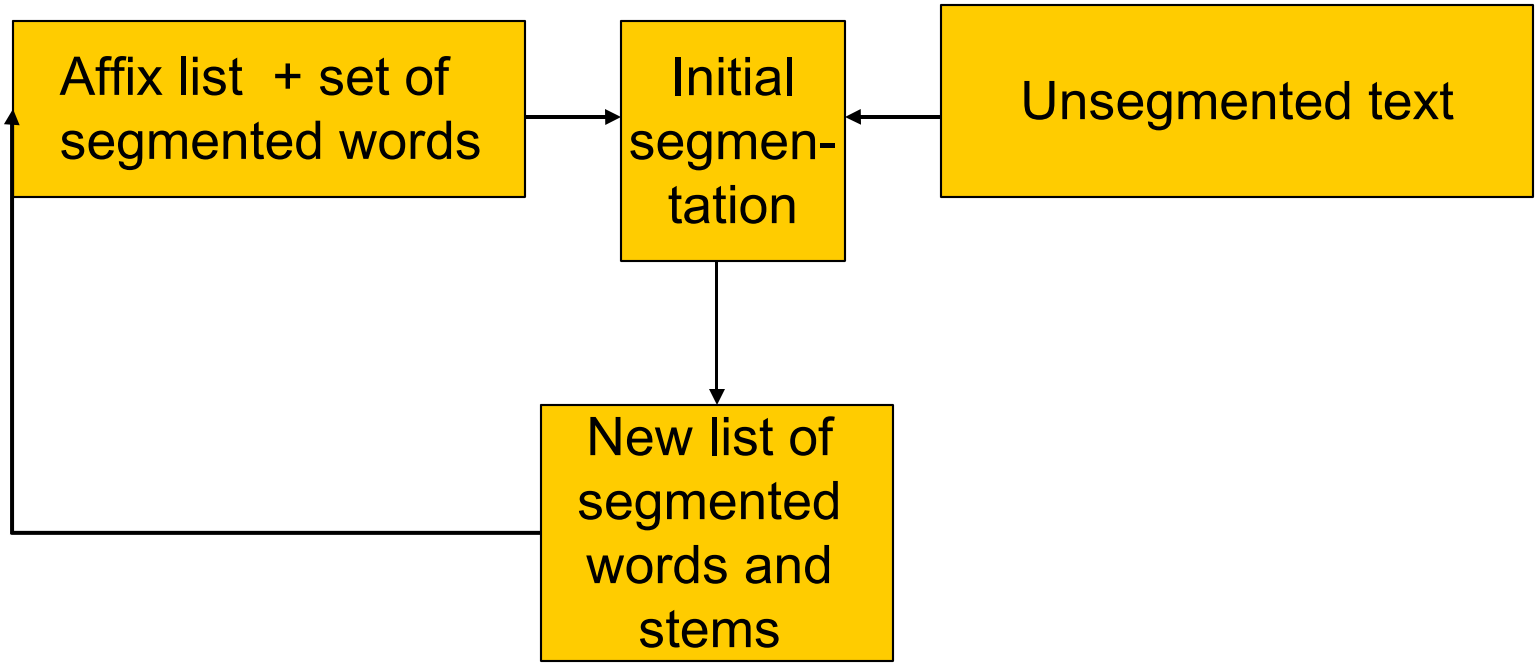
Linguistic Tokenization



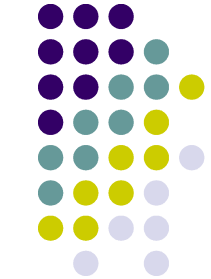
- Columbia University MADA/TOKAN tools:
  - Buckwalter stemmer to suggest different morphological analyses for each word
  - Statistical disambiguation based on context
  - Splitting off of word-initial particles and definite article
- - Involves much human labour
- - difficult to port to new dialects/languages

Semi-supervised tokenization

- Based on [Yang et al. 2007]: segmentation for dialectal Arabic



Semi-supervised tokenization



- Needs fewer initial resources
- Produces potentially more consistent segmentations
- Once trained, much faster than linguistic tools

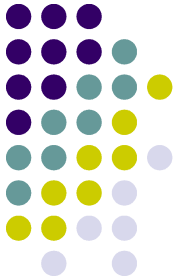
Method	BLEU/PER
MADA/TOKAN	22.5/50.5
SemiSup – initialized on MSA	23.0/50.7
SemiSup – initialized on Iraqi Arabic	21.6/51.1

Effect of out-of-domain data

Phrase coverage rate (%) and translation performance on dev5 set, clean text

	1	2	3	4	5	6	BLEU/PER
BTEC	**69.3**	**34.3**	**12.9**	**3.6**	**1.2**	**0.3**	**22.5/50.5**
News	**60.2**	**30.4**	**11.5**	**2.7**	**0.9**	**0.2**	**----**
combined	**82.6**	**46.7**	**20.1**	**5.6**	**1.9**	**0.5**	**24.6/47.5**

Effect of different techniques



Technique	BLEU(%)/PER
Baseline	22.5/50.5
OOD data	24.6/48.2
Rescoring	24.6/47.5
Postprocessing	23.4/48.5

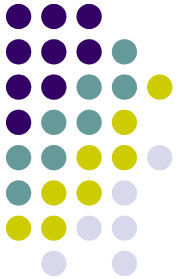
Dev5 set, clean text

Evaluation results

Performance on eval 2007 sets

	BLEU(%)
IE – clean text	26.5
IE – ASR output	25.4
AE – clean text	41.6
AE – ASR output	40.9

Conclusions



- Adding out-of-domain data helped in both clean text and ASR conditions
- Importance of stylistically matched data
- Semi-supervised word segmentation for Arabic comparable to supervised segmentation, uses fewer resources
- Cross-dialectal bootstrapping of segmenter possible