HKUST System Description
Toward Integrating Word Sense and Entity Disambiguation into SMT

Marine CARPUAT Yihai SHEN Xiaofeng YU Dekai WU

HKUST
Human Language Technology Center
Department of Computer Science
University of Science and Technology
Hong Kong

{marine, shenyh, xfyu, dekai}@cs.ust.hk

The HKUST submission

Goals for our first-time IWSLT participation:

- Test our current system, designed primarily for Chinese-English text translation
  - on various data sets and input conditions
    - Chinese-English text, read speech, spontaneous speech
  - on various language pairs from different language families
    - Arabic-English, Chinese-English, Italian-English, Japanese-English
- Investigate integration of semantic processing

Outline

- System description
  - Core phrase-based SMT engine
  - Word Sense Disambiguation for lexical choice
  - Named-Entity translation
- IWSLT results
  - Chinese-English
  - Other language pairs

System description
Core MT engine uses phrase-based SMT

- Baseline is a phrase-based log-linear model
- Phrasal bilexicon
  - learned from intersection of IBM4 alignments
  - Following Koehn [2003], base features are:
    - conditional translation probabilities in both directions
    - lexical weights derived from word translation probabilities
- Decoder
  - Pharaoh [Koehn 2004]
- Language model
  - standard 3-gram model trained using SRI LM toolkit [Stolcke 2002]

Integration of semantic processing
1) Word Sense Disambiguation for lexical choice

- Phrase-based SMT makes little use of context information
- In contrast, WSD approaches generalize across rich contextual features to choose a word sense

- Previous work:
  - Senseval WSD models do not help translation quality when integrated into a word-based SMT model [Carpuat & Wu 2005]

- In this new version, we repurpose the WSD models for SMT:
  - WSD "senses" are exactly same as SMT translation candidates
  - WSD training data is exactly same as SMT training data
  - WSD scores are added to log linear model feature set

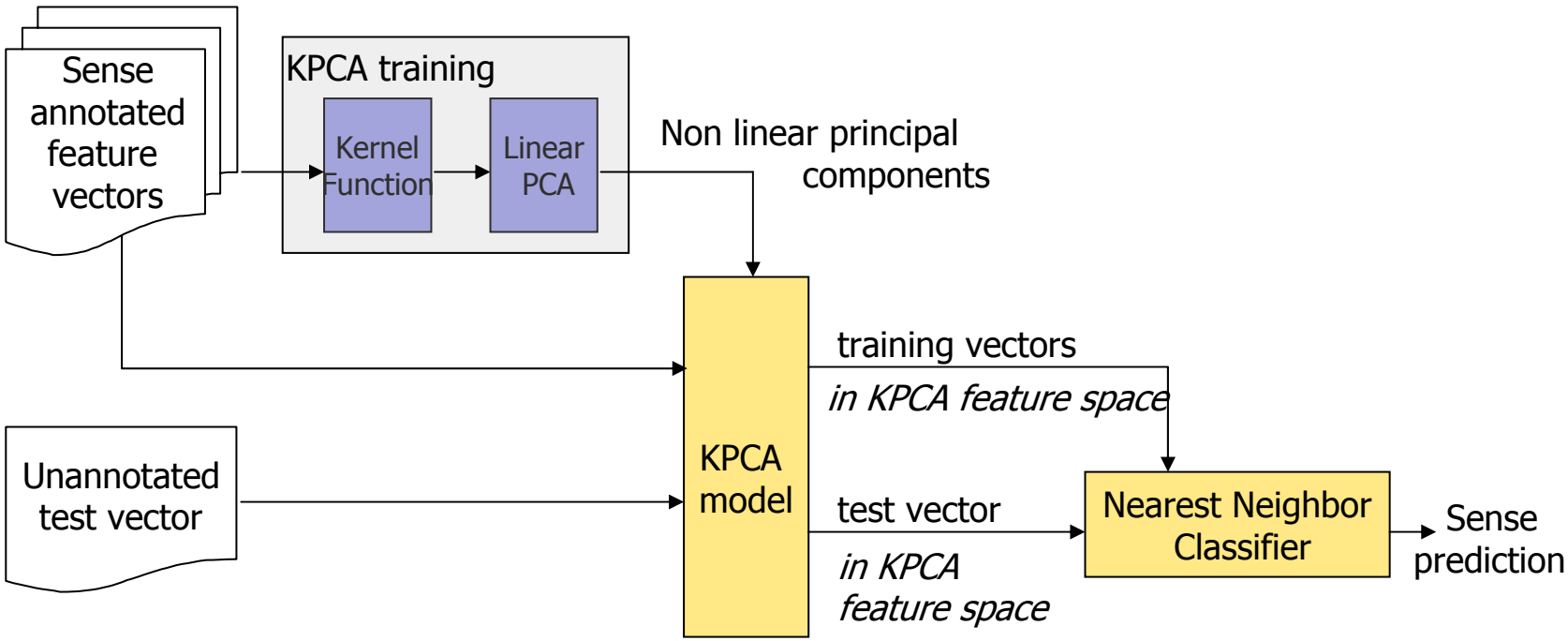
The HKUST WSD System

- Proved highly effective at Senseval-3
  - Placed first on Chinese lexical sample
  - Placed second on Multilingual lexical sample (translation)
  - 71.4% on English lexical sample (median 67.2, best 72.9)
- Classifier ensemble:
  - naïve Bayes [Yarowsky & Florian 2002]
  - maximum entropy [Klein & Manning 2002]
  - boosting [Carreras *et al.* 2002; Wu *et al.* 2002]: we use boosted decision stumps
  - Kernel PCA model [Wu *et al.* 2004]

The HKUST WSD System

KPCA model is a good candidate to augment ensemble

- We introduced the new Kernel Principal Component Analysis (KPCA) model for WSD in Wu *et al.* [2004]



Overview of our KPCA model for WSD

The HKUST WSD System

KPCA model is a good candidate to augment ensemble

- Is well suited to WSD:
  - can handle high dimensional problems
  - inherently takes into account combinations of predictive features
- Achieves high accuracy as a stand-alone model:
  - outperforms maximum entropy, naïve Bayes, and SVMs on Senseval-2 English lexical sample
- Has a different prediction bias than the 3 other voters

The HKUST WSD System
Contextual Features

- Feature set includes:
  - Bag-of-words context
  - Position sensitive local collocational features
  - Syntactic features
- A WSD model using these features yielded the best classification accuracy in Yarowsky & Florian [2002]

Preliminary result

Using WSD helps translation quality on **all** IWSLT dev test sets.

Evaluation results on IWSLT 2006 Chinese-English Test Sets

Test set	SMT system	BLEU	NIST
Dev. Test 1	Baseline	40.76	7.938
	+ WSD	**41.28**	**7.981**
Dev. Test 2	Baseline	39.81	8.153
	+ WSD	**39.85**	**8.175**
Dev. Test 3	Baseline	49.26	9.117
	+ WSD	**49.81**	**9.152**
Dev. Test 4	Baseline	16.13	5.725
	+ WSD	**16.27**	**5.757**

Integration of semantic processing
2) Named-Entity Translation

- NE translation is a particular case of WSD that requires specific handling
  - Names are rare and often never seen during training
  - Whether a phrase is a NE is context-dependent
  - Specific translation patterns

- Our approach
  - NE recognition: tag both NE boundaries and type
    - We distinguish Person, Organization, Location
  - Rule-based translation approach for each category using
    - Name-lists
    - Transliteration scheme
  - Integration:
    - NE translation candidates are added to phrasal bilexicon using Pharaoh XML markup scheme for input text

The HKUST NER system

- Extensively evaluated on European languages at CoNLL shared tasks [Wu et al. 2002, 2003]

- Classifier ensemble:
  - boosting [Carreras *et al.* 2002; Wu *et al.* 2002]: we use boosted decision stumps
  - Support Vector Machines [Boser *et al.* 1992]
  - Transformation-based learning [Brill 1995]

The HKUST NER system
Feature set

- **Lexical features** within a window of 2 words around the current word
  - Word
  - Lemma
- **Morphological features** for the current word
  - Prefixes and suffixes of up to 4 characters from the current word
  - Capitalization: is the first letter or the whole word capitalized?
- **Syntactic features**
  - POS tag within a window of 2 words around the current word
  - 2 previous NE tags history
- **Lexicosyntactic features**
  - Conjunctions of POS tags and words

Outline

- System description
  - Core phrase-based SMT engine
  - Word Sense Disambiguation for lexical choice
  - Named-Entity translation
- IWSLT results
  - Chinese-English
  - Other language pairs

IWSLT tasks

- Chinese-English speech translation
  - Correct Recognition vs. Read Speech vs. Spontaneous Speech
  - We do not perform ASR and simply used provided 1-best recognition results
- Text and read speech only
  - Arabic-English
  - Italian-English
  - Japanese-English

Language-specific data preprocessing was kept minimal

- **English** data was tokenized and case-normalized
- **Italian** data was processed as if it were English
- **Chinese** data was word segmented using LDC segmenter
- **Japanese** data was used directly as provided
- **Arabic**
  - Converted to Buckwalter romanization scheme
  - Tokenized with ASVMT Morphological Analysis toolkit [Diab 2005]

Chinese-English evaluation results

Metric	HKUST result **correct recog.**	result range **correct recog.**	HKUST result **read speech**	result range **read speech**	HKUST result **spont. speech**	result range **spont. speech**
BLEU	18.04	12.84–24.23	15.45	10.37–21.11	14.41	05.85–18.98
NIST	5.3615	4.0658–6.0961	4.7769	3.6384–5.4154	4.6365	3.5755–5.1513
METEOR	49.15	46.01–50.33	**44.56**	37.29–44.56	**42.38**	31.43–42.38

Chinese translations for different input conditions: **Example 1**

Input (text):	可以 请 把 你 在 日本 的 地址 写 下来 好吗
Output:	Could you please write down the address in Japan, please.
Input (read):	可以 请问 您 在 日本 的 地址 写 下来 好吗
Output:	Could you please write down the address in Japan, please.
Input (spontaneous):	可以 请 办理 给 人 的 地址 写 了 好吗
Output:	May handle, deal with the address of the please.

- Translation of both text and read speech are acceptable

Chinese translations for different input conditions: **Example 1**

Input (text):	可以 请 把 你 在 日本 的 地址 写 下来 好吗
Output:	Could you please write down the address in Japan, please.
Input (read):	可以 请问 您 在 日本 的 地址 写 下来 好吗
Output:	Could you please write down the address in Japan, please.
Input (spontaneous):	可以 请 办理 给 人 的 地址 写 了 好吗
Output:	May handle, deal with the address of the please.

- Translation of both text and read speech are acceptable
- But SMT can't recover ASR error in spontaneous speech transcription

Chinese translations for different input conditions: **Example 2**

Input (text):	你 晚上十点 以前 必须 登记
Output:	You must check in by ten o'clock in the evening.
Input (read):	您 玩儿十点 以前 必须 登记
Output:	You must check in at ten before.
Input (spontaneous):	您 晚上十点 以前 必须 登记
Output:	You must check in by ten o'clock in the evening.

- Some ASR transcription errors do not affect the translation

Chinese translations for different input conditions: **Example 2**

Input (text):	你 晚上 十点 以前 必须 登记
Output:	You must check in by ten o'clock in the evening.
Input (read):	您 玩儿 十点 以前 必须 登记
Output:	You must check in at ten before.
Input (spontaneous):	您 晚上 十点 以前 必须 登记
Output:	You must check in by ten o'clock in the evening.

- Some ASR transcription errors yield same translation
- Here, read speech transcription is worse than spontaneous speech transcription, which affects the translation

Additional language pair results

Evaluation Metric	Arabic (text)	Arabic (read)	Italian (text)	Italian (read)	Japanese (text)	Japanese (read)
BLEU	16.63	14.77	29.64	23.74	15.60	15.23
NIST	3.8863	3.3318	7.1816	6.0956	0.1560	0.1523
METEOR	42.88	39.20	62.39	54.03	45.79	42.83
WER	67.57	69.16	58.08	63.07	72.48	72.39
PER	56.47	59.48	43.40	49.38	57.86	58.18

Comparison of translation quality across languages: **Example 1**

	SMT Input	**SMT Output**
Reference		*It is about twenty kilometers away from here.*
Arabic	انها على بعد حوالي عشرين كيلو متر من هنا	On in about twenty kilometers from here.
Chinese	离 这儿 大概 二 十 公里	About twenty kilometers from here.
Italian	è distante circa venti chilometri da qui	It's about twenty kilometers far from here.
Japanese	ここ から 二 十 キロ ほど です	About two – kilometers from here.

Comparison of translation quality across languages: **Example 2**

	SMT Input	**SMT Output**
Reference		*This wine is from France. It's very famous.*
Arabic	هذا النبيذ من فرنسا وهو مشهور جدا	This wine from France and is very popular.
Chinese	这 是 法国产 的 葡萄酒 非常 有名	This is very famous French made wine.
Italian	questo vino viene dalla francia è molto famoso	This wine comes from France is very popular.
Japanese	こ れ は フ ラ ン ス ワ ィ ン で す と て も 有名 で す	This is 's very famous.

Comparison of translation quality across languages: **Example 3**

	SMT Input	**SMT Output**
Reference		*Yes. We also have blue, red, yellow and pink.*
Arabic	نعم لدينا ايضا ازرق احمر اصفر ووردي	Yes, we have a red and my.
Chinese	有 还 有 蓝色 红色 黄色 还 有 粉色	Do you have any blue red yellow and pink.
Italian	sì abbiamo anche blu rosso giallo e rosa	Yes, we have red yellow blue and pink.
Japanese	はい 青 赤 黄色 ピンク も ございます	Yes, we have red green yellow pink.

Conclusion

- We have described the design of HKUST system which integrates semantic processing into SMT
- We presented results on 4 different IWSLT-06 tasks
  - On Chinese-English, our system achieved reasonable performance, despite being a text-based system
  - We also reported results on 3 other language pairs from different language families