Phrase-Based Statistical MT for MANOS System

Prof. *Bo Xu*

Institute of Automation
Chinese Academy of Sciences (CASIA)
xubo@nlpr.ia.ac.cn

Phuket Thailand, Sept 15, 2005

Content

- MANOS/MT/S2S
- 1st SMT workshop in China
- Phrase-Based Statistical MT
- Conclusion and Future directions

1、 MANOS/MT/S2S

MANOS Framework

City

2008 Event

Information Service Platform

Multilingual Support

Role of MT(1/2)

- MANOS is a service platform that try to integrate commercial available or will be available services in market for events
- Multilingual is one of key features across applications
  - in business model, we could view as value-add services

Role of MT (2/2)

- Multilingual Support Tools and multilingual Interface: fast and accurate (maybe manually assistance) translation for multilingual information distribution and display
- Multilingual Interaction: Multilingual also could be independent application such as a Speech TRANSLATOR device etc. ***Mainly domain specific.***
- Not only MT also S2S,especially in PDA or Mobile phone.

MT and S2S

- MT – Traditionally refer as General domain and Text-to-text translation
  - Relatively long R&D history
  - Huajian, Xiamen, Haerbing Polytech, ICT, Beijing University, CCID….
- S2S– Traditionally refer as Specific domain of course Speech-to-Speech translation
  - From 1986 ATR and mainly from ASR community
  - CASIA …

Merging of MT/S2S

- Penetrating mutually
  - More application or domain-specific MT are preferred
  - Besides domain specific, S2S extend to connect LVCSR and unlimited domain MT(TC-STAR)
- Convergence of Research Community
  - MT Summit X
  - IWSLT
- Convergence point SMT
  - Especially the rapid progress in SMT

2、 1st SMT workshop in China

Purpose and Main participants

- Purposes:
  - To enhance the SMT research in China
  - Specifically as in beginning stage, algorithm and methods implementation and understanding
  - Planned from Oct. 2004
- Participants:
  - Institute of Automation, CAS(CASIA)
  - Institute of Computing Technology, CAS (ICT)
  - Computer Department , Xiamen University

CASIA S2S(MT) Research

- Focus on some limited domain, all are corpus-based approach
  - IF(Interlingual exchange Format)(1998-2002)
  - Word-based SMT(1998-2002)
  - EBMT(2002-2004)
  - Phrase-based SMT(2005-)
  - ……

Characteristic of Spoken Translation(1/3)

- Spontaneous speech
  - Ellipsis and fragments, High function/low content term,Vagueness, Anaphora, Juncture, repair, ….

Characteristic of Spoken Translation(2/3)

- Technology related
  - Spoken style, variable speed, non-speech utterance, break, anaphora, repair
  - Ungrammatical, recognition error, no-boundary between sentences
  - personalized voice, accent and intonation

Characteristic of Spoken Translation(3/3)

- Advantage under SMT framework
  - Simple structure
  - Short Sentences (in Chinese average around 9 characters, 7 words)
  - Though variable but still limited expression
  - For real application – balance between expressive and simplicity
- We realize the importance of solving limited domain problem using unlimited domain technology is absolute necessary

Institute of Computing Technology(ICT)

- Lead by Prof. Qun LIU
- Is expertise in NLP, including different Tools, analysis, corpus construction and MT(Some are available in open source)
- They setup three systems called PBT, AT and SBTG under SMT framework and same phrase dictionary

Xiamen University

- Lead by Prof. Xiaodong Shi
- One of most famous MT system in China
- Rule-based translation R&D for more than 20 years
- Expertise in algorithm
- Began SMT research from Summer of 2004

Corpus Preparing-- training

- CASIA50K: 50K bilingual corpus in travel domain by CASIA
- ICT150K:150K bilingual corpus of movie caption
- XMU200K: 200K bilingual corpus of movie caption
- All sentences are not very long, not very large because the purpose of the workshop and some copyright problem

Corpus Preparing-- testing

- CASIA1500: 1500 test corpus with every sentence 5 translations by CASIA
- 863-03 and 863-04 standard dialogue test in previous 2 years
  - 863-03: 350 (4 translation)
  - 863-04: 400 (4 translation)

About Workshop

- Held from July 13-14
- Email discussion and result exchange before workshop
- Two-days workshop
  - On site evaluation
  - System technical report for every group
  - Discussing

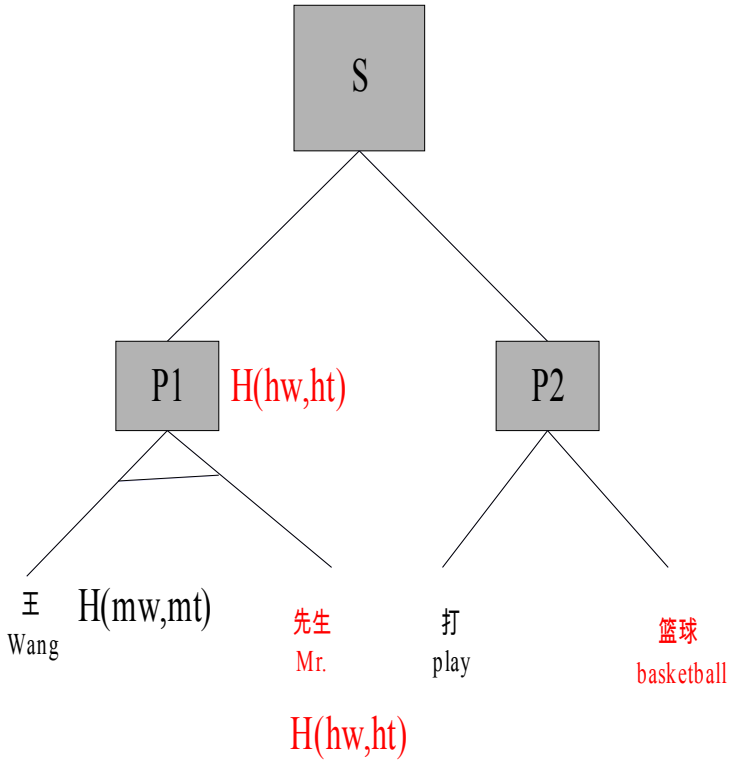
Final 5 systems

- All are phrase-based
- PBT1-3: Phrase-based decoding, no other additional syntax and semantic knowledge added
- AT– Alignment template
- SBTG--Stochastic BTG

Alignment Templates

- Try to increase the generalization ability of system, quite preliminary implementation for participating system
- Monolingual clustering through MKCLS or Bilingual Clustering
- Decoding: Phrase template applied when no phrase corresponding are available
- Class ： 2 3 14 15 20 pro: 0.33
- (发表 声明 issued a statement)
- Template is : 2 3 -> 14 15 20 0.33

SBTG Framework



- PCFG Syntax parsing to source language (Collins)
- Using extracted phrase information in process of search
- Make decision in every node mono|invert
  - P(e|e') * p_m or
  - P(e'|e) * (1-p_m)
  - P(e|e') is language model, P_m are trainned from PenTree Bank

T1: Training: casia50k, Test:casia1500

Results	Nist	bleu	GTM	mWER	mPER
PBT1	6.673	0.268	0.623	0.597	0.486
PBT2		0.3143			
AT	6.6871	0.3146	0.6627	0.5054	0.4149
SBTG	7.0086	0.3283	0.6765	0.5157	0.4177
PBT3	7.0647	0.3511	0.6897	0.4769	0.3969

T2: Training: casia50k test:863_03

Results	Nist	bleu	GTM	mWER	mPER
PBT1	6.078	0.222	0.638	0.676	0.517
PBT2		0.2679			
AT3	6.6844	0.2696	0.6637	0.6060	0.4612
SBTG4	6.2962	0.2425	0.6500	0.6367	0.4857
PBT3	6.4168	0.2833	0.6338	0.5885	0.4840

T3: Training casia50K, Test: 863_04

Results	Nist	bleu	GTM	mWER	mPER
PBT1	4.79	0.179	0.573	0.677	0.569
PBT2		0.1225			
AT	4.5544	0.1287	0.5575	0.6811	0.5521
SBTG	4.5842	0.1191	0.5385	0.7650	0.5997
PBT3	3.8910	0.1105	0.5207	0.6944	0.5899

Remarks(1)

- The best result of PBT is superior than other two systems (AT and SBTG)
- PBT3 performance better in casia1500 and 863-03,however, PBT1 is much better than 863-04. The reason is that 863-04 have more questioning sentence and also long sentences,
- PBT1 decoding is with ability of re-ordering during search; PBT2 and PBT3 adopt relatively simply decoding(PBT1 take 8m, PBT2 and PBT3 only take 4s for casia1500 testing)

T4:training ： ict150k Test: casia1500

Results	Nist	bleu	GTM	mWER	mPER
PBT1	**4.15**	**0.10**	**0.46**	**0.82**	**0.69**
PBT2		**0.0816**			
AT	**3.6340**	**0.0771**	**0.4150**	**0.8190**	**0.7088**
SBTG	**4.6469**	**0.1265**	**0.4649**	**0.8062**	**0.6896**
PBT3	**3.5894**	**0.0948**	**0.4258**	**0.7965**	**0.7027**

T5 ： training: xmu200k, test casia1500

Results	Nist	bleu	GTM	mWER	mPER
PBT1	4.63	0.117	0.487	0.808	0.675
PBT2		0.1266			
AT	3.9264	0.1213	0.4480	0.7778	0.6766
SBTG	4.0516	0.0927	0.4258	0.8440	0.7259
PBT3	4.0930	0.1416	0.4680	0.7507	0.6594

T6: training(ict150k+xmu200k), test: casia1500

Results	Nist	bleu	GTM	mWER	mPER
PBT1	4.597	0.116	0.486	0.813	0.677
PBT2		0.1274			
AT	3.9923	0.1292	0.4566	0.7647	0.6658
SBTG	4.5816	0.1275	0.4584	0.8114	0.6966
PBT3	4.1439	0.1445	0.4705	0.7466	0.6586

T7 : training(ict150k+xmu200k+casia50k), Test: casia1500

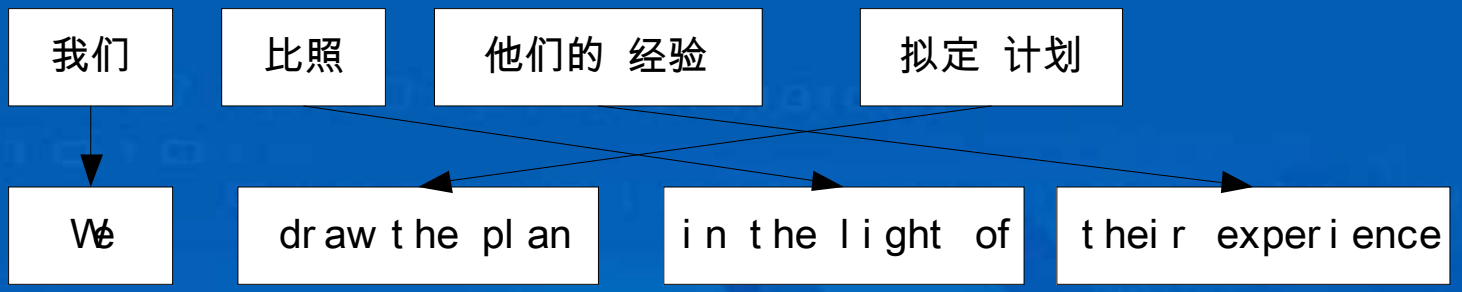
Results	Nist	bleu	GTM	mWER	mPER
PBT1	**4.66**	**0.117**	**0.483**	**0.818**	**0.685**
PBT2		**0.1279**			
AT	**4.0219**	**0.1298**	**0.4598**	**0.7648**	**0.6652**
SBTG	**4.5463**	**0.1276**	**0.4555**	**0.8134**	**0.6953**
PBT3	**4.1298**	**0.1459**	**0.4806**	**0.7386**	**0.6493**

Remarks(2)

- SBTG has certainly more generalization ability in some cases, but not robust enough to benefits from different corpus
- AT are not in stable condition
- PBT1-3 have steadily performance when changing the training corpus
- Through exchange of results before workshop, there is so “surprising” result among three groups --- reach to general normal state

3、 Phrase-based Statistical Translation

Phrase-Based Translation Model



- Source input is segmented into all possible phrases
  - Any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Extraction of phrase translation

- **Integrated segmentation and phrase alignment (ISA, Zhang, 2003)**
- **Extracting phrase pairs from HMM word alignment model (HMM, Vogel et al., 1996)**
- **Phrase-extraction using Inversion Transduction Grammar (BTG, Wu, 1997)**
- **Phrases from bi-direction Word-Based Alignment (WBA, Och et al, 1999).**
- **……**

Probability of Phrase Pairs

- We need a probability distribution over collected phrase pairs

  → Possible Choices:

  - using lexical translation probabilities
  - relative frequency of collected phrases

$$\phi(f/e) = \frac{count(f,e)}{\sum_{f} count(f,e)}$$

- or, conversely $\phi(e/f)$

Phrase Translation

- Phrase translation for “我 要 买”

English Phrase	$\phi(f/e)$
I want to buy	0.386
I would like to buy	0.234
I will buy	0.119
I wanna buy	0.108
I wan to get	0.101
……	……

Beam-Search Decoding

- **Look up possible phrase translations**
  - •many different ways to segment words into phrases
  - •many different ways to translate each phrase

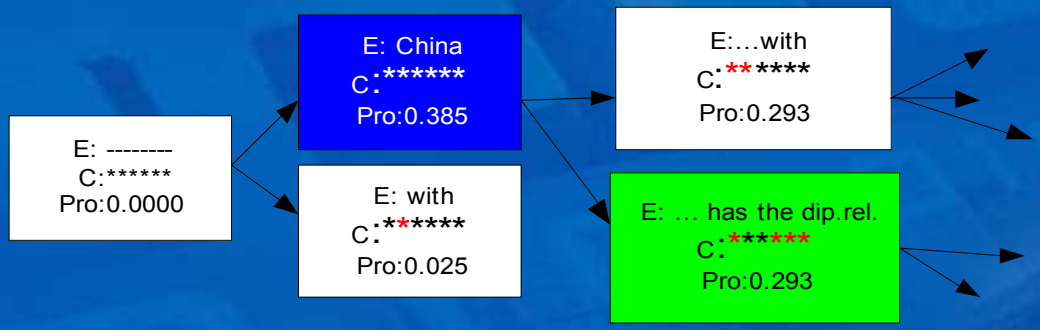
中国	与	北朝鲜	有	外交	关系
China	with	North Korea	has	diplomatic	relationships

with North Korea

diplomatic relationships

has the diplomatic relationships

China has the diplomatic relationships with North Korea



IWSLT2005 evaluation(1/2)

- Training Corpus
  - 1,000K domain-specific corpus(C-Star,CASIA …. etc)
  - 500K general domain news corpus(863,973, HIT)

IWSTL2005 Evaluation(2/2)

Track (C-E)	Data condition	Bleu4	NIST	Meteor	WER	PER
Manual transcription	unrestricted	0.5279	10.2499	0.7214	0.4160	0.3366
ASR Output	unrestricted	0.3845	8.0406	0.5802	0.5788	0.4770

PBT vs. Other methodology(1/2) 863-03 Test

Training and Test corpus	NIST	BLEU
nlpr50K, 863-03Test	6.0	0.22
cstar130K training, 863-03 Test	6.82	0.28
IWSLT2005,863-03 test	7.28	0.32
Best performance in 03 evaluation(rule-based or hybrid)	7.77	0.36

PBT vs. Other methodology(2/2) 863-04 Test

Training and Test corpus	NIST	BLEU
IWSLT2005	5.91	0.22
Best performance in 863-04 evaluation	6.12	0.21

Remarks(3)

- Our Phrase-based translation is still a bit lower but comparable than best rule-based
  - We only less than one year real experiences in SMT
  - World-class commercial Rule-based systems

Phrase-template

- Weakness of Phrase-based Model
  - lack of generalization
  - difficult to estimate the quality and quantity of phrases extracted

->Phrase-Template

- can contain phrase variables
- enables dedicated modeling

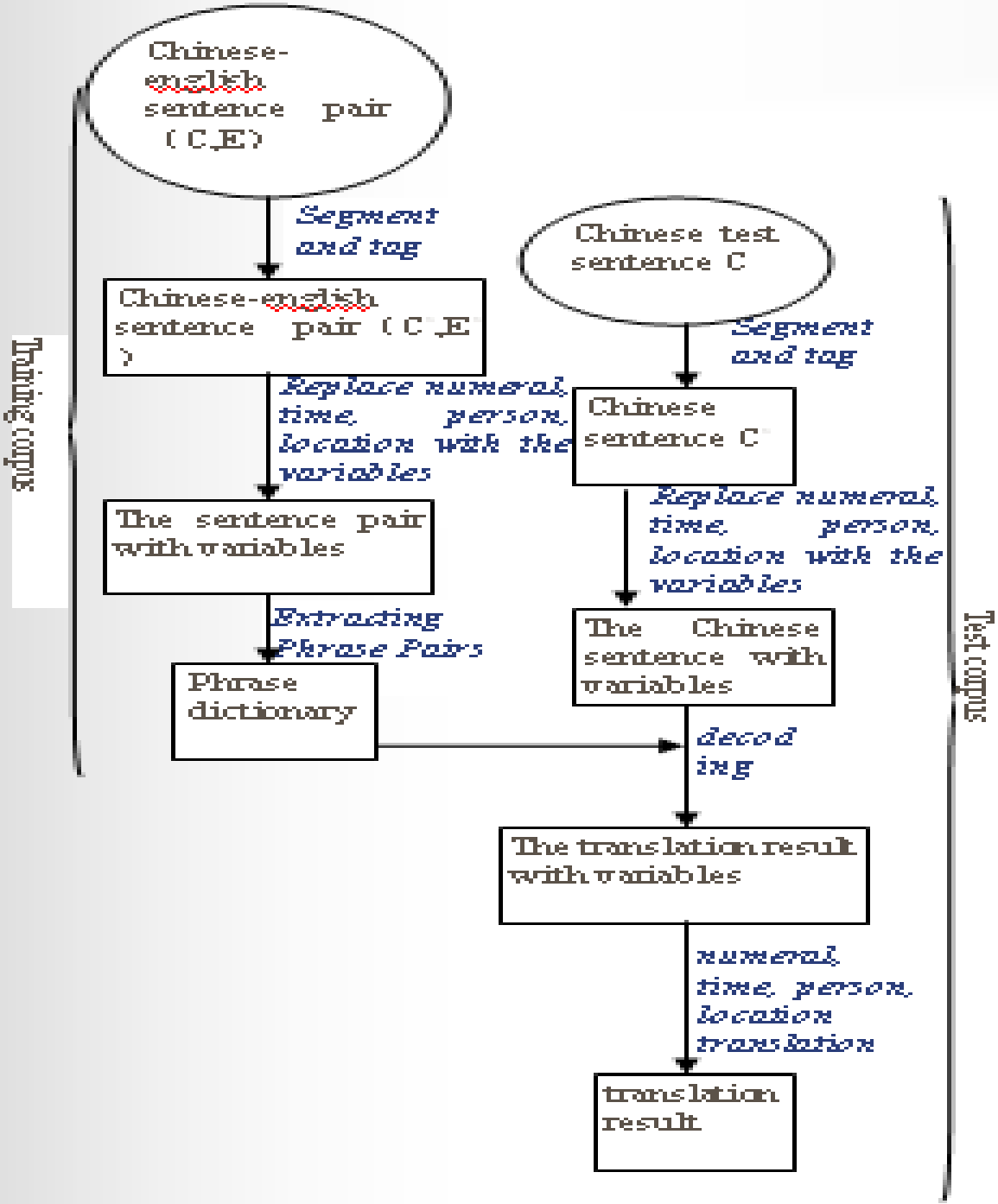
Phrase variables(1/2)

- Variable Selection:
  - Alignment template: word clustering: very difficult to balance the robustness and accuracy
  - Numeral, time, person, location(Name entity)
  - Phrase with (Numeral, time, person and location) are phrase-template with variables
- 上海 到 北京 - 》 from shanghai to beijing
  - LOC 到 LOC -> from LOC to LOC 0.51

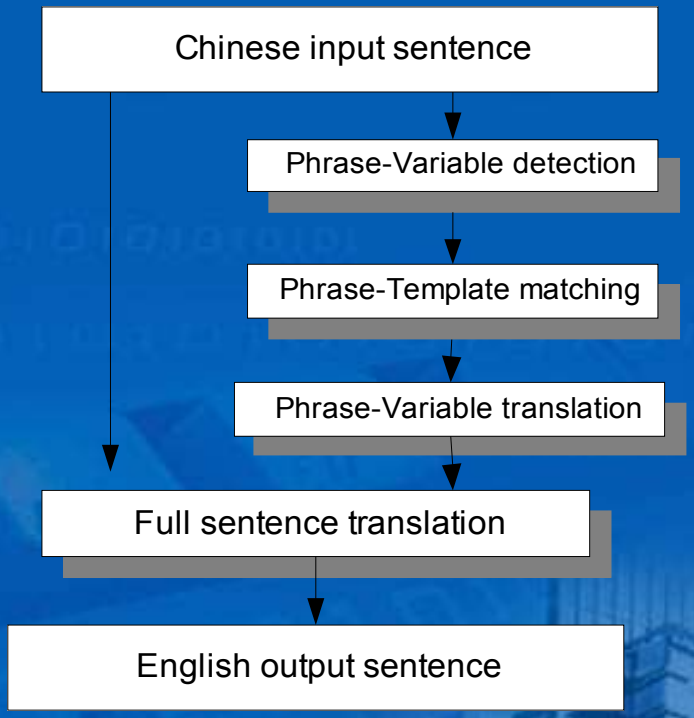
Phrase variables(2/2)

- **Variable are translated independently**
  - **When we have the translation result with variables, we can use dictionary of person and location to translate person and location variables, and use numeral model translate numeral and time.**
- 我要去上海 -〉我 要 去 _loc -> I want to go to _loc->I want to go to shanghai.

Shallow Parsing(maybe latter deep paring) is necessary to extract that variables.



Phrase template detection and translation



- About 5% extracted phrases which contain variable (phrase-variables focus on Name entity: Time, Number, Location and Person)

Phrase-template result for 863-04

- IWSLT2005 :
  - NIST 5.9143 BLEU 0.2241 GTM 0.6369 mWER0.6353 mPER05106
- Phrase-Template:
  - NIST score = 5.9882 BLEU score = 0.2320 GTM score = 0.6408 mWER score = 0.6366 mPER score = 0.5098

4、 Conclusion and direction

Conclusion

- MT and S2S are converging
- Statistical MT has been initially investigated in China that preliminary result is comparable with the state-of-art rule-based or hybrid system
- Phrase-based has shown to be superior to other system by now in view of implementation and accuracy
- Phrase with variables or phrase template has been initially tried to have some improvement in accuracy

Future direction

- Way of merging EBMT and SMT
  - phrase template
  - What kind of words or parameter could be variables
    - Besides time\number,name entity like \name\location …..
  - Need to integrate more advanced preprocessing (shallowing paring to medium-depth parsig)
- But Systematical integration of structure knowledge-morphological, syntax and so on ….

Thanks ！