Generating Chinese Couplets using a Statistical MT Approach

Ming Zhou

Microsoft Research Asia

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Chinese Couplet (对联)

- A special type of poetry
  - Composed of two sentences with same length and in a regular form
- One of the most important Chinese cultural heritages
  - Originates as early as the Five Dynasties (907 AD - 960 AD)
- An example:
  - First sentence (FS): “海 阔 凭 鱼 跃” (sea wide allow fish jump)
  - Second sentence (SS): “天 高 任 鸟 飞” (sky high permit bird fly)
  - The sea is wide enough so that fish can jump unrestrictedly; and the sky is high enough so that bird can fly at their pleasure.

Problem Definition

- Given the FS, to generate the SS so that the two sentences can form a qualified Chinese couplet
  - For example
    - Input: “海 阔 凭 鱼 跃” (sea wide allow fish jump)
    - Output: “天 高 任 鸟 飞” (sky high permit bird fly)

- A popular language game since one thousand years ago

- The difficulty of this problem comes from the principles of Chinese couplet

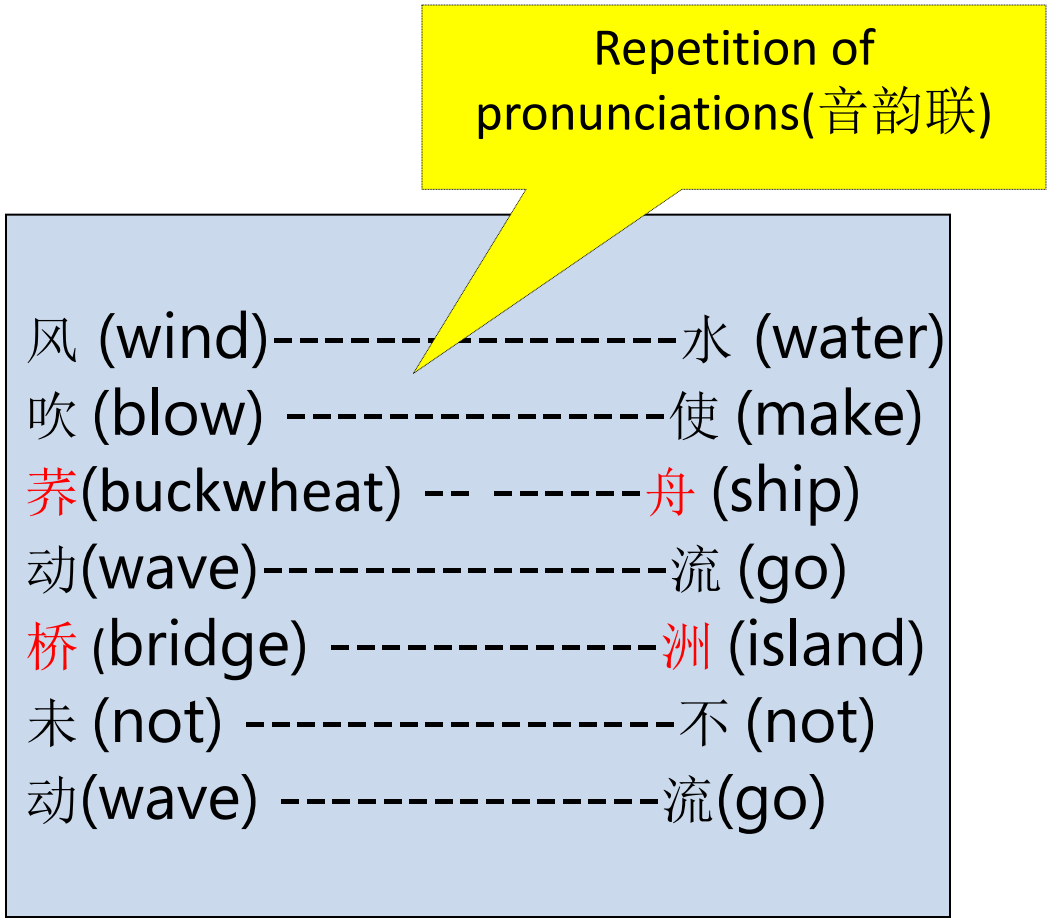
Chinese Couplets (http://duilian.msra.cn)

CCTV

中国中央电视台
CHINA CENTRAL TELEVISION

http://video.sina.com.cn/v/b/10937201-1452530713.html

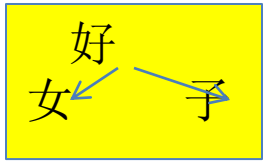
FS and SS Share the Same Style

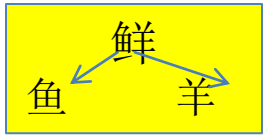


FS and SS Share the Same Style

Decomposition of characters (拆字联)

有 (have)---------- 缺 (lack)
子 (son) ----------鱼 (fish)
有 (have) ----------缺 (lack)
女 (daughter)----------羊 (mutton)
方 (so) ----------敢 (dare)
称 (call) ----------叫 (call)
好(good) ----------鲜(fresh)





FS and SS Share the Same Style

Person name (人名联)

Palindrome (回文联)

板桥(Banqiao)---------- 东坡 (Dongpo)
造(produce) ----------居 (live)
桥(bridge) ----------坡 (mountain)
板(board)----------东(east)

- Banqiao(板桥) and Dongpo(东坡) are famous litterateurs
- Reading from top to down is identical to down to top

Principle 1

- The FS and SS should agree in length and word segmentation
  - FS: 知识 能 致 富 (knowledge can bring richness)
  - SS: 勤劳 可 兴 家 (work can raise family)
  - English translation: knowledge can make one rich and hard work can make one's family live better.

Principle 2

- Corresponding words in FS and SS should agree in their part of speech

海	阔	凭	鱼	跃
sea	wide	allow	fish	jump
天	高	任	鸟	飞
sky	high	permit	bird	fly
noun	adjective	conjunction	noun	verb

Principle 3

- The contents of the FS and SS should be related but cannot be duplicated
  - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)
  - SS: 天 高 任 鸟 飞 (sky high permit bird fly)
  - Examples in different situations: fish in the sea and bird in the sky
  - Same truth that only in broad space one can use all one's talents

Principle 4

- The last character of the FS should be pronounced in “仄”(Ze) tone
- The last character of the SS should be in “平”(Ping) tone

  - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)
  - SS: 天 高 任 鸟 飞 (sky high permit bird fly)

  - 跃-> “仄”(Ze)
  - 飞-> “平”(Ping)

Principle 5

- The writing styles of the FS and SS should be identical
  - Character repetition
  - Pronunciation repetition
  - Character decomposition

    - FS: “有 女 有 子 方 称 好”(have daughter have son so call good)
    - SS: “缺 鱼 缺 羊 敢 叫 鲜”(lack fish lack mutton dare call delicious)

MT vs. SS Generation

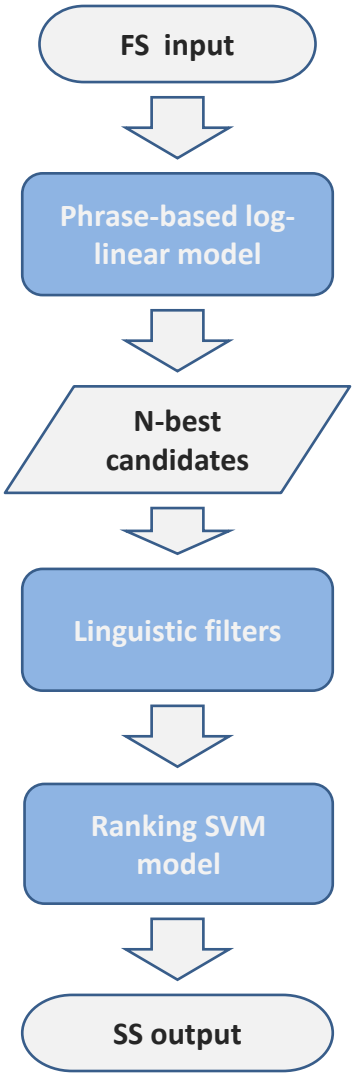
- Machine translation
  - He sent her a bunch of flowers .
  - 他 给 她 送 了 一 束 花 。
- SS generation
  - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)
  - SS: 天 高 任 鸟 飞 (sky high permit bird fly)
  - No word insertion, deletion and reordering

Outline

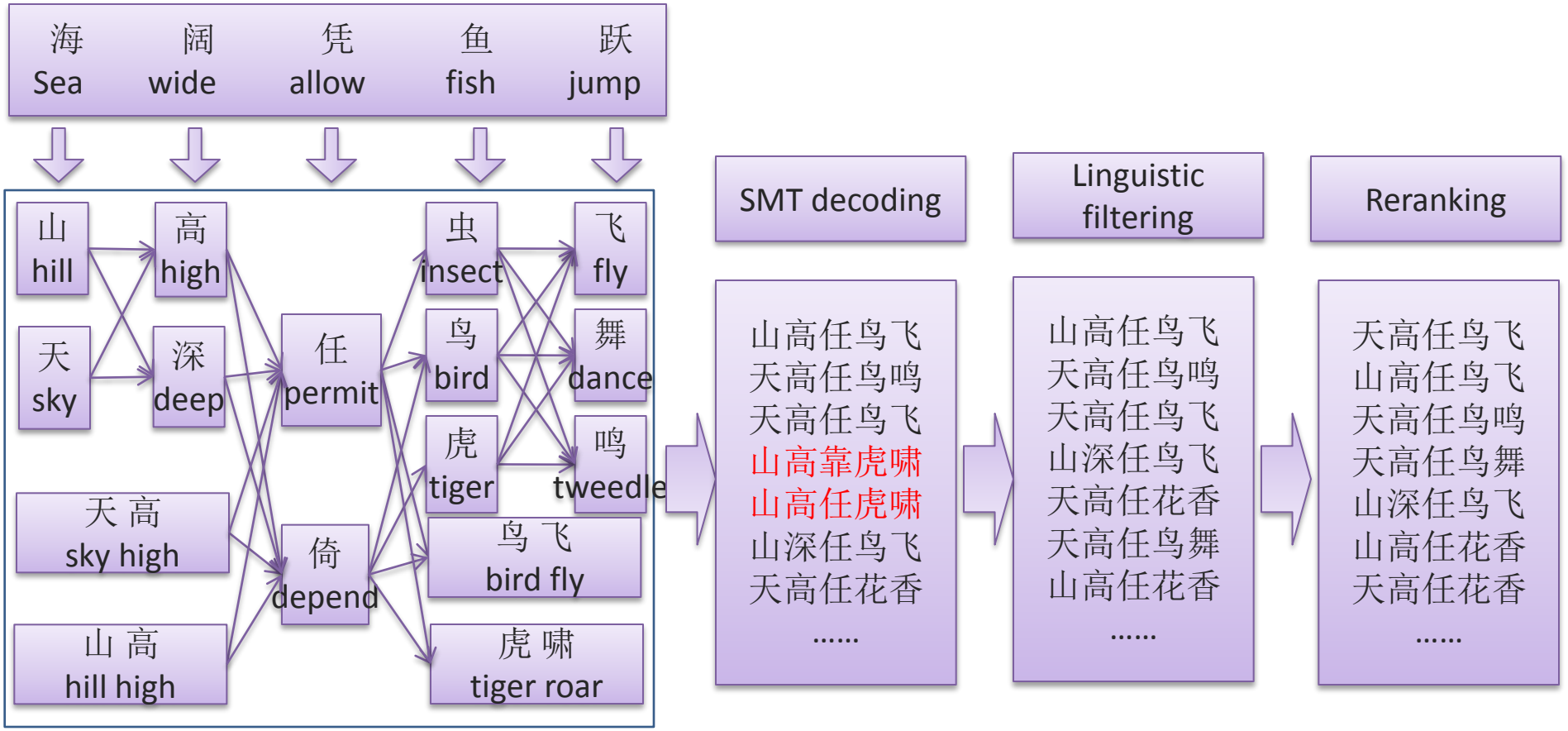
- Introduction
- Couplet generation model
- Experimental results
- Conclusions

SS Generation Approach

- A multi-phase SMT approach
  - Phase1: a phrase-based log-linear model
  - Phase2: some linguistic filters
  - Phase3: a Ranking SVM



SS Generation Process



Phrase-based Log-linear Model

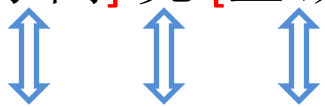
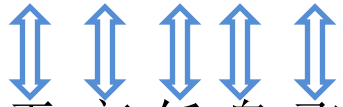
- Given a FS denoted as *F={f1, f2, .. .,fn}*, to seek a SS denoted as *S*={s1, s2, …, sn}* that satisfies

$$S^* = \arg\max_{S} \sum_{i=1}^{M} \lambda_i \log h_i(S, F)$$

  - Where f_i and s_i are Chinese characters

- Five feature functions
  - Phrase translation model (PTM)
  - Inverted PTM
  - Character translation model (CTM)
  - Inverted CTM
  - Language model

Feature Functions

- Phrase translation model and inverted PTM
  - FS: [海 阔] 凭 [鱼 跃] ([sea wide] allow [fish jump])
  - 
  - SS: [天 高] 任 [鸟 飞] ([sky high] permit [bird fly])
- Character translation model and inverted CTM
  - FS: 海 阔 凭 鱼 跃 (sea wide allow fish jump)
  - 
  - SS: 天 高 任 鸟 飞 (sky high permit bird fly)
- Language model
  - Character-based trigram model
    - p(海 阔 凭 鱼 跃)= p(海|START)*p(阔| START 海)*p(凭|海 阔)*p(鱼|阔 凭)*p(跃|凭 鱼)

Training Data

- Couplet Data
  - Classic Chinese couplets
    - From books
    - From the web
  - Extract sentence pairs from ancient Chinese poems
  - From couplet forums
  - 970,000 couplets obtained finally

- LM training
  - Chinese poems besides the couplet data for LM training
  - 1,600,000 sentences in total

Linguistic Filters

- Only SMT model can not guarantee the SS has
  - The same writing styles as the FS
  - Correct tone for the last character

- Four filters
  - Character repetition filter
  - Pronunciation repetition filter
  - Character decomposition filter
  - Phonetic harmony filter

Linguistic Filters 1

- Character repetition filter
  - The FS
    - “有 女 有 子 方 称 好” (have daughter have son so call good)
  - SS candidates
    - “缺 鱼 缺 羊 敢 叫 鲜” (lack fish lack mutton dare call delicious)

      √
    - “缺 鱼 少 羊 敢 叫 鲜” (lack fish miss mutton dare call delicious)

      ×

Linguistic Filters 2

- Pronunciation repetition filter
  - The FS
    - “风 吹 荞 动 桥 未 动” (wind blow buckwheat wave bridge not wave)
  - SS candidates
    - “水 使 舟 流 洲 不 流” (water make ship move island not move)

      √
    - “水 使 舟 流 岛 不 流” (water make ship move island not move)

      ×

Linguistic Filters 3

- Character decomposition filter
  - The FS
    - “有 女 有 子 方 称 好” (have daughter have son so call good)
  - SS candidates
    - “缺 鱼 缺 羊 敢 叫 鲜” (lack fish lack mutton dare call delicious)

      √
    - “缺 鱼 缺 牛 敢 叫 鲜” (lack fish lack beef dare call delicious)

      ×

Linguistic Filters 4

- Phonetic harmony filter
  - The FS
    - “海 阔 凭 鱼 跃” (sea wide allow fish jump)
    - 跃 : Ze
  - SS candidates
    - “天 高 任 鸟 飞” (sky high permit bird fly)
    - 飞: Ping

      √
    - “山 高 任 虎 啸” (mountain high permit tiger roar)
    - 啸: Ze

      ×

Candidate Re-ranking

- Ranking SVM for re-ranking SS candidate
  - To leverage long-distance features

$$f_{\vec{w}}(\vec{x}) = \langle \vec{w}, \vec{x} \rangle$$

- Two more features
  - Mutual information (MI)
  - MI-based structural similarity (MISS)

- Parameter estimation
  - Tool: SVM Light
  - Training data: 200 FSs and each of them has 50 SSs labeled as positive or negative by human

Candidate Re-ranking (con't)

- Mutual information (MI)
  - Motivation
    - Candidate 1: "天 高 任 鸟 飞" (sky high permit bird fly)
    - Candidate 2: "天 高 任 狗 叫" (sky high permit dog bark)
    - Candidate 1 is better
      - MI(天,鸟) > MI(天,狗) (MI(sky, bird) > MI(sky, dog))
  - To measure the semantic consistency of words in a candidate SS

$$MI(S) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(s_i, s_j) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \log \frac{p(s_i, s_j)}{p(s_i)\, p(s_j)}$$

Candidate Re-ranking (con't)

- MI-based structural similarity (MISS)
  - Motivation: structural similarity in Chinese couplets
    - 海 阔 凭 鱼 跃 (sea wide allow fish jump)

    - 天 高 任 鸟 飞 (sky high permit bird fly)

  - To measure the structural similarity

$$MISS(F,S)=\cos(V_f,V_s)=\frac{V_f \bullet V_s}{|V_f|\times|V_s|}$$

    - Given the FS $F=\{f_1, f_2,..., f_n\}$, we first build its MI vector

$$V_f=\{MI_{12}, MI_{13},.., MI_{1n}, MI_{23}.., MI_{n-1n}\}$$

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Automatic Evaluation of SS

- BLEU for SS evaluation

$$BLEU = BP \bullet \exp(\sum_{n=1}^{N} w_n \log p_n)$$

  - N = 3; BP = 1
  - P_n is position-sensitive
- Data set
  - 1051 FSs and their SSs mined from couplet forums
    - 24 references for each FS on average
  - 600 as development set and 451 as test set

Translation Unit Setting

- Experiment setting
  - Only translation model and language model are used
  - Same training data, same linguistic filters and no re-ranking

Translation Unit setting	BLEU
character-based	0.236
word-based	0.261
phrase-based	0.276

Feature Evaluation

- Incrementally adding new features
- Same linguistic filters for all settings

	Features	BLEU
Baseline	Phrase TM(PTM) + LM	0.276
Phrase-based SMT Model	+ Inverted PTM	0.282
	+ Character TM (CTM)	0.315
	+ Inverted CTM	0.348
Ranking SVM	+ Mutual information (MI)	0.356
	+ MI-based structural similarity	0.361

Overall Performance

- By human evaluation
  - 100 FSs
  - Output 10 best SSs by our best system for each FS
  - Human labeling: acceptable or not
  - Metric
    - *top-n inclusion rate* is defined as the percentage of the test sentences whose top-n outputs contain at least one acceptable SS.

	Top-1	Top-10
Top-n inclusion rate	0.21	0.73

新浪推出基于微软专利的手机对联服务

CNET中国·PChome.net·编译 作者: 责编:江海明 时间:2009-01-08 **标签:** 手机短信 MSRA 春节短信

2009年1月7日，美国雷德蒙及中国北京——微软公司今天与新浪公司共同宣布，双方签署了一项有关“中文对联生成器”技术的专利许可协议，它是微软亚洲研究院（MSRA）在自然语言处理领域的尖端创新之一。此项专利授权将增强新浪在中国市场提供创新性移动增值服务的能力。

根据这项许可协议，微软的专利技术将用于新浪的全新移动电话对联服务，手机用户可以将自己编写的上联以短信形式发送至新浪的服务器，服务器上运行的微软中文对联生成引擎将自动构造下联以及横批，并以彩信或短信的形式发送给用户。这项服务将于2009年1月6日起投入运作，迎接新春佳节（2009年1月26号），届时人们将发送数十亿短消息互致新年问候。新浪还计划利用微软的核心技术自行开发其他创新产品，为这项技术寻找新用途，并且使移动增值服务更加具有个性、互动性和娱乐性。

“对联的分享与展示是中国千百年来的一项春节传统。我们非常高兴，因为此次与微软的全新合作将帮助我们把这一习俗搬到电脑和移动电话上。”新浪公司副总裁，新浪无线总经理王高飞表示。“通过直接与微软研究院合作，我们实现了先进的机器学习和语言技术与我们自有创新的成功结合，扩展了这项基于Web的技术，用于手机和短信，进而提供一项对于中国乃至全球华人客户十分有意义的服务。”

“我们感到高兴的是，新浪选择了微软研究院自然语言处理技术的使用许可证。”位于北京的微软亚洲研究院院长洪小文博士说：“这项协议是一个很好的例子，我们向来致力于通过知识产权的合作提升消费者体验。这也是我们与中国本地企业展开更密切合作的重要步骤之一，让中国IT产业更具活力。”


万物生辉
心向蓝天风和日丽
面朝大海春暖花开
微软亚洲研究院


春兰秋菊
花客樵夫过客
竹人菜农旅人
微软亚洲研究院


绰有余裕
赢马更赢酒
输人又输球
微软亚洲研究院


风月无涯
蚕为天下宝
鸿是江边鸟
微软亚洲研究院

User log for Model Enhancement

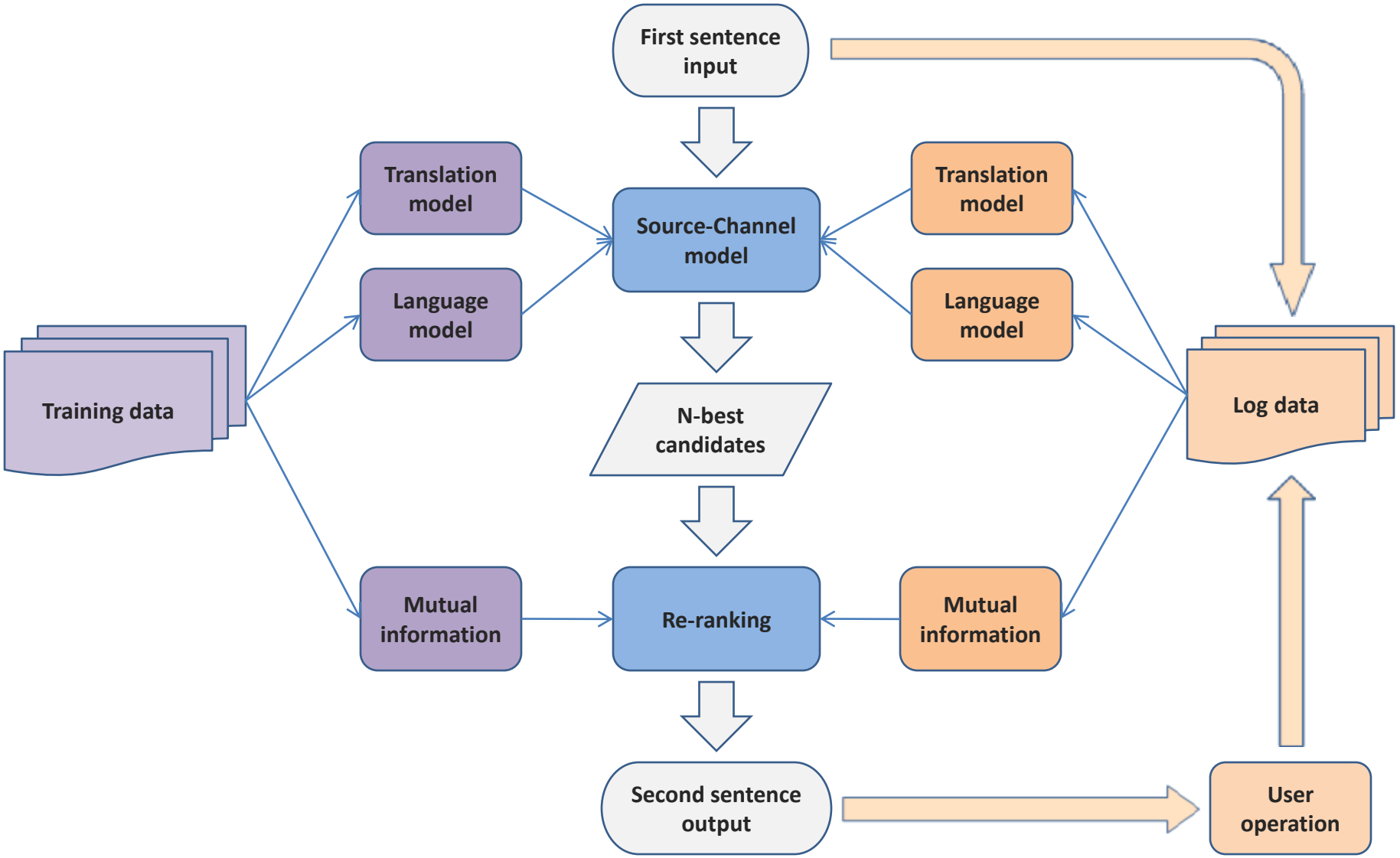
- Motivat
  - Traini
  - While
- What lo
  - User i
  - User
    - Se
    - U

User's Log Analysis

Number of input sentences	12,322
Number of unique input sentences	6,698
Users directly select from system output	3,459
User manual modify system output	606
Save as favorite couplets	109
Invalid user input	618
No second sentence generated	2,211
Banner generation	2,687
Select the generated banner as favorite	428
No banner output	265

- Data Source
  - Log from http://couplet.msra.cn
- Time period
  - Aug. 31-Oct. 9, 2006

New Framework with Log Data



Extended to Quatrains(绝句)

- 感归

零落泣鬼神
秋来愁人心
肝肠断挥洒
不思归日吟

- 春兴

残花飘黄叶
细雨落青山
蝶飞红杏里
燕舞绿杨湾

- 从军北征

雁字风月一时清
天书云山千里远
锦字凭谁寄笔力
人生何日归来晚

- 望洞庭

移舟雨逐行云水
一路风随日月天
云破长江万里船
风来一水千山川

Outline

- Introduction
- Couplet generation model
- Experimental results
- Conclusions

Conclusions

- Conclusions
  - A multi-phase SMT approach for generating the second sentence given a first sentence of a Chinese couplet
  - Promising experimental results obtained
  - A popular website (http://couplet.msra.cn)
    - 50,000 visitors / day during the peak time
- Future work
  - Word clustering for smoothing TM
  - As a extension, work on Chinese poetry generation

Thanks!
Questions?

The Game of Second Sentence Generation

创大业一帆风顺

展宏图万事胜意

Couplet Web Service

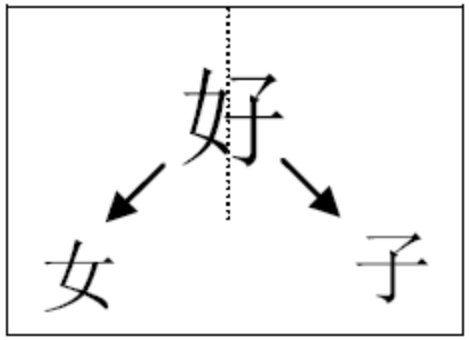
- S

Couplet Web Service (con't)

- St
  a

Character Decomposition

- Character decomposition is an interesting language phenomenon in Chinese: some Chinese characters can be decomposed into other characters.
- For example, “好” (good) can be decomposed into “女” (daughter) and “子” (son).



Some Reference Examples

- The FS
  - 品 茶 不 为 渴 (degust tea not because thirstiness)
- References:
  - 弹 曲 却 因 情 (play zither but for feeling)
  - 踏 雪 只 因 梅 (trample snow only for plum)
  - 醉 酒 总 关 情 (drink wine always relate feeling)
  - …

Couplet in Poems

- Eight-line Poem

春望

国破山河在，
城春草木深。
感时花溅泪，
恨别鸟惊心。
烽火连三月，
家书抵万金。
白头搔更短，
浑欲不胜簪。