A Phonetic-Based Approach to Chinese Chat Text Normalization

Kam-Fai Wong

Dept of Systems Engineering & Engineering Management,
The Chinese University of Hong Kong, China

Table of Content

Part 1: Web 2.0

Part 2: Network Informal Languages

Part I: Web 2.0

1. What is Web 2.0?
2 Resemblance to our Society
3 Design Considerations
4 Example
5 Research Challenges

1. What is Web2.0?

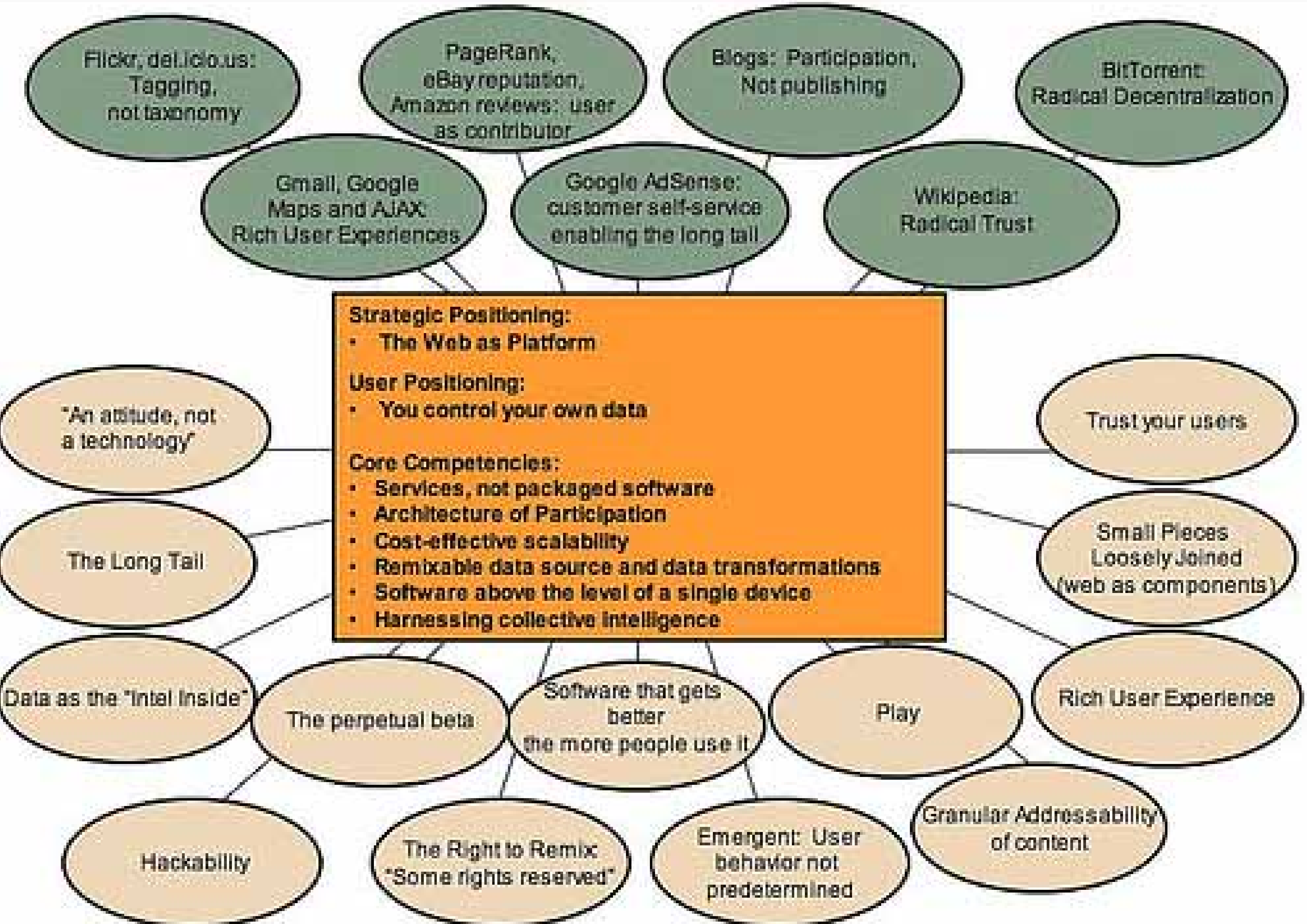
- Web 2.0 = Ubiquitous Knowledge Base
- More than "Web as Platform" (Tim O'Reilly), "Web is Life"
- Knowledge (a) owned by people; (b) accessible anywhere; (c) available anytime.
- Operations: retrieval, extraction, integration, sharing

2. Resemblance to our Society

Core Competencies

- Services industry (服務工業)
- Syndicalism (工團主義)
- Mutual benefits (互利互惠)
- Division of labor (分工合作)
- Wisdom of crowds (集體智慧)
- Adoptability (適應能力)
- Collective intelligence (一人計短，二人計長)
- Information Anarchy
- Evolution of knowledge

Web 2.0 Meme Map



Flickr, del.icio.us: Tagging, not taxonomy

PageRank, eBay reputation, Amazon reviews: user as contributor

Blogs: Participation, Not publishing

BitTorrent: Radical Decentralization

Gmail, Google Maps and AJAX: Rich User Experiences

Google AdSense: customer self-service enabling the long tail

Wikipedia: Radical Trust

Strategic Positioning:

- **The Web as Platform**

User Positioning:

- **You control your own data**

Core Competencies:

- **Services, not packaged software**
- **Architecture of Participation**
- **Cost-effective scalability**
- **Remixable data source and data transformations**
- **Software above the level of a single device**
- **Harnessing collective intelligence**

"An attitude, not a technology"

Trust your users

The Long Tail

Small Pieces Loosely Joined (web as components)

Data as the "Intel Inside"

The perpetual beta

Software that gets better the more people use it

Play

Rich User Experience

Hackability

The Right to Remix: "Some rights reserved"

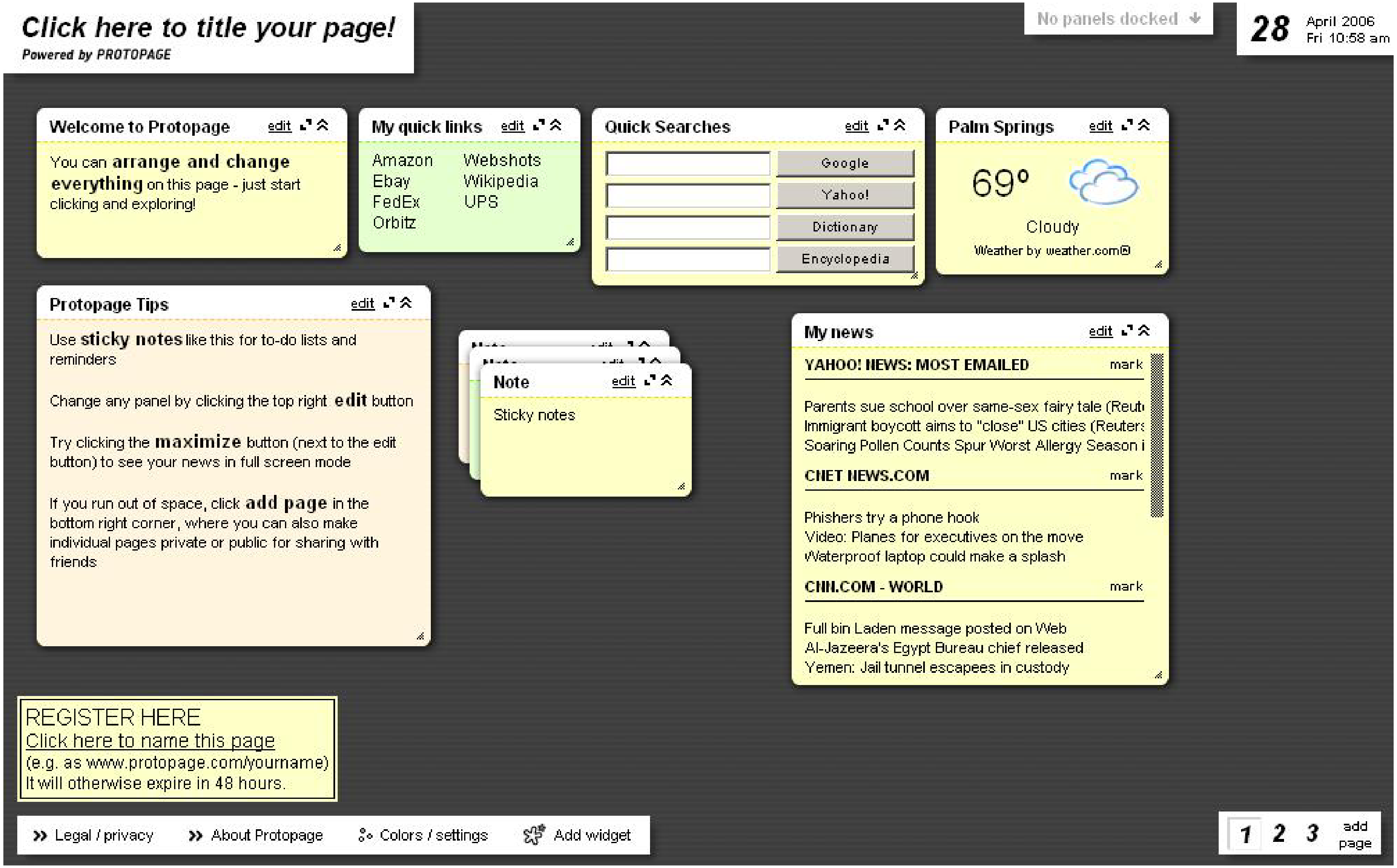
Emergent: User behavior not predetermined

Granular Addressability of content

3. Application Design Considerations

- Interoperability (data, process)
- Lower re-use barrier
- Perpetual beta
- Some rights reserved
- Flat hierarchy
- Resource sharing (content & process)
- Fine grain content ownership
- Short development cycle
- Widely open communications
- Criticism handling
- Media sensitivity

4. Web 2.0 Example Snapshot



Click here to title your page!
Powered by PROTOPAGE

No panels docked

28 April 2006
Fri 10:58 am

Welcome to Protopage edit

You can **arrange and change everything** on this page - just start clicking and exploring!

My quick links edit

Amazon	Webshots
Ebay	Wikipedia
FedEx	UPS
Orbitz	

Quick Searches edit

Google
Yahoo!
Dictionary
Encyclopedia

Palm Springs edit

69°
Cloudy
Weather by weather.com®

Protopage Tips edit

Use **sticky notes** like this for to-do lists and reminders

Change any panel by clicking the top right **edit** button

Try clicking the **maximize** button (next to the edit button) to see your news in full screen mode

If you run out of space, click **add page** in the bottom right corner, where you can also make individual pages private or public for sharing with friends

Note edit

Sticky notes

My news edit

YAHOO! NEWS: MOST EMAILED mark

Parents sue school over same-sex fairy tale (Reut
Immigrant boycott aims to "close" US cities (Reuters
Soaring Pollen Counts Spur Worst Allergy Season i

CNET NEWS.COM mark

Phishers try a phone hook
Video: Planes for executives on the move
Waterproof laptop could make a splash

CNN.COM - WORLD mark

Full bin Laden message posted on Web
Al-Jazeera's Egypt Bureau chief released
Yemen: Jail tunnel escapees in custody

REGISTER HERE
Click here to name this page
(e.g. as www.protopage.com/yourname)
It will otherwise expire in 48 hours.

» Legal / privacy   » About Protopage   Colors / settings   Add widget

1 2 3 add page

http://www.protopage.com

5. Research Challenges

- Knowledge definition = Micro-content, e.g. XML data, RSS, wiki, blog
- Retrieval: dynamic knowledge (e.g. RSS, blogs), tagging-based clustering
- Extraction: Network Language, Opinion Mining
- Integration: personalization (e.g. services)
- Sharing: social network, e.g. privacy, ethics, etc.

Remark

A Government is obliged to look after her citizens. Web is a growing society The e-Citizen should be looked after by the e-Government. Otherwise, watch out for e-crime: e vandalisms, e-riots, … etc. leading to an insecure workplace

Remark

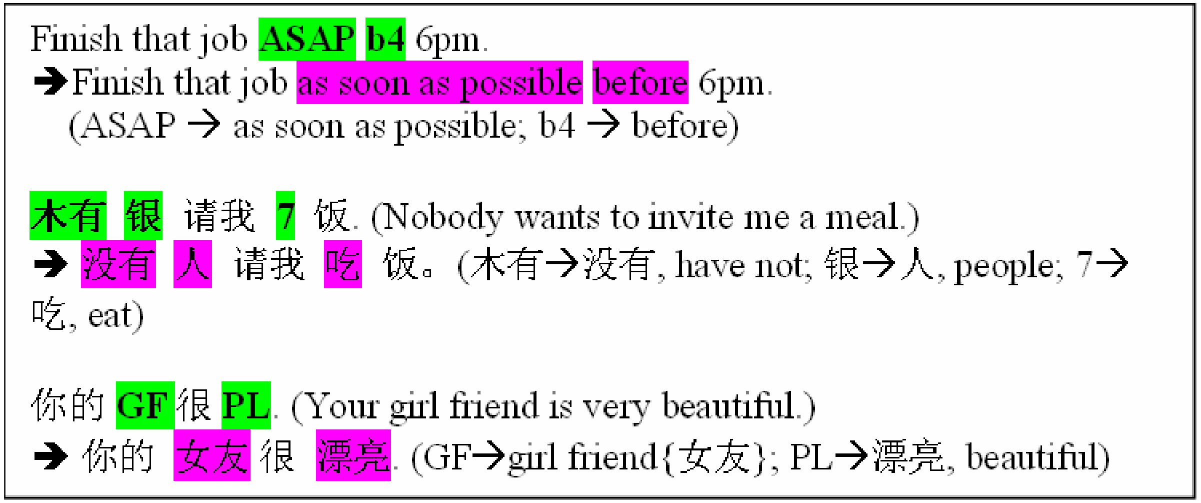
The Web will be our future e-society. As educators we have the obligation to nurture creative and constructive as well as responsible and ethical e-citizens.

Part II: Network Informal Language

1. Background
2. Related Works
3. Source Channel Model
4. Phonetic Mapping Model
5. Extended Source Channel Model
6. Evaluation
7. Conclusion

1. BACKGROUND

- What does chat text look like? – Typical examples.



Finish that job ASAP b4 6pm.
➔Finish that job as soon as possible before 6pm.
(ASAP → as soon as possible; b4 → before)

木有 银 请我 7 饭. (Nobody wants to invite me a meal.)
➔ 没有 人 请我 吃 饭。(木有→没有, have not; 银→人, people; 7→吃, eat)

你的 GF 很 PL. (Your girl friend is very beautiful.)
➔ 你的 女友 很 漂亮. (GF→girl friend{女友}; PL→漂亮, beautiful)

- Where is chat text found? – Sources.
  - Online chat rooms (in most ISP websites)
  - P2P chat tools (ICQ/QQ, MSN, etc)
  - Online BBS forums

1. Background (cont'd)

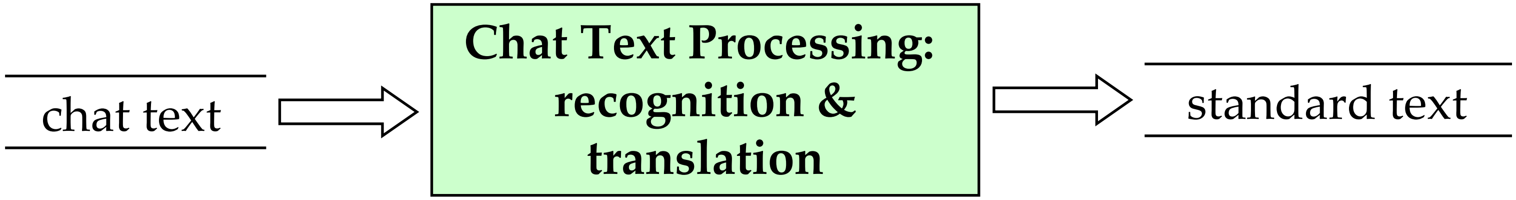
- Why is investigation of chat text worthwhile?
  - Chat text is found daily and in huge volume.
  - Important information is hidden within chat text in
    - CRM chat records: customer concentration
    - Online education log: student learning habit
    - Security in online chat room: pornography, crime and terrorism.

- Traditional NLP tools are ineffective

- Challenges: Chat text is anomalous and dynamic in nature

1. Background (cont'd)

- What is the objective of this work?
  - recognize chat language terms and
  - translate them to standard language words



chat text

Chat Text Processing: recognition & translation

standard text

- How are the problems addressed? – Our proposal
  - The dynamic character mappings are replaced by the stable phonetic mappings.
  - The stable phonetic mapping models are constructed with standard language corpus.
  - The phonetic mapping models are inserted into source channel model to achieve robustness in chat term normalization.

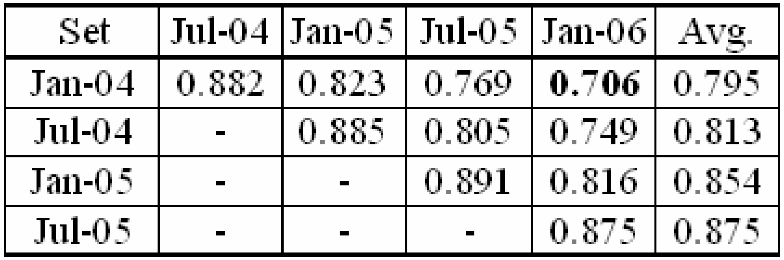
2. RELATED WORKS

- The **anomalous nature** of Chinese chat language is investigated in (Xia et al., 2005) . Two types of anomalies are mentioned:
  - Anomalous entries to standard dictionaries.
    - “介里(*jie4 li3*)” → “这里(here, *zhe4 li3*)”
    - “介里” is not a standard word.
  - Anomalous meanings of entries to standard dictionaries.
    - “偶(even, *ou3*)” → “我(me, *wo2*)”
    - “偶” is a standard word used to describe even numbers.

  - Problems caused:
    - New dictionary is required to provide knowledge for these terms. Or corpus is needed by statistical techniques.
    - Ambiguity occurs when the second type of anomalies are found.

2. Related Works (cont'd)

- The **dynamic nature** is investigated in (Xia et al., 2006a).
  - Chat term re-occurring rates based on five chat term sets constructed in ever half year from Jan 2004 to Jan 2006.
  - 29.4% of chat terms changed in two years
  - 18.5% changed within one year



Set	Jul-04	Jan-05	Jul-05	Jan-06	Avg.
Jan-04	0.882	0.823	0.769	**0.706**	0.795
Jul-04	-	0.885	0.805	0.749	0.813
Jan-05	-	-	0.891	0.816	0.854
Jul-05	-	-	-	0.875	0.875

  - Problems caused:
    - Static dictionary and corpus become outdated very quickly.
    - Performance of NLP techniques drops on new chat text

2. Related Works (cont'd)

- The first chat language corpus, NIL corpus (Xia et al., 2006b), is constructed by CUHK.
  - Covering chat text in YESKY BBS system from Dec 2004 to Feb 2005.
  - 22,432 pieces of chat text.
  - 451,193 Chinese words
  - 22,648 Chinese chat terms.

3. SOURCE CHANNEL MODEL

- Original Source Channel Model

$$\hat{C} = \arg\max_{C} p(C \mid T) = \arg\max_{C} p(T \mid C)\, p(C)$$

  - Standard character string C={c_i}
  - input chat text character string T={t_i}

- Problems
  - Data sparseness problem
    - NIL corpus contains only 12,112 NIL sentences
  - Poor training effectiveness due to the dynamic nature

4. PHONETIC MAPPING MODEL

- **Assumption**
  - Chat terms are mainly formed via phonetic mappings so that - phonetic mapping models are helpful to resolve dynamic problems to most extent. Thus:
    - Every chat term and its normal counterpart can be mapped to each other via phonetic transcription, i.e. Chinese pinyin in our case. E.g.:

      [银] ← [yin2(.)ren2]→[人]

      "银" means silver while "人" means human in Chinese.

    - The phonetic mapping is probabilistic, e.g.

      [银] ← [yin2(0.537)ren2]→[人]

4. Phonetic mapping model (cont'd)

- **Formalism**
  - phonetic mapping model (PMM)

    $$< t, c, pt(t), pt(c), p_{pm} >$$

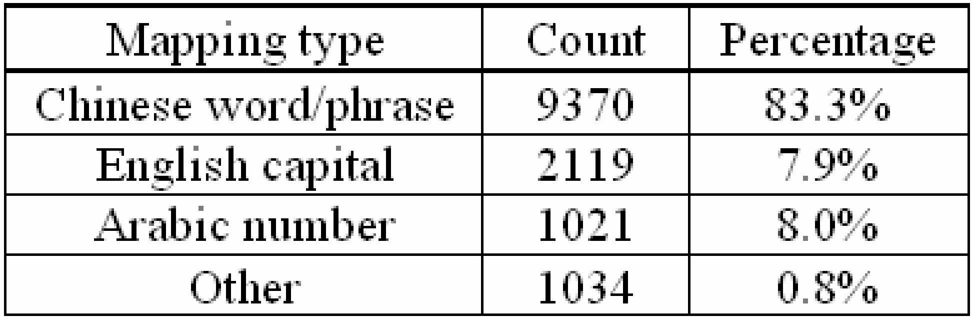
    - *t* denotes character in chat terms and *c* the corresponding character in standard word. Any character in Chinese can be *t* or *c* provided that they are phonetically similar.
    - *pt(t)* and *pt(c)* denote phonetic transcription of *t* and *c* respectively.
    - p_{pm} denotes phonetic mapping probability.
  - character mapping model (CMM)

    $$< t, c, p_{cm} >$$

    - p_{cm} denotes character mapping probability.
- **Comparison**
  - The character mapping model is constructed with **chat language corpus** and hence changes quickly
  - The phonetic mapping model is constructed with **standard language corpus** and therefore relatively stable.

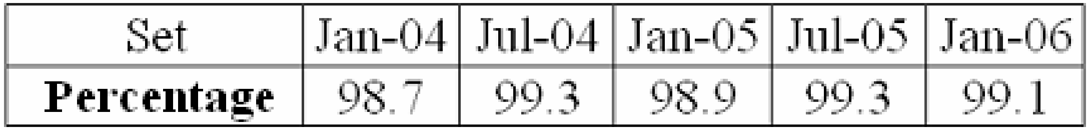
4. Phonetic mapping model (cont’d)

- Justification I: 99.2% chat terms are created via phonetic mappings



Mapping type	Count	Percentage
Chinese word/phrase	9370	83.3%
English capital	2119	7.9%
Arabic number	1021	8.0%
Other	1034	0.8%

- Justification II: % of phonetic mappings in each set covered by the standard set constructed with CNGIGA remains stable.



Set	Jan-04	Jul-04	Jan-05	Jul-05	Jan-06
Percentage	98.7	99.3	98.9	99.3	99.1

4. Phonetic mapping models (cont'd)

- Parameter estimation
  - **Phonetic mapping probability,** e.g. p_{pm}, between two characters a and a^*

$$p_{pm}(a,a^*) = \frac{\left(fr_{slc}(a^*)\times ps(a,a^*)\right)}{\sum_i \left(fr_{slc}(a_i)\times ps(a,a_i)\right)}$$

  - $\{a_i\}$ is the character set each of which is similar to character *a* in terms of phonetic transcription.
  - $ps(a, a^*)$ denotes phonetic similarity and $fr_{slc}(a^*)$ character frequency.

  - **Phonetic similarity** is product of similarities between corresponding initials and finals.

$$\begin{aligned} ps(A, A^*) &= Sim(py(A), py(A^*)) \\ &= Sim(initial(py(A)), initial(py(A^*))) \\ &\quad \times Sim(final(py(A)), final(py(A^*))) \end{aligned}$$

  - py – Chinese pinyin
  - initial – *shengmu of* Chinese pinyin.
  - final – *yunmu of* Chinese pinyin.

5. EXTENDED SOURCE CHANNEL MODEL

- The extended source channel model

$$\hat{C} = \arg\max_{M,C} p(T, M \mid C)\, p(C) = \arg\max_{M,C} p(T \mid M, C)\, p(M \mid C)\, p(C)$$

  - M denotes phonetic mapping models
- chat term normalization model <u>*p(T|M,C)*</u>

$$p(T \mid M, C) = \prod_i p(t_i \mid m_i, c_i)$$

- phonetic mapping model <u>*p(M|C)*</u>

$$p(M \mid C) = \prod_i p_{pm}(t_i, c_i)$$

- chat language model <u>*p(C)*</u>

6. EVALUATION

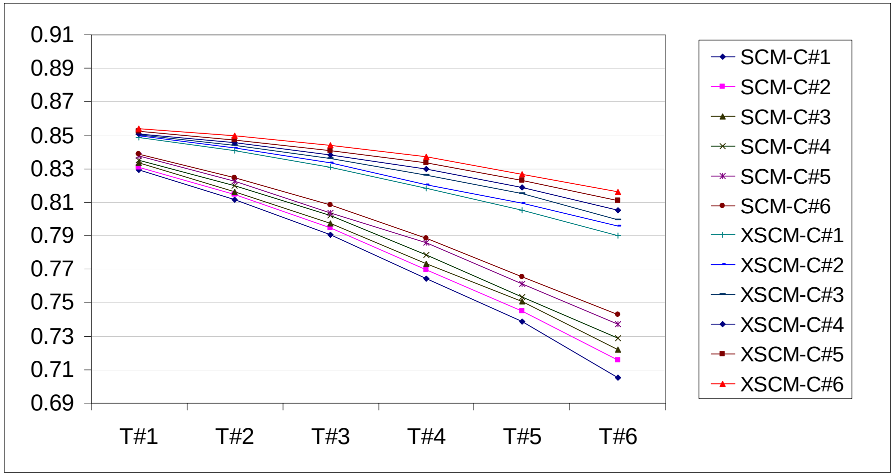
- Training sets
  - Standard Chinese corpus
    - to construct phonetic mapping model
    - Xinhua News Agency in LDC Chinese Gigaword v.2 (CNGIGA)
  - Chat language corpus
    - to construct the chat term normalization model and chat language model
    - NIL corpus (Xia et al., 2006b)
    - size-varying chat language corpus C#1(6056) ~ C#6(12,113)

- Test sets
  - 6 time-varying test sets, T#1 ~ T#6, comprising of monthly texts, 8/05–1/06
  - Normalized sentences are created by hand.

- Evaluation criteria
  - Recognition: precision, recall, f-1 measure
  - Normalization: accuracy

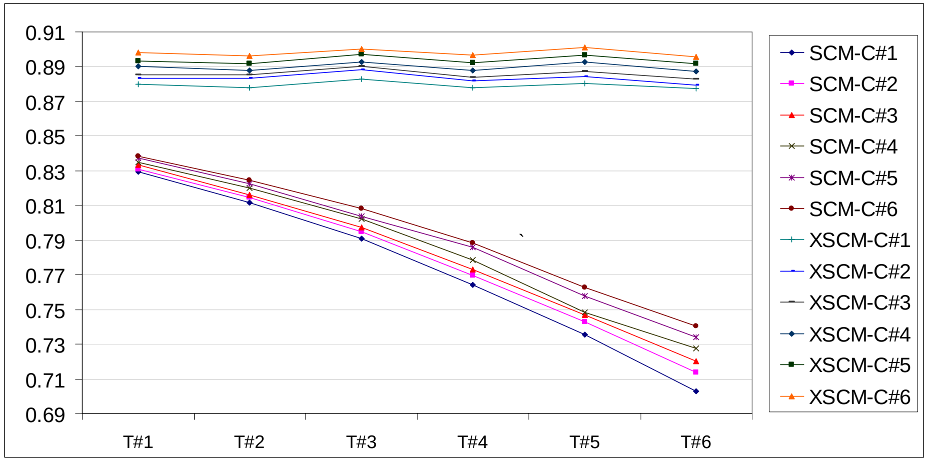
Experiment I: SCM vs. XSCM Using C#1 ~ C#6

- **Training:** using C#1 ~ C#6 as both standard corpus and chat language corpus.
- **Testing:** using test sets T#1 ~ T#6
- **Observation 1**: f-1 measure in both methods drops on time-varying test sets
- **Observation 2**: f-1 measure of both methods on same test sets drops when trained with size-varying chat language corpora
- **Observation 3**: f-1 measure gaps between SCM and XSCM becomes bigger when test set becomes newer



Experiment II: SCM vs. XSCM Using C#1 ~ C#6 and CNGIGA

- **Training:** using CNGIGA as standard corpus and C#1 ~ C#6 as chat language corpus
- **Testing**: using test sets T#1 ~ T#6
- **Observation 1**: f-1 measure of SCM drops on time-varying test sets, but XSCM trained using CNGIGA and same training chat language corpora was rather consistent
- **Observation 2**: on the same test sets, both methods produced best result with C#6, i.e. the biggest training chat language corpus



7. CONCLUSION

Error Analysis

- **Err.1** Ambiguous chat terms
  - Example-1: 我还是8米
  - ==> "我还是不明 (I still don't understand)"
  - 'eight meters' or 'don't understand' ?

- **Err.2** Chat terms created in manners other than phonetic mapping
  - Example-2: 忧虑ing
  - ==> "(正在)忧虑 (worrying)"
  - English phenomena: ing

  - Example-3: 表害怕
  - ==> "不要害怕(Don't be afraid)"
  - multiple mapping:表→不要 (do not)

7. Conclusion (cont'd)

Contribution:

We propose a new phonetic-based translation method for handling chat terms. We show that:

- XSCM outperforms SCM with same training data, this proves phonetic mapping models work.
- XSCM produces higher performance consistently on time-varying test sets
- both SCM and XSCM perform best with biggest training chat language corpus

Thank you ☺

Contact me:

Email: kfwong@se.cuhk.edu.hk

Homepage: http://www.se.cuhk.edu.hk/~kfwong