Chinese Informal Word Normalization: an Experimental Study

Aobo Wang[1], Min-Yen Kan[1,2]

[1]Web IR / NLP Group (WING)
[2]Interactive and Digital Media Institute (IDMI)

Daniel Andrade[3], Takashi Onishi[3] and Kai Ishikawa[3]

[3]Knowledge Discovery Research Laboratories
NEC Corporation, Nara, Japan

Introduction

- **Informal words in microtext**

Twitter @xxx

"The song is koo, doesnt really showcase anyones talent though."

koo	➔	cool
doesnt	➔	doesn't
anyones	➔	anyone's

Weibo @vvv

"排n久连硬座都木有了"

排	?	排队 [queue]
n久		很久 [long time]
木有		没有 [no]

- –Normalization is an important pre-processing step
- –Benefit downstream applications
  - ➢*e.g.*, translation, semantic parsing, word sense disambiguation

Outline

- Introduction
- Data Analysis
  - Data Annotation
  - Chanels & Motivations
- Related Work
- Methodology
- Experiment Result
- Conclusion

- **Data Set Preparation**
  - Crawling data from Sina Weibo
    - PrEV (Cui et al., 2012)
  - Crowdsourcing annotations using *Zhubajie*
    - informal words
    - normalization
    - sentiment
    - motivation
  - 1036 unique informal–formal pairs with informal contexts

- **Major Channels of Informal Words**

Channel (%)	Informal to formal	Translation
Phonetic Substitutions (63)	河蟹 (he2 xie4) ➔ (he2 xie2) 和谐 木有 (mu4 you3) ➔ (mei2 you3) 没有 bs ➔ (bi3 shi4) 鄙视	harmonious no despise
Abbreviation (19)	手游 ➔ 手机 游戏 网商 ➔ 网络 商城	mobile game online shopping mall
Paraphrase (12)	萌 ➔ 可爱 暴汗 ➔ 非常 尴尬	cute very embarrassed

- **Motivation of informal Words**

Motivation	%	Example
To avoid (politically) **sensitive** words	17.8	“财产**公式**是一种态度” [property formula indicates the attitude] 公式 [formula] ➔ （gong1 shi4）➔公示 [publicity] “财产公示是一种态度” [property publicity indicates the attitude]
To be **humorous**	29.2	鸭梨 [pear] ➔ （ya1 li2）➔ （ya1 li4） 压力 [pressure]
To hedge criticism using **euphemisms**	12.1	bs ➔ (bi3 shi4) 鄙视 [despise]
To be **terse**	25.4	剧透 ➔ 剧情 透露 [tell the spoilers]
To **exaggerate** the posts’ mood	10.5	暴汗 ➔ 非常 尴尬 [very embarrassed]
Others	5.0	乘早 ➔ 趁早 [as soon as possible]

Outline

- Introduction
- Data Analysis
- **Related Work**
  - –Li and Yarowsky (2008)
  - –Xia et al. (2008)
- **Methodology**
- **Experiment Result**
- **Conclusion**

- Li and Yarowsky (2008)
  - –Mining informal-formal pairs from the web blog
    - ➢Query: “GF 网络语言” [internet language]
      ⬇ Search Engine
    - ➢Definition: “GF是女朋友的意思” [GF refers to Girl Friend]

  - o Assume the formal and informal equivalents co-occur nearby
  - o Works for highly frequent and well defined words.
  - o Relies on the quality of search engine

  - – Our goal
  - ✓ Relax the strong assumption
  - ✓ React to the **evolution** of informal words

- Xia et al. (2008)
  - –Normalize informal words from chats
    - ➢Extend source-channel model with phonetic mapping rules

  - o Only deal with the *Phonetic Substitutions* channel
  - o Manually weighting similarity is time-consuming but inaccurate

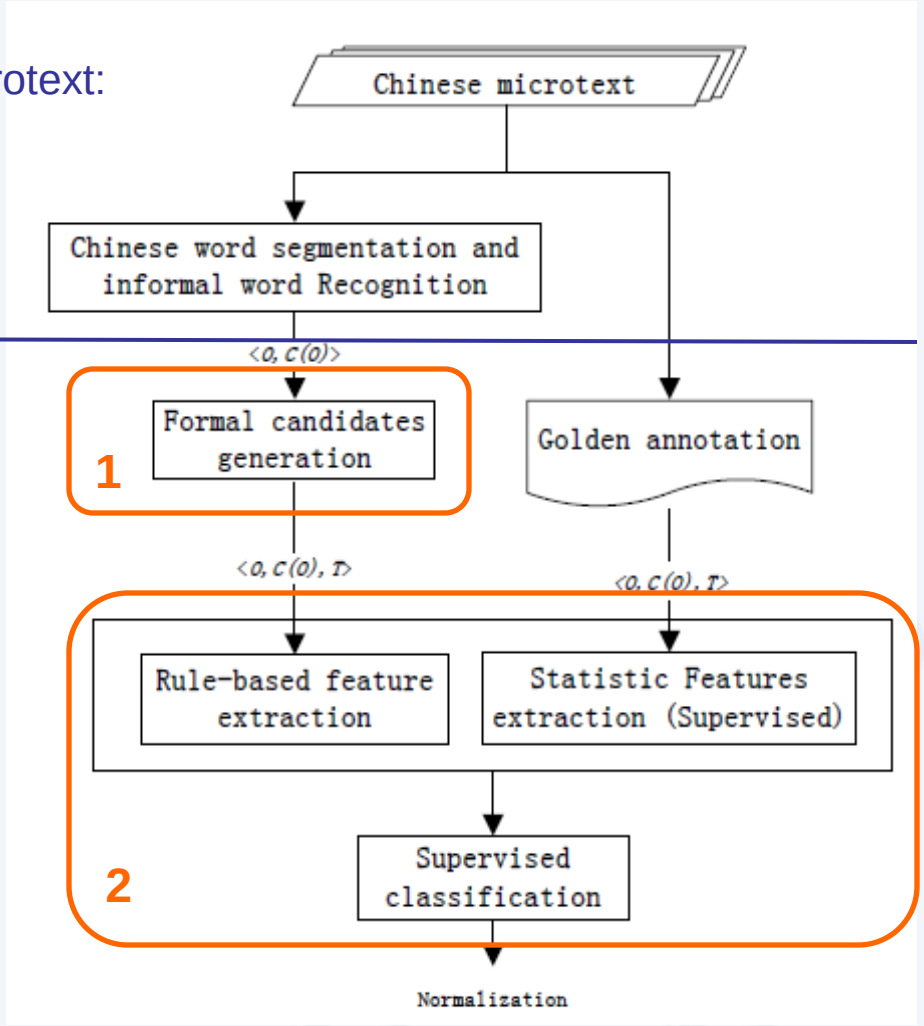
  - – Our Goal
  - ✓ Deal with three major channels
  - ✓ Learn the similarity automatically

Outline

- Introduction
- Data Analysis
- Related Work
- **Methodology**
  - –Candidates generation
  - –Candidates classification
- **Experiment Result**
- **Conclusion**

- **Pre-processing**
  - –Mining Informal Language from Chinese Microtext: Joint Word Recognition and Segmentation
  - –Wang and Kan, ACL 2013

- **Normalization**
  - ***O**:* observed informal words
  - ***C(O)**:* context of the informal words
  - ***T***: target formal candidates



- **Step 1: Candidate Generation**
  - The informal word and its formal equivalents share similar contextual collocations.

➢ … 建设 **河蟹** 社会 … **O**bservation
➢ … 建设 **和谐** 社会 … **T**arget
[build the **harmonious** society]

–Search for formal candidates from Google Web1T Corpus
➢Generate lexicon patterns from context (***C(O)***)

➢Use patterns as queries to search for candidates (***T***)
➢***<O, C(O), T>***

- **Step 1: Candidate Generation**
  - *<O, C(O), T>*

… 建设	河蟹	社会 …	O	[build the harmonious society]
… 建设	和谐	社会 …	T1	
… 走向	中国	社会 …	T2	
… 建设	未来	社会 …	T3	

  - Noise filtering
    - ➢Rank the candidates by **word trigram probability**
    - ➢Keep the top **N=1000 candidates**

Channel	Loss Rate (%)
Phonetic Substitution	14
Abbreviation	15
Paraphrase	17

- **Step 2: Candidates Classification**

  –**Feature Extraction** *F (<O, C(O), T>)*

Rule-based
O contains valid Pinyin script
O contains digits
O is a potential Pinyin acronym
T contains characters in *O*
The percentage of characters common be- tween *O* and *T*

Statistical
N-Gram Probabilities
Pinyin Similarity
Lexicon and Semantic Similarity

- **Pinyin Similarity**

$$PYSim(T|O) = \prod PYSim(t_i|o_i)$$

$$PYSim(t_i|o_i)) = \mu P(py(t_i)|py(o_i)) + \lambda P(ini(t_i)|py(o_i)) + \eta P(fin(t_i)|py(o_i))$$

pinyin script of character (t) *initial part of py(t)* *initial part of py(t)*

- **Lexicon and Semantic Similarity**

$$\hat{T} = \arg\max_{T} P(T|O) = \arg\max_{T} P(O|T)P(T)$$

$$P(O|T) = \prod P(o_i|t_i)$$

–Extend the Source-Channel model with POS mapping model

$$P(O|T) = \prod P'(o_i|t_i)$$

$$P'(o_i|t_i) = \alpha P(o_i|t_i) + \beta P(o_i|pos(t_i), pos(o_i))$$

–Use synonym dictionaries to further address the data sparsity

- ➢TYC Dict – datatang.com
- ➢Cilin – HIT IR lab

Outline

- Introduction
- Motivation
- Related Work
- Methodology
- **Experiment Result**
  - –E1: Informal words Normalization
  - –E2: Formal domians synonym acquisition
- **Conclusion**

- **E1: Informal word Normalization**
  - –Data from all the channels are merged together
  - –5-fold cross validation
  - –Weka 3
  - ✓ Decision Tree performs best

Classifier	Pre	Rec	F_1
SVM	.646	.273	.383
LR	.567	.340	.430
DT (C4.5)	.886	.443	.590

Table 4: Performance comparison using different classifiers.

  - –Final loss rate 64.1%
  - –Less than 70% estimated in Li and Yarowsky (2008)

E1: Informal word Normalization

- Phonetic Substitution Channel is relatively easy
- Semantic similarity is difficult to measure

Channel	System	Pre	Rec	F_1
Phonetic Substitution	OurDT	.956	.822	.883
	LY Top1	.754	—	—
	LY Top10	.906	—	—
Abbreviation	OurDT	.807	.665	.729
	LY Top1	.118	—	—
	LY Top10	.412	—	—
Parapharse	OurDT	.754	.331	.460
	LY Top1	—	—	—
	LY Top10	—	—	—

– Loss comparison with Li and Yarowsky (2008)

E1: Informal word Normalization

- The sparsity is lessened with synonym dictionaries
- The upper-bound performance is still significantly higher

Feature set	Pre	Rec	F_1
w/o	.886	.443	.590
w	**.895**	**.583**	**.706**
w + channel	.915	.638	.752

Table 6: Performance over different feature sets. "w" ("w/o") refers to the model trained with (without) features from formal synonym dictionaries. "channel" refers to the model trained with the correct channel given as an input feature.

E2: Formal Domain Synonym Acquisition

- –Trained with Cilin and Weibo data
- –Tested with TYC Dict
- –The contexts are extracted from Chinese Wikipedia

- –Performance
  - ➢F_1 69.9%
  - ➢Precision 94.9%
  - ➢Recall 55.4%

Conclusion

- Informal words are created through three major channels with different motivations
- Propose a two-stage candidate generation-classification method for normalization
- It can also be applied to synonym acquisition task in the formal domain

Thank You