Be Appropriate and Fun:
Automatic Entity Morph Encoding

Pole, Brother Huang, The Boy, The Wanted, Kim Warrior

Authentic Text, Sunshine, Godfather, The Spy
Starring

Boliang Zhang [Pole]  Hongzhao Huang [Brother Huang]  Xiaoman Pan [The Boy]

Heng Ji [The Wanted]  Kevin Knight [Kim Warrior]  Zhen Wen [Authentic Text]

Yizhou Sun [Sunshine]  Jiawei Han [Godfather]  Bulent Yener [The Spy]
The Secret Weapon: “Morphing”
Morphs by Intentions

(1) Avoid censorship 7%

(2) Express strong sentiments, sarcasm, emotions 16%

(3) Be humorous or make descriptions more vivid 27%

Mixture of (1) and (2) 12%

Mixture of (1) and (3) 9%

Mixture of (2) and (3) 5%

Others 24%

Tender Beef Pentagon = Yang Mi
Morphs by Encoding Methods

- M1: Phonetic Substitution - 13%
- M2: Spelling Decomposition - 1%
- M3: Nickname Generation - 12%
- M4: Translation and Transliteration - 3%
- M5: Semantic Interpretation - 20%
- M6: Historical Figure Mapping - 4%
- M7: Characteristics Modeling - 21%
- M8: Reputation and public perception - 26%

Examples:
- Chairman Mao
- First emperor
- Bird
- Larry Bird
M1: Phonetic Substitution

比尔盖茨 (Bill Gates)

[Bi Er Gai Ci]

盖子 (Lid)

[Bi Er Gai Zi]

- Replace the phonetically similar part of the entity name
- Prefer candidates including more negative words (derived from HowNet (Dong and Dong, 1999)) or rare words (Valitutti et al., 2013)
M2: Spelling Decomposition

- Decompose complex character to simple radicals.

M3: Nickname Generation

- In baby talk, parents give kids lovely nick name by repeating the last character of the name.
M4: Translation & Transliteration

布什 (Bush) → Bush → bush → 灌木 (shrub)

M5: Semantic Interpretation

金日成 (Kim Il-sung) → 日 → 太阳 (sun) → 金太阳 (Kim Sun)

Chinese character dictionary

• Interpret one character of the entity name based on Xinhua character dictionary.
M6: Historical/Fictional Figure Mapping

- They both governed the west of China and started a rebellion and were defeated at last.

- Collected 38 famous historical figures and their descriptions. Applied morph resolution approach (Huang et al., 2013) to rank candidates based on semantic contexts.
We compute the semantic relationship between the query entity and each word from a positive and negative words corpora by using word2vec (Mikolov et al., 2013).

- We compute the semantic relationship between the query entity and each word from a positive and negative words corpora by using word2vec (Mikolov et al., 2013).
M8: Reputation & Public Perception

苏亚雷斯(Suarez)

苏牙(Sua-tooth)
Data and Evaluation

• Data
  • 1,553,347 tweets from Sina Weibo 05/01/2013-06/30/2013

• 55 person names
  • Human created 187 morphs
  • System created 382 morphs

• Human Evaluation
  • 9 Chinese native speakers to help evaluate morphs based on Perceivability, Funniness and Appropriateness

• Automatic Evaluation
  • Use each system created morph to replace its corresponding human created morphs in tweets and form a “morphed” data set
  • Apply a morph decoder: Candidate identification based on anomaly analysis + morph resolution (Huang et al., 2013)
Translation & transliteration: system outperforms human in perceivability because system can search larger vocabulary, similar observation to (Knight and Graehl, 1998)

Only 64 human created morphs and 72 system created morphs are perceivable by all human assessors
Human Evaluation: Funniness

- Spelling Decomposition: human created morphs are much more funny
- Radicals reflect character meaning or reflect some characteristic of the entity
- The radicals are funny and vivid, express strong sentiment/sarcasm
Human Evaluation: Appropriateness

![Bar chart showing the appropriateness of human and system-generated morphs.](chart.png)
Human Evaluation: Overall

- Our system achieves 66% of the human performance
- The assessors were asked to recite the morphs after the survey: 20.4% remembered morphs are generated by our system
Automatic Evaluation

- Human morphs are discovered more easily because the decoder was trained based on human morph related features.

Resolution Acc@K (%)

- System generated morphs are more easily resolved than human generated ones because they are more implicit.
Related Work

- Our pronunciation, lexical and semantic similarity measurements were inspired from the methods to map between Chinese formal and informal words (Xia et al., 2005&2006; Li and Yarowsky, 2008; Wang et al., 2013; Wang and Kan, 2013)

- Some selection criteria were inspired from previous work on generating humors (Valitutti et al., 2013; Petrovic and Matthews, 2013)
Conclusions and Future Work

• Proposed a new problem of encoding entity morphs and developed a wide variety of novel automatic approaches

• Future Work
  • Improve the language-independent approaches based on historical figure mapping and culture and reputation modeling
  • Extend to other types of information including sensitive events, satires and metaphors to generate fable stories
  • Track morphs over time to study the evolution of Internet language
  • Online applicatio
3Q, Bricks?