Be Appropriate and Fun: Automatic Entity Morph Encoding

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

Authentic Text, Sunshine, Godfather, The Spy

Rensselaer

Starring

Boliang Zhang [**Pole**]

Hongzhao Huang [**Brother Huang**]

Xiaoman Pan [**The Boy**]

Heng Ji [**The Wanted**]

Kevin Knight [**Kim Warrior**]

Zhen Wen [**Authentic Text**]

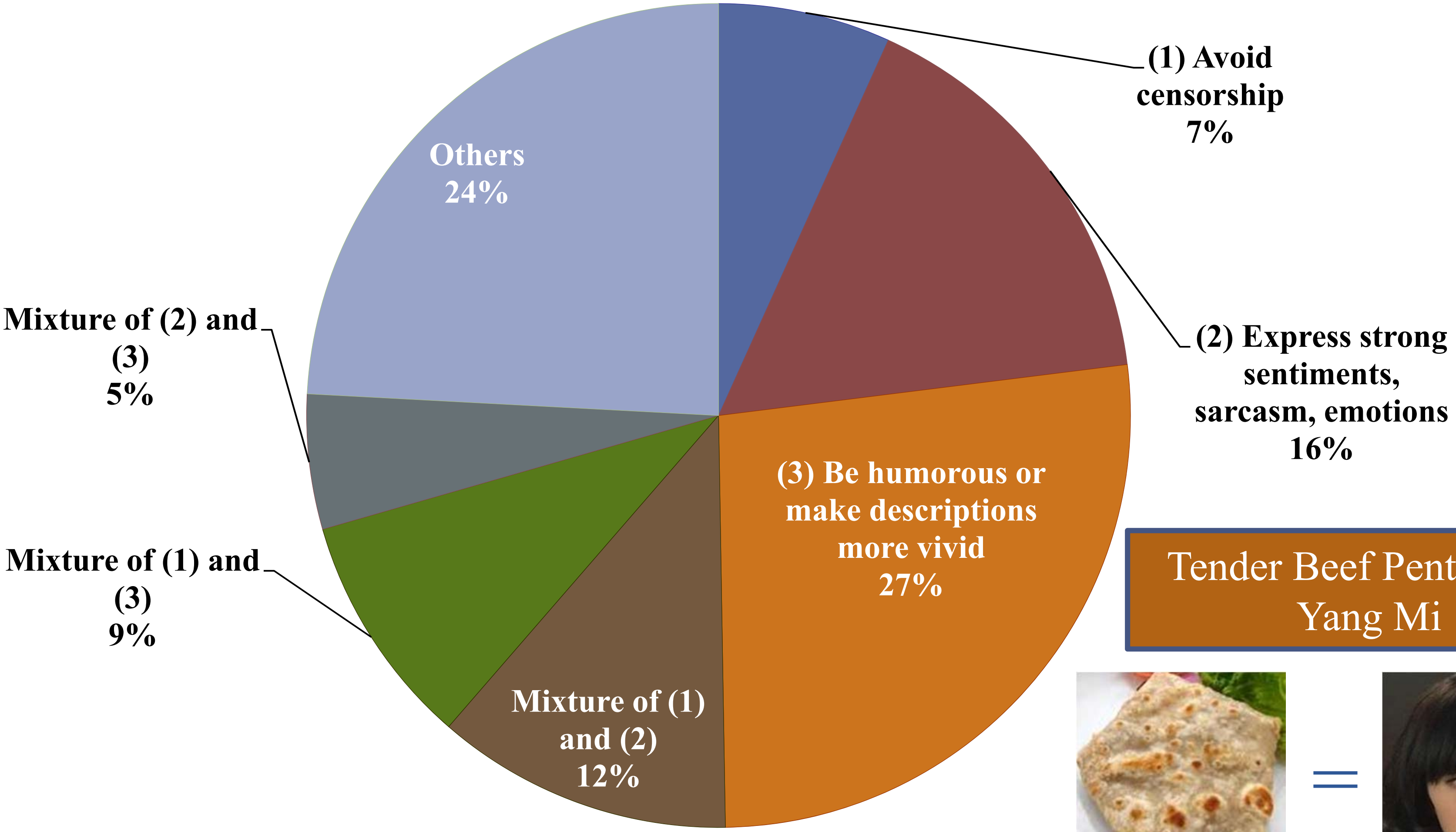
Yizhou Sun [**Sunshine**]

Jiawei Han [**Godfather**]

Bulent Yener [**The Spy**]

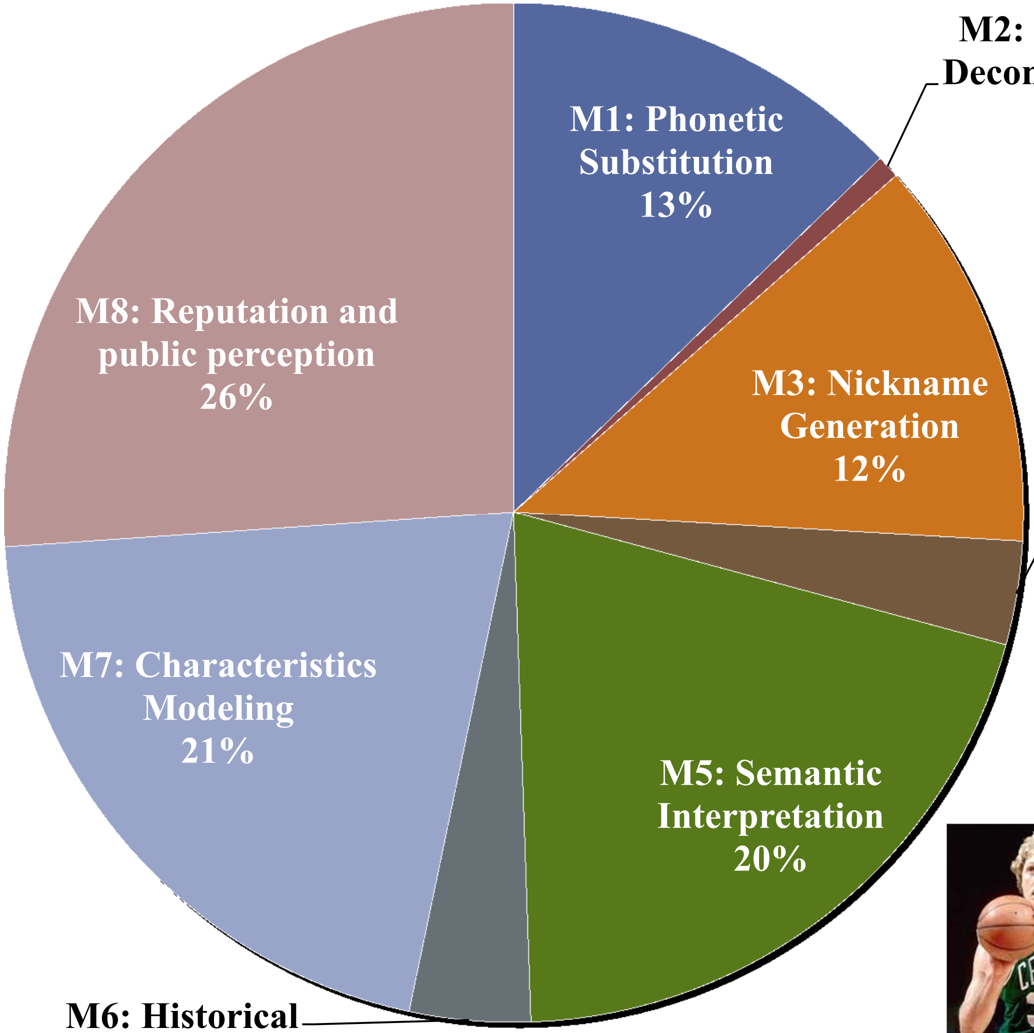
The Secret Weapon: “Morphing”

Morphs by Intentions



Tender Beef Pentagon = Yang Mi

Morphs by Encoding Methods



Chairman Mao

First emperor

Larry Bird

Bird

M1: Phonetic Substitution

比尔盖茨 (Bill Gates)

[Bi Er Gai Ci]

Gai Zi

term frequency table

less common word

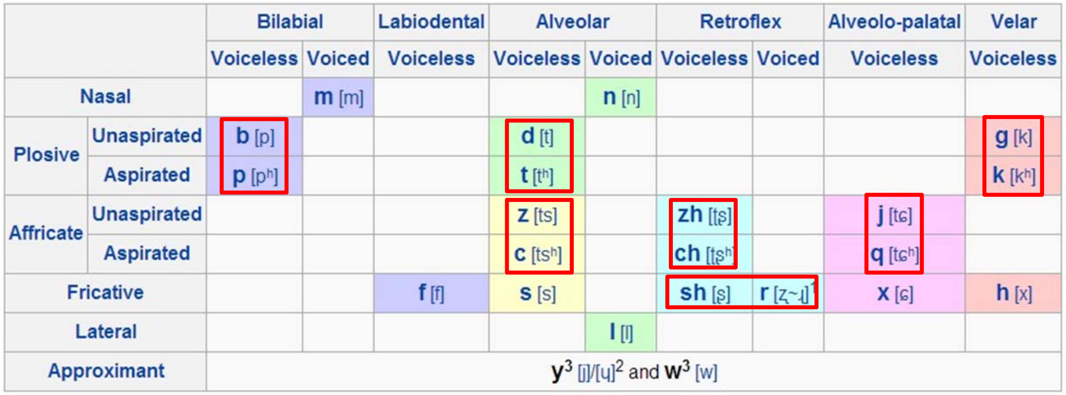
same pronunciation

盖子 (Lid)

[Gai Zi]

比尔盖子 (Bill Lid)

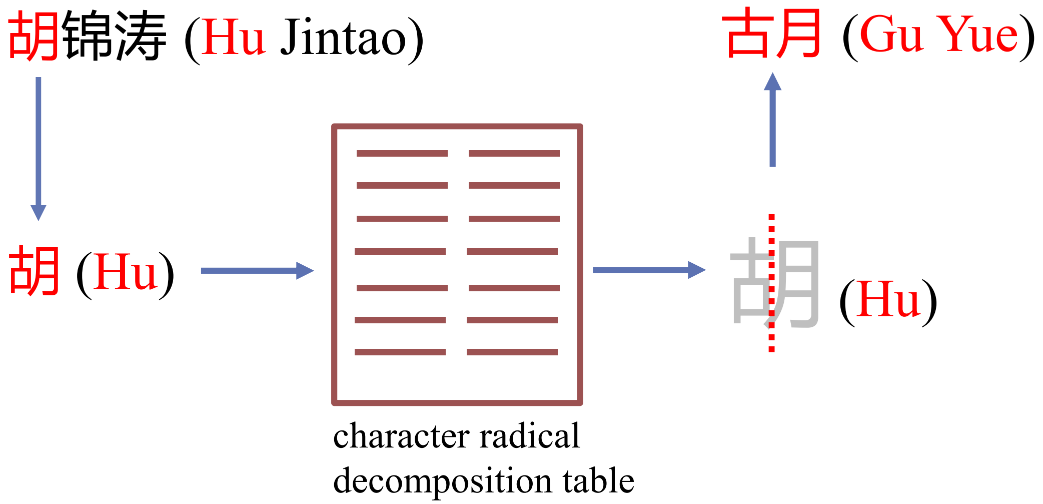
[Bi Er Gai Zi]



		Bilabial		Labiodental	Alveolar		Retroflex		Alveolo-palatal	Velar
		Voiceless	Voiced	Voiceless	Voiceless	Voiced	Voiceless	Voiced	Voiceless	Voiceless
Nasal			m [m]			n [n]				
Plosive	Unaspirated	b [p]			d [t]					g [k]
	Aspirated	p [pʰ]			t [tʰ]					k [kʰ]
Affricate	Unaspirated				z [ts]		zh [ʈʂ]		j [tɕ]	
	Aspirated				c [tsʰ]		ch [ʈʂʰ]		q [tɕʰ]	
Fricative				f [f]	s [s]		sh [ʂ]	r [ʐ~ɻ][1]	x [ɕ]	h [x]
Lateral						l [l]				
Approximant		y[3] [j]/[ɥ][2] and w[3] [w]								

- 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



胡锦涛 (Hu Jintao)

胡 (Hu)

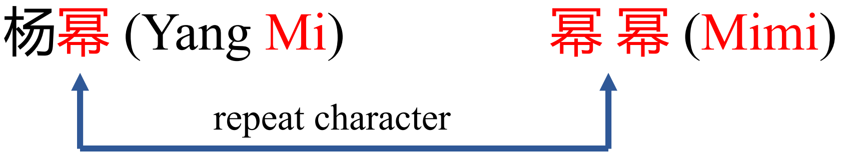
character radical decomposition table

胡 (Hu)

古月 (Gu Yue)

- Decompose complex character to simple radicals.

M3: Nickname Generation



杨幂 (Yang Mi)

repeat character

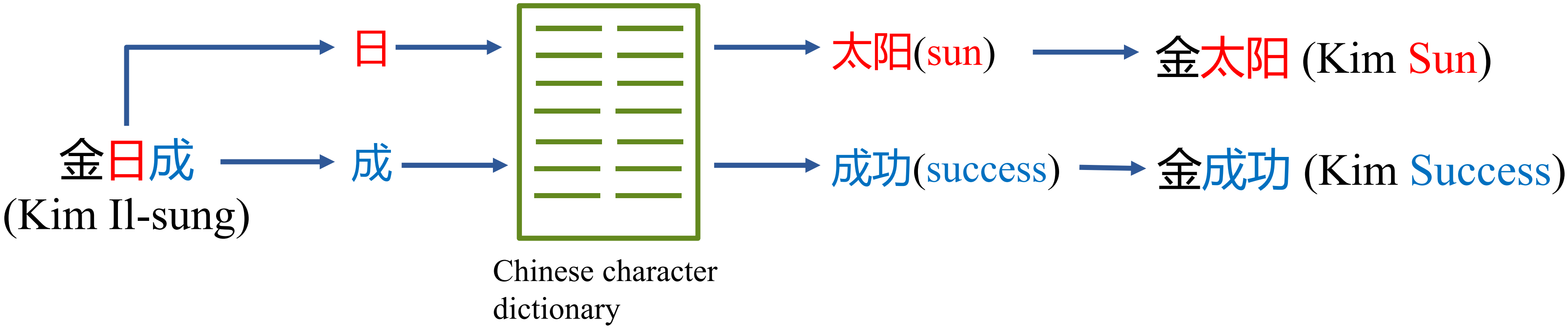
幂 幂 (Mimi)

- In baby talk, parents give kids lovely nick name by repeating the last character of the name.

M4: Translation & Transliteration

布什 (Bush) → Translated to English → Bush → bush → Translated back to Chinese → 灌木 (shrub)

M5: Semantic Interpretation



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

金日成 (Kim Il-sung) → 成 → Chinese character dictionary → 成功(success) → 金成功 (Kim Success)

- Interprete one character of the entity name based on Xinhua character dictionary.

M6: Historical/Fictional Figure Mapping

薄熙来 → 平西王

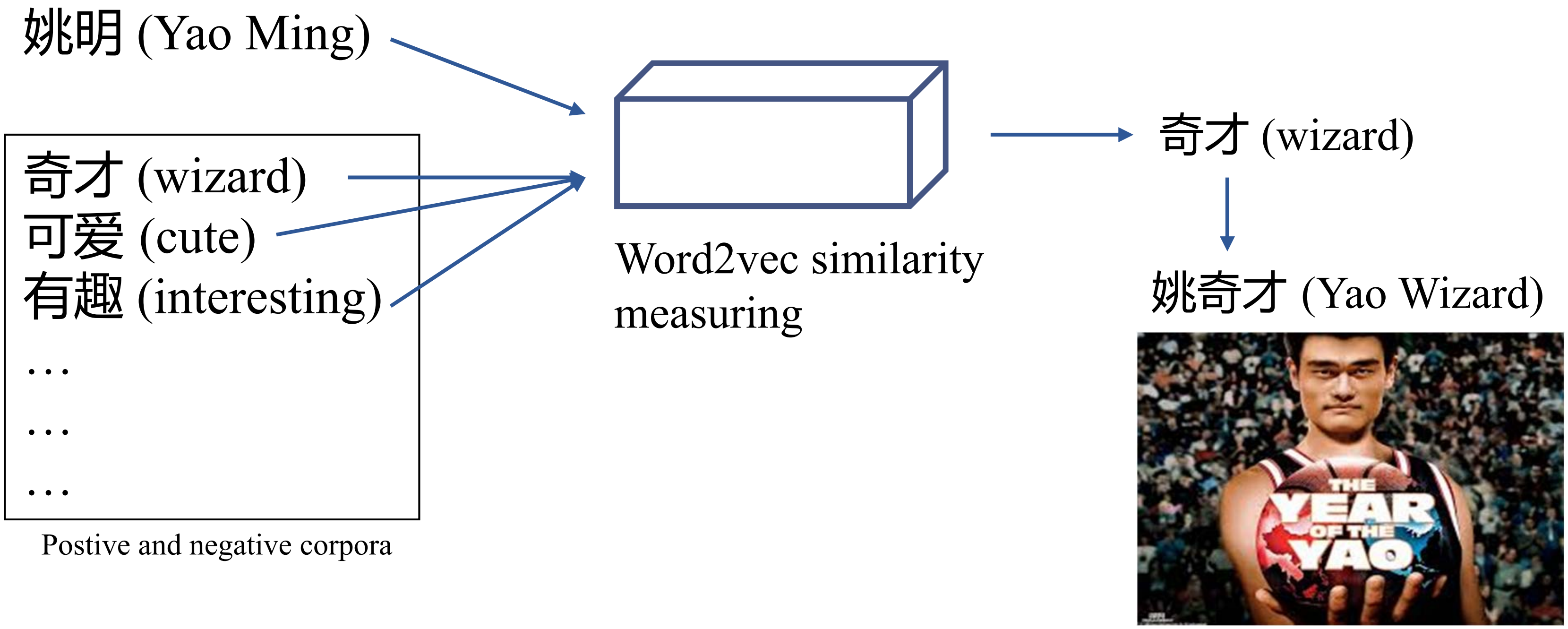
(Bo Xilai) (Conquer West King)

Chris Christie → the Hutt

- 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.

M7: Characteristics Modeling



姚明 (Yao Ming)

奇才 (wizard)
可爱 (cute)
有趣 (interesting)
…
…
…

Postive and negative corpora

Word2vec similarity measuring

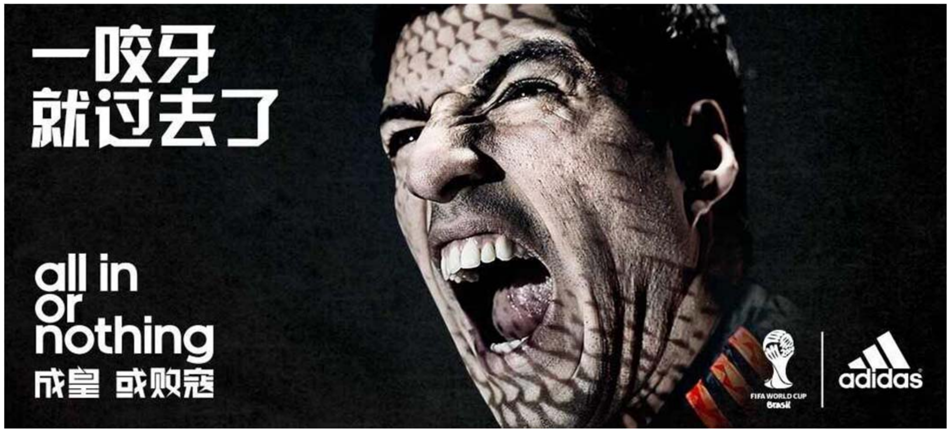
奇才 (wizard)

姚奇才 (Yao Wizard)

- 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).

- 金正恩 (Kim Jong-un) → 金胖子 (Kim Fat)

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

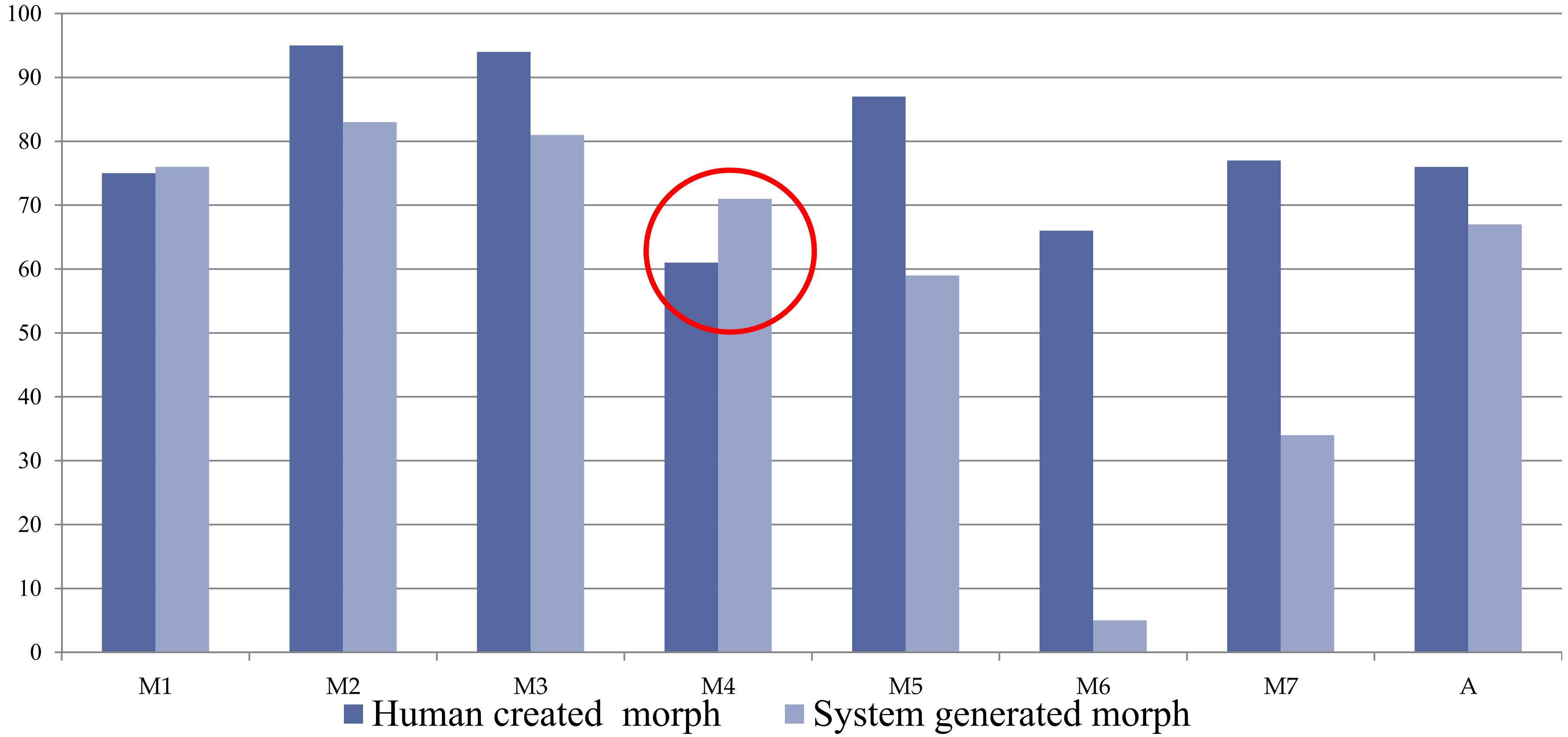
- 55 person names
    - Human created 187 morphs
    - System created 382 morphs

- 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)

Human Evaluation: Perceivability

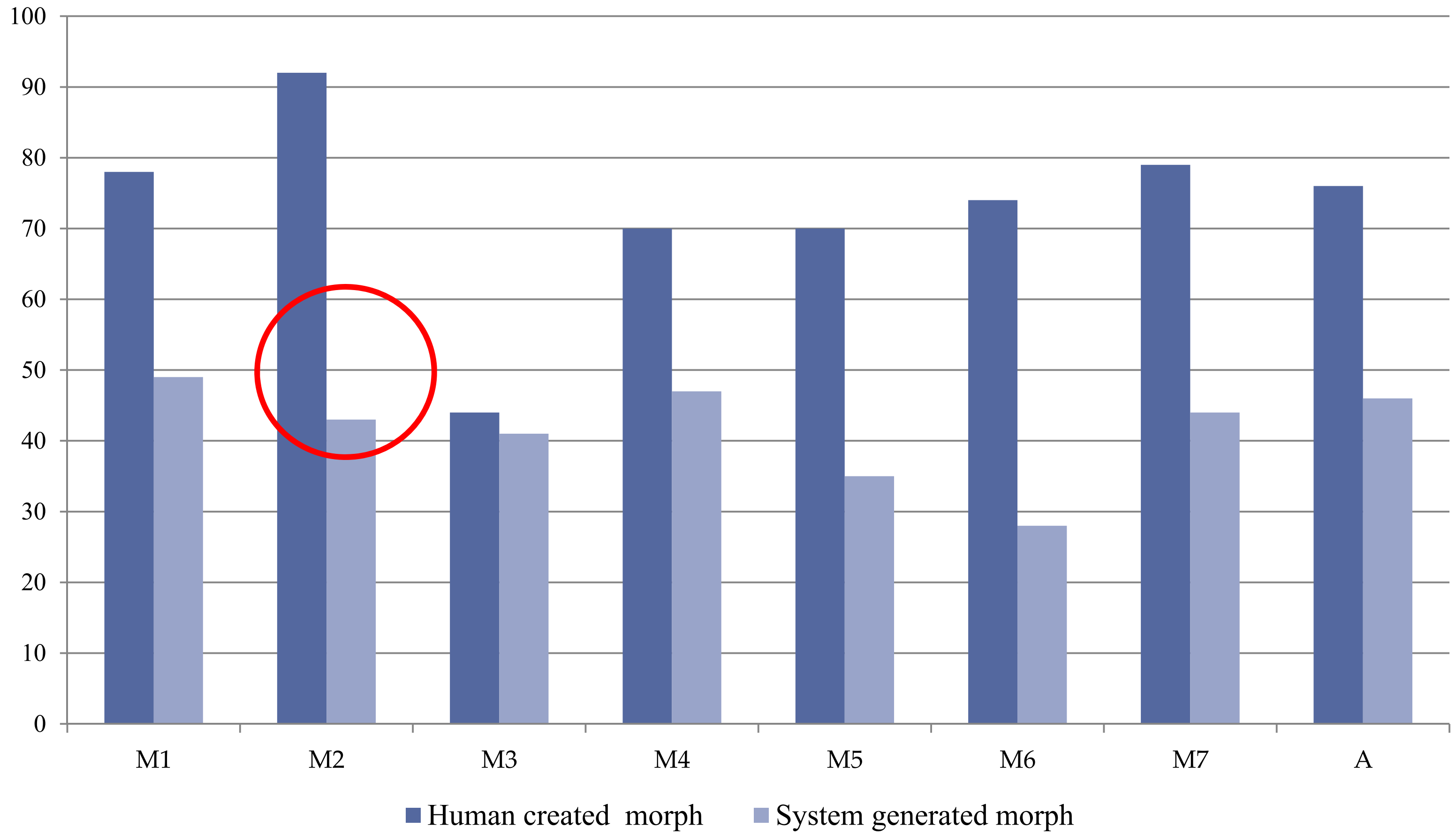
Rensselaer



- 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

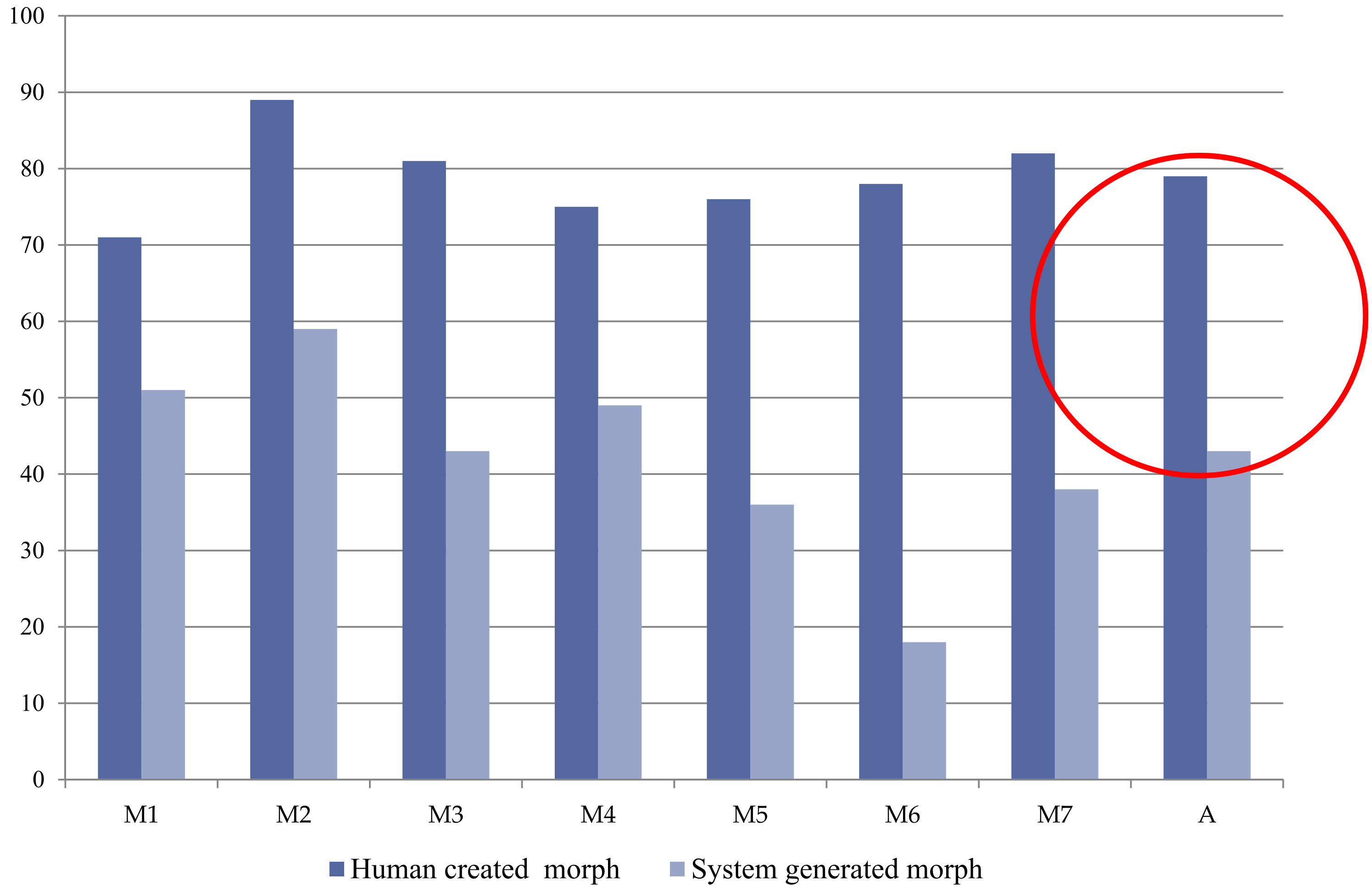
Rensselaer



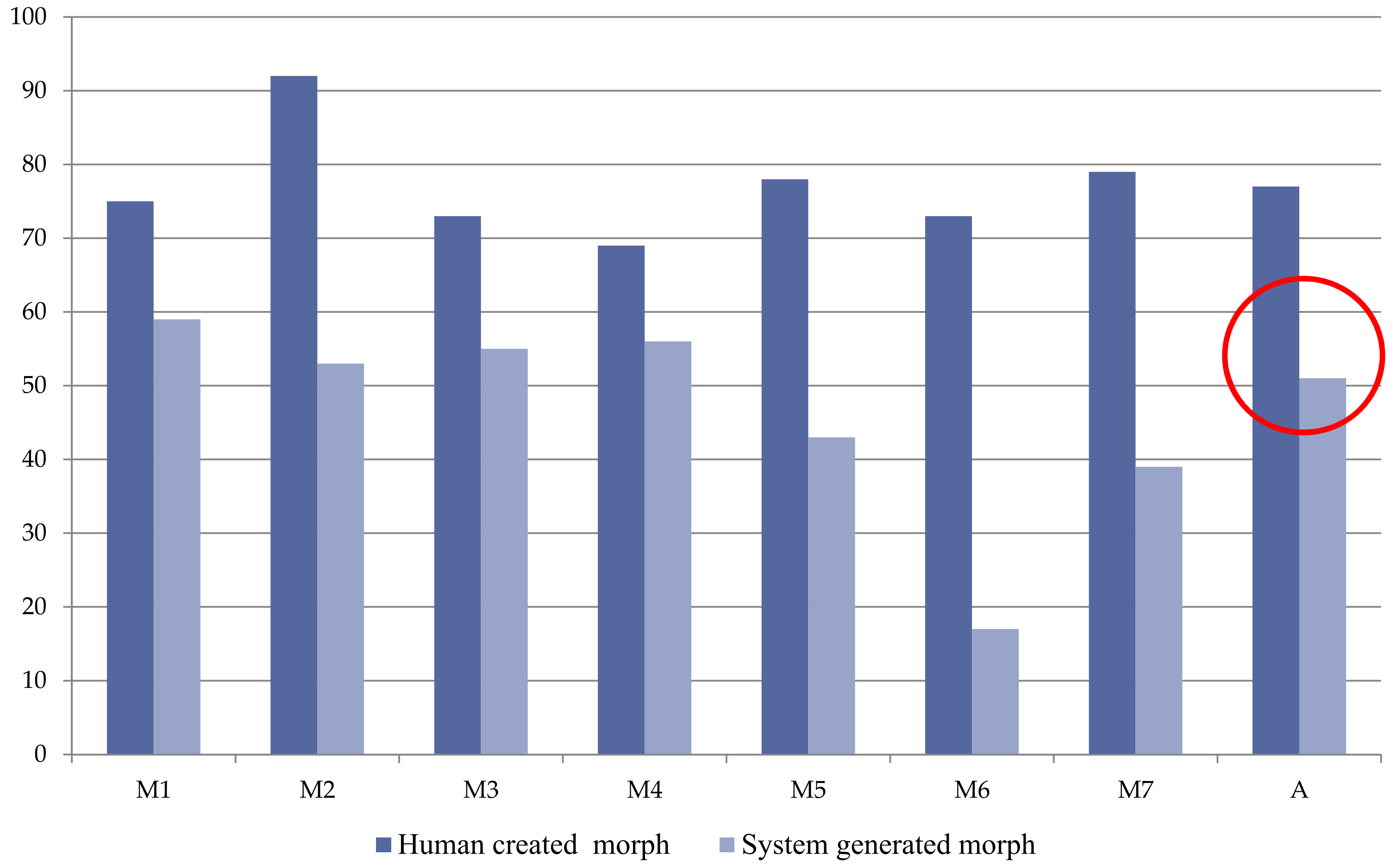
- Spelling Decomposition: human created morphs are much more funny
- Radicals reflect character meaning or reflect some characteristic of the entity
- The radicals are funny and vivid, express strong sentiment/sarcasm

Human Evaluation: Appropriateness

Rensselaer

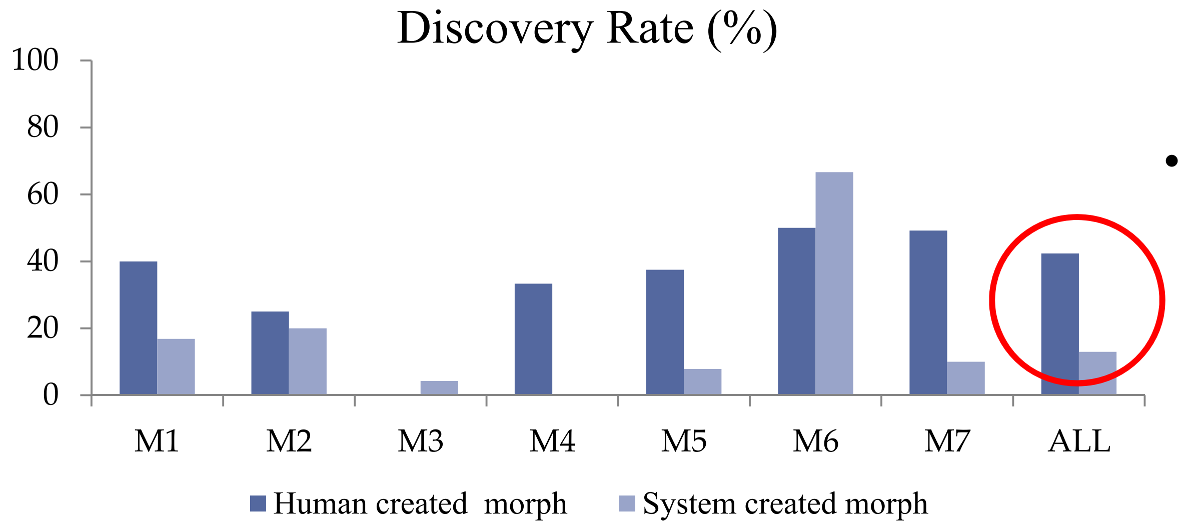


Human Evaluation: Overall

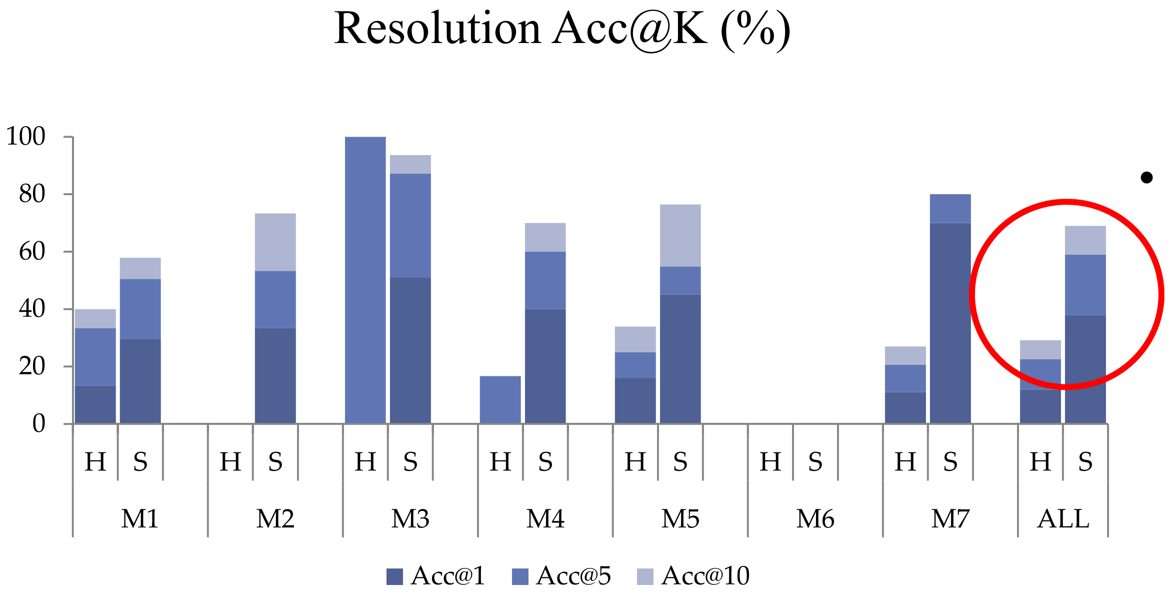


- Our system achieves 66% of the human performance
- The assessors were asked to recite the morphs after the survey: 20.4% remembered morphs are generated by our system

Automatic Evaluation



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



- 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?