NLP for Social Media

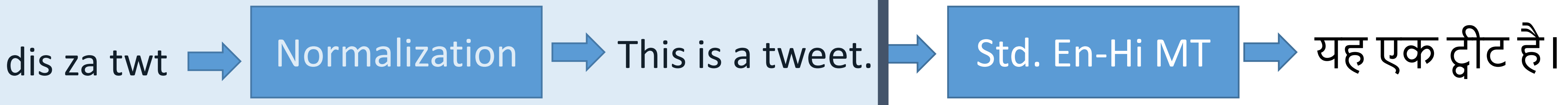
Lecture 2: Text Normalization

Monojit Choudhury

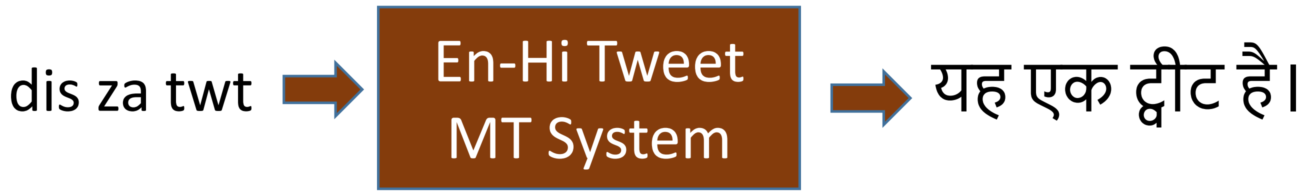
Microsoft Research Lab, *monojitc@microsoft.com*

Two Approaches for SMD Processing

- Normalization



- Systems/techniques specifically built for SMD.



What will we learn?

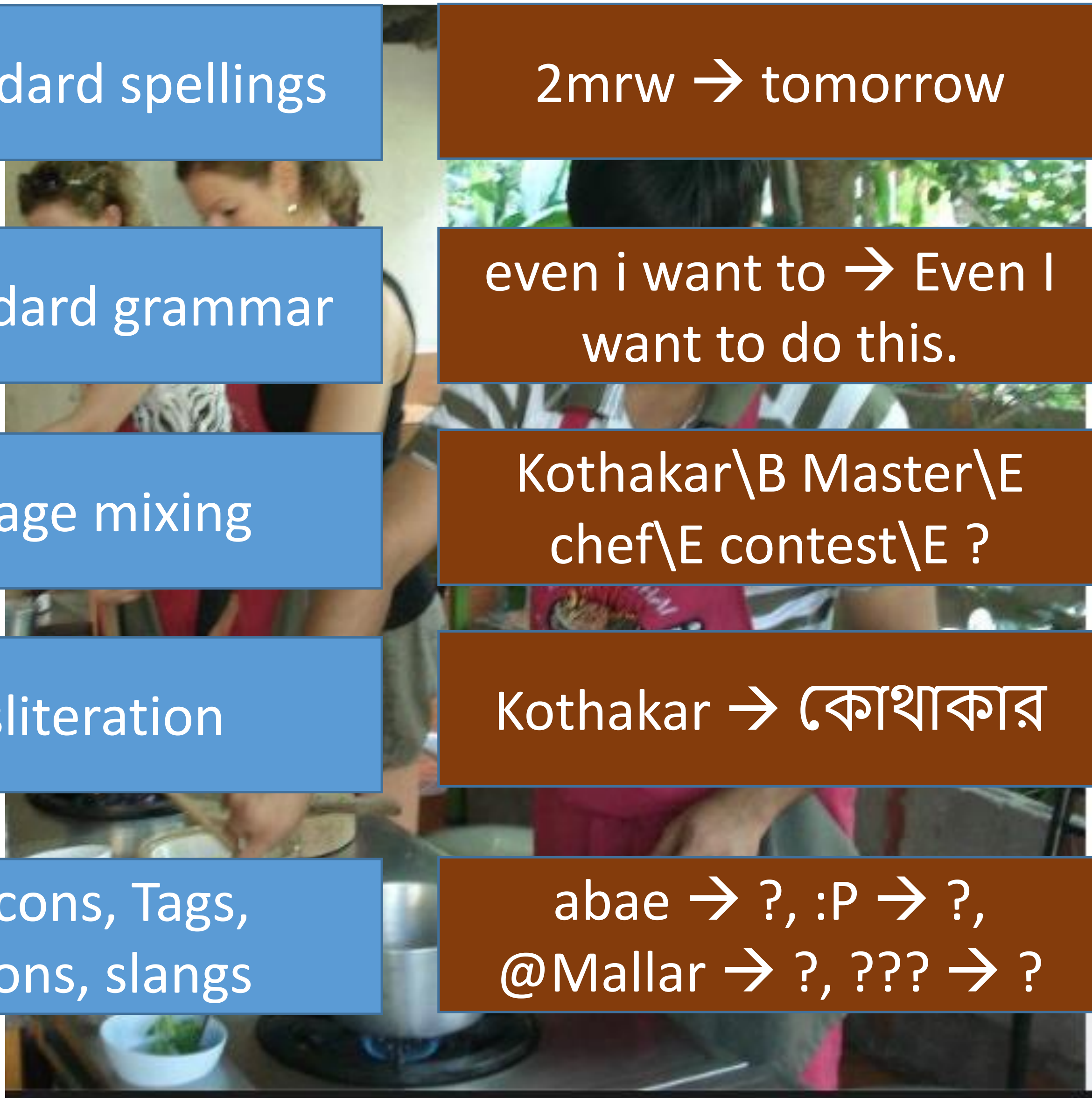
- What does “normalization” entail?
- Unintentional Spelling changes & Edit Distance
- Intentional spelling changes
- Patterns of intentional spelling changes

What will we learn?

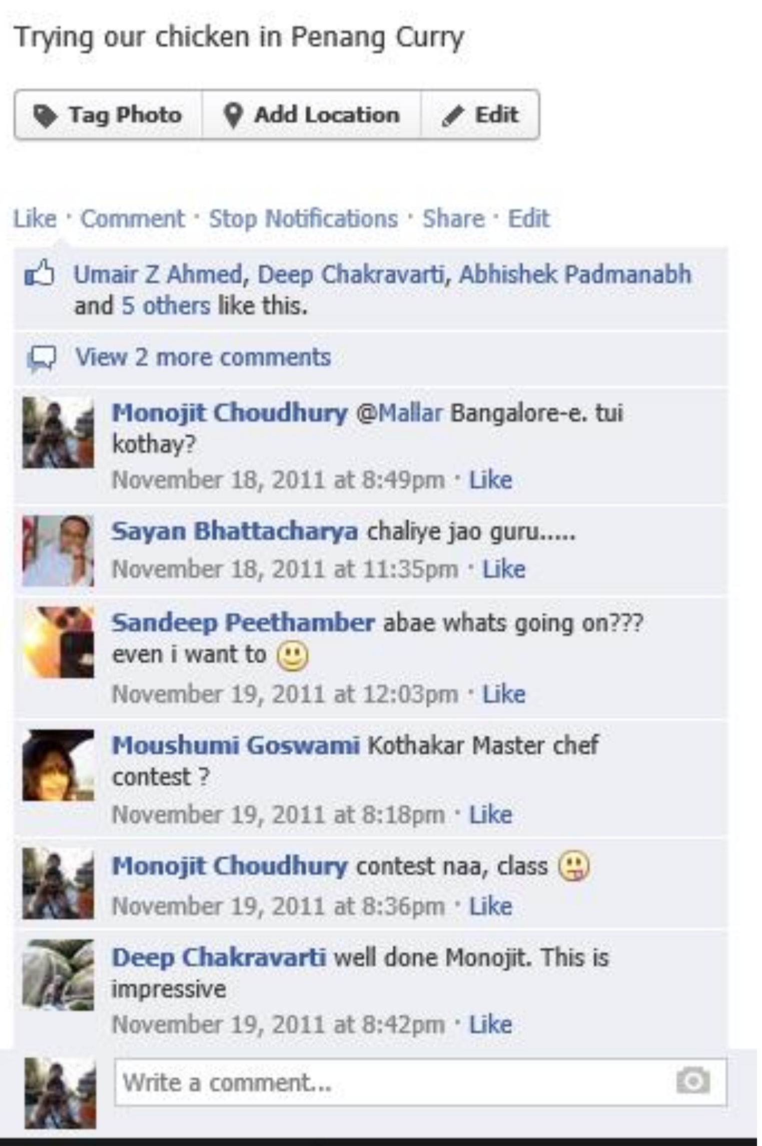
- What does “normalization” entail?
- Unintentional Spelling changes & Edit Distance
- Intentional spelling changes
- Patterns of intentional spelling changes

Anything that is needed to convert the non-standard text to an *equivalent* standard text which is processable by a standard language NLP system.

Classifying non-standard usage



Non-standard spellings	2mrw → tomorrow
Non-standard grammar	even i want to → Even I want to do this.
Language mixing	Kothakar\B Master\E chef\E contest\E ?
Transliteration	Kothakar → কোথাকার
Emoticons, Tags, mentions, slangs	abae → ?, :P → ?, @Mallar → ?, ??? → ?



Trying our chicken in Penang Curry

Umair Z Ahmed, Deep Chakravarti, Abhishek Padmanabh and 5 others like this.

View 2 more comments

Monojit Choudhury @Mallar Bangalore-e. tui kothay?
November 18, 2011 at 8:49pm · Like

Sayan Bhattacharya chaliye jao guru.....
November 18, 2011 at 11:35pm · Like

Sandeep Peethamber abae whats going on??? even i want to 🙂
November 19, 2011 at 12:03pm · Like

Moushumi Goswami Kothakar Master chef contest ?
November 19, 2011 at 8:18pm · Like

Monojit Choudhury contest naa, class 😛
November 19, 2011 at 8:36pm · Like

Deep Chakravarti well done Monojit. This is impressive
November 19, 2011 at 8:42pm · Like

Write a comment...

Classifying non-standard usage

Non-standard spellings	2mrw → tomorrow	Spelling Normalization	L2 & 3
Non-standard grammar	even i want to → Even I want to do this.	Grammar correction	
Language mixing	Kothakar\B Master\E chef\E contest\E ?	Language Detection	L4
Transliteration	Kothakar → কোথাকার	Machine Transliteration	L3
Emoticons, Tags, mentions, slangs	abae → ?, :P → ?, @Mallar → ?, ??? → ?	Special Treatment	L*

Old wine in new bottle

- Speech processing (ASR & TTS) requires normalization:
    - 2nd = *second*, 5.24% = *five point two four percent*, dr. = *doctor*
    - Rule based generation, with some rule-based or statistical disambiguation
    - (Sproat et al., 2001)
- Spelling and grammar correction
    - Spell checking (Kukich, 1992)
    - L2 error modeling and correction (Rozovskaya and Roth, 2011)
- SMS Normalization
    - (Aw et al., 2006)
    - (Choudhury et al., 2007)

What will we learn?

- What does “normalization” entail?
- Unintentional Spelling changes & Edit Distance
- Intentional spelling changes
- Patterns of intentional spelling changes

Spelling Errors

TOMORROW

- Tomorow
- Tommorow
- Tommorrow

Phonetic/Cognitive Errors

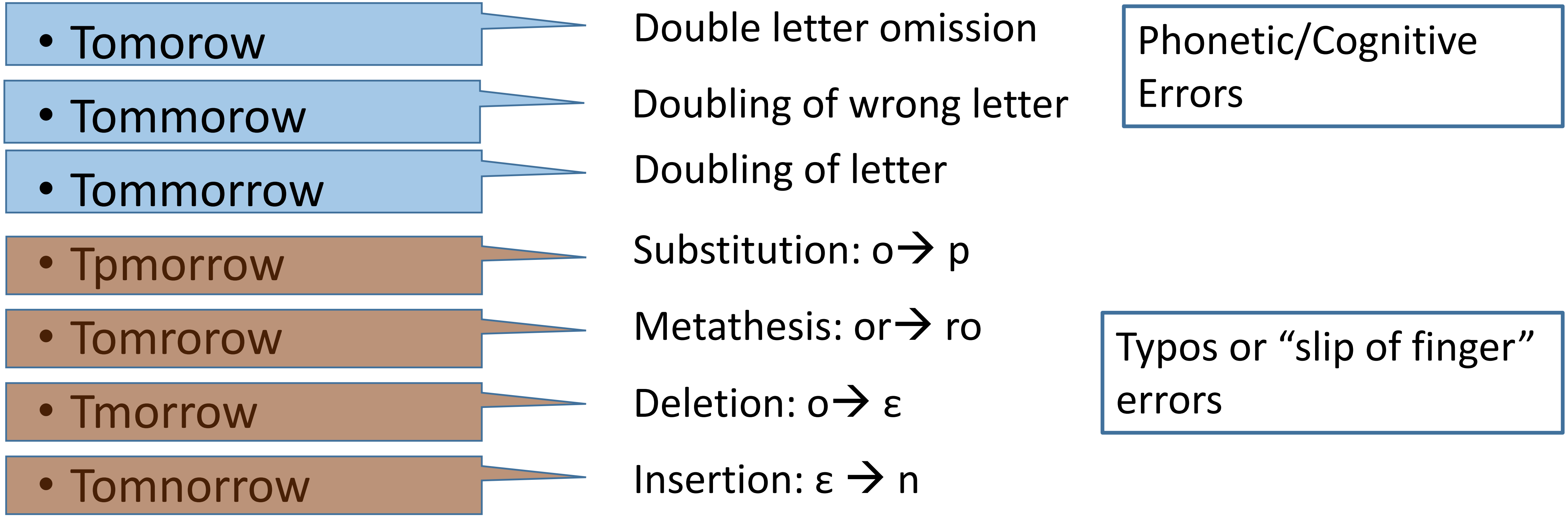
- Tpmorrow
- Tomrorow
- Tmorrow
- Tomnorrow

Typos or “slip of finger” errors

Unintentional Errors

Types of Unintentional Spelling Errors

TOMORROW



• Tomorow	Double letter omission	Phonetic/Cognitive Errors
• Tommorow	Doubling of wrong letter	
• Tommorrow	Doubling of letter	
• Tpmorrow	Substitution: o→ p	Typos or “slip of finger” errors
• Tomrorow	Metathesis: or→ ro	
• Tmorrow	Deletion: o→ ε	
• Tomnorrow	Insertion: ε → n	

What will we learn?

- What does “normalization” entail?
- Unintentional Spelling changes & Edit Distance
- Intentional spelling changes
- Patterns of intentional spelling changes

Edit Distance

- Cost of **Edit Operations**:
  - Insertion(ε → *c*): 1
  - Deletion (*c* → ε): 1
  - Substitution: (*c* → *c'*): 1 or 2

Metathesis (*cc'* → *c'c*) is either modeled as a single edit operation (cost = 1) or as a deletion-insertion pair (*cc'* → ε*c'* → *c'c*), and therefore cost of 2.

- **Edit Distance** between two strings **s**:$c_1c_2c_3 \ldots c_n$ and **s'**:$c'_1c'_2c'_3 \ldots c'_n$ is defined as the minimum value of the sum of the cost of a sequence of edit operations required to convert **s** to **s'**.
  - *engine* & *begin*, elevator & evaluator, east & csar

- Dynamic Programming Algorithm

What will we learn?

- What does “normalization” entail?
- Unintentional Spelling changes & Edit Distance
- Intentional spelling changes
- Patterns of intentional spelling changes

What about spelling errors in Social Media?

The shorter → the faster
Constraint: understandability

24

dis is n eg 4 txtin lang

39

This is an example for Texting language

Other factors: Coolness, group-membership, accommodating

Tomorrow never dies!!!

- 2moro (9)
- tomoz (25)
- tomoro (12)
- tomrw (5)
- tom (2)
- tomra (2)
- tomorrow (24)
- tomora (4)
- tomm (1)
- tomo (3)
- tomorow (3)
- 2mro (2)
- morrow (1)
- tomor (2)
- tmorro (1)
- moro (1)

Spell-checkers, such as Aspell, perform very poorly on such data (<22%)

Data from (Choudhury et al., 2007)

Patterns or Compression Operators

- Phonetic substitution (phoneme)
  - psycho → syco, then → den
- Phonetic substitution (syllable)
  - today → 2day , see → c
- Deletion of vowels
  - message → mssg, about → abt
- Deletion of repeated characters
  - tomorrow → tomorow

Data from (Choudhury et al., 2007)

Patterns or Compression Operators

- Truncation (deletion of tails)
  - introduction → intro, evaluation → eval
- Common Abbreviations
  - Kharagpur → kgp, text back → tb
- Informal pronunciation
  - going to → gonna
- Emphasis by repetition:
  - Funny → fuunnnnyyyyyy

Data from (Choudhury et al., 2007)

Successive Application of Operators

- Because → cause (informal usage)
- cause → cauz (phonetic substitution)
- cauz → cuz (vowel deletion)

Data from (Choudhury et al., 2007)

Summary

- Normalization involves transforming the non-standard input text to the standard forms (which then makes it possible to apply the standard NLP tools on the text).
- Normalization for Social Media text includes: orthographic normalization, grammar correction, language detection, transliteration, and handling of emoticons/hashtags etc.
- Unintentional spelling changes or errors are either because the user doesn’t know the correct spelling or due to “slip of fingers”.
- Orthographic Edit Distance is an efficient way to model and correct unintentional spelling errors.
- Motivation behind intentional spelling changes could be to type faster, emphasis, group identity and accommodation.
- Most of the changes are phonetically governed.

Suggested Readings

For Orthographic Patterns in Computer Mediated Communication:

Choudhury, Monojit, et al. "Investigation and modeling of the structure of texting language." *International Journal of Document Analysis and Recognition (IJDAR)* 10.3-4 (2007): 157-174.

For Spelling Correction Techniques and Algorithms:

Kukich, Karen. "Techniques for automatically correcting words in text." *ACM Computing Surveys (CSUR)* 24.4 (1992): 377-439.

References

- Sproat, Richard, et al. "Normalization of non-standard words." *Computer Speech & Language* 15.3 (2001): 287-333.
- Rozovskaya, Alla, and Dan Roth. "Algorithm selection and model adaptation for ESL correction tasks." *ACL*, 2011.
- Kukich, Karen. "Techniques for automatically correcting words in text." *ACM Computing Surveys (CSUR)* 24.4 (1992): 377-439.
- Choudhury, Monojit, et al. "Investigation and modeling of the structure of texting language." *International Journal of Document Analysis and Recognition (IJDAR)* 10.3-4 (2007): 157-174.
- Aspell: http://aspell.net/