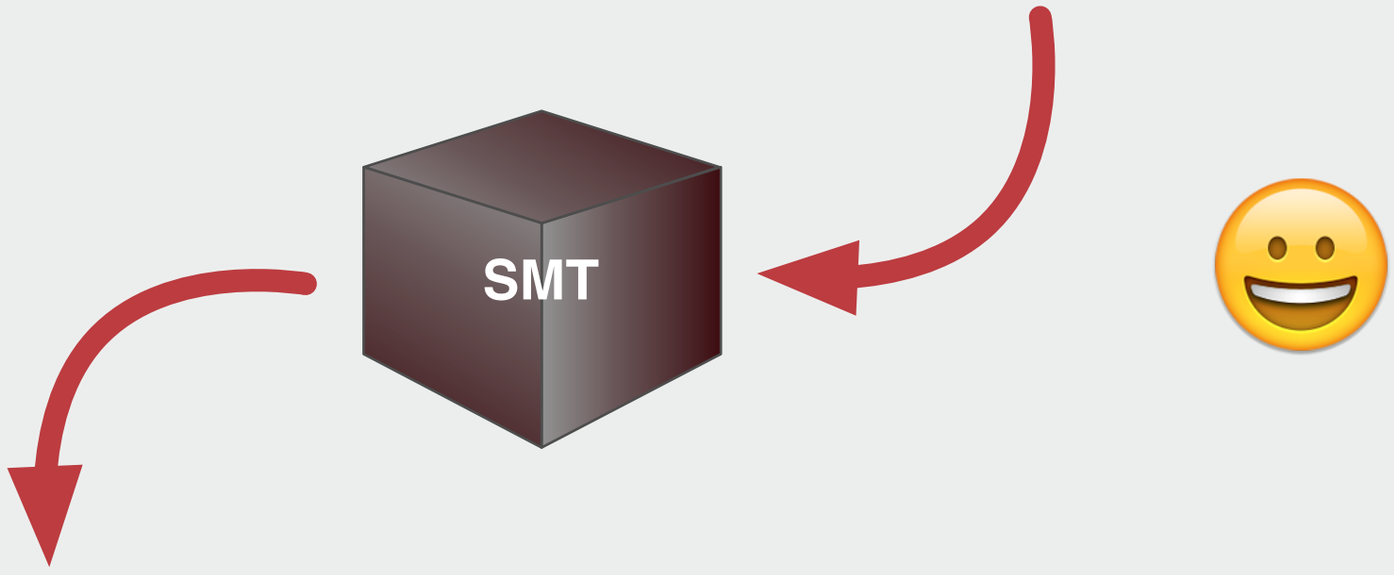
UNIVERSITY OF AMSTERDAM

# Five Shades of Noise: Analyzing Machine Translation Errors in User-Generated Text

Marlies van der Wees, Arianna Bisazza, Christof Monz

# Statistical Machine Translation

News sentence: 印度金融中心孟买亦受到波及。
(mumbai, india's financial center, was also affected.)

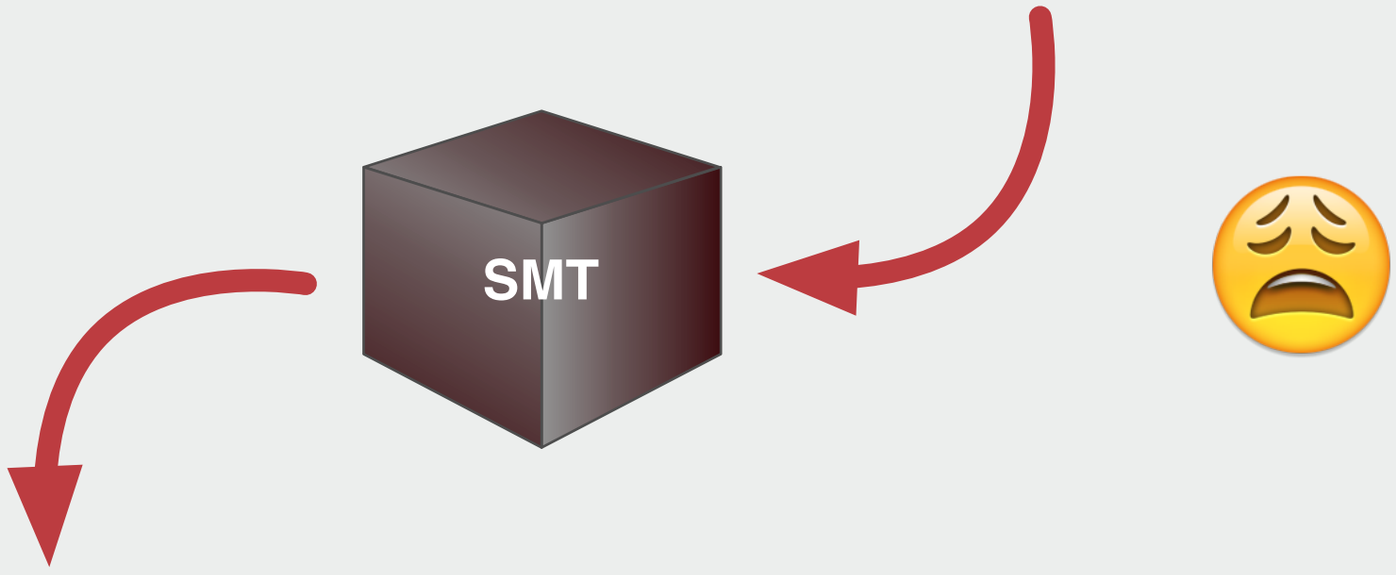



india's financial center mumbai also affected.

# Statistical Machine Translation

SMS sentence: 你路上慢点

(be careful on your way / take your time)




you are on the road to slow points

# SMT for user-generated text is often bad

- Reference
  - and if i go out, i will stop by your place
  - i could not bring it to you
  - i've never seen a pig there
  - you're too delighted to be homesick
- SMT output
  - and if i went.
  - into its enemies.
  - i am seen pig there.
  - anytime you

# Towards improving SMT quality for UG

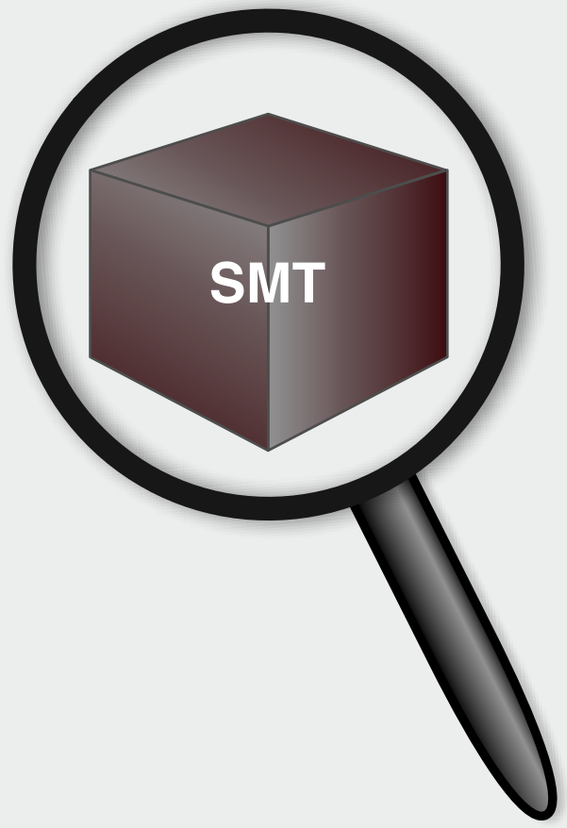
- To target specific error types, we need to know why mistakes are made:
  - in UG versus formal text
    - contrast UG with newswire
  - in different types of UG
    - five shades of noise: weblogs, comments, speech (CTS), SMS, and chat messages
  - in different language pairs
    - Arabic-English & Chinese-English

# Analyzing SMT errors in UG text

- What translation choices were made by the SMT system?

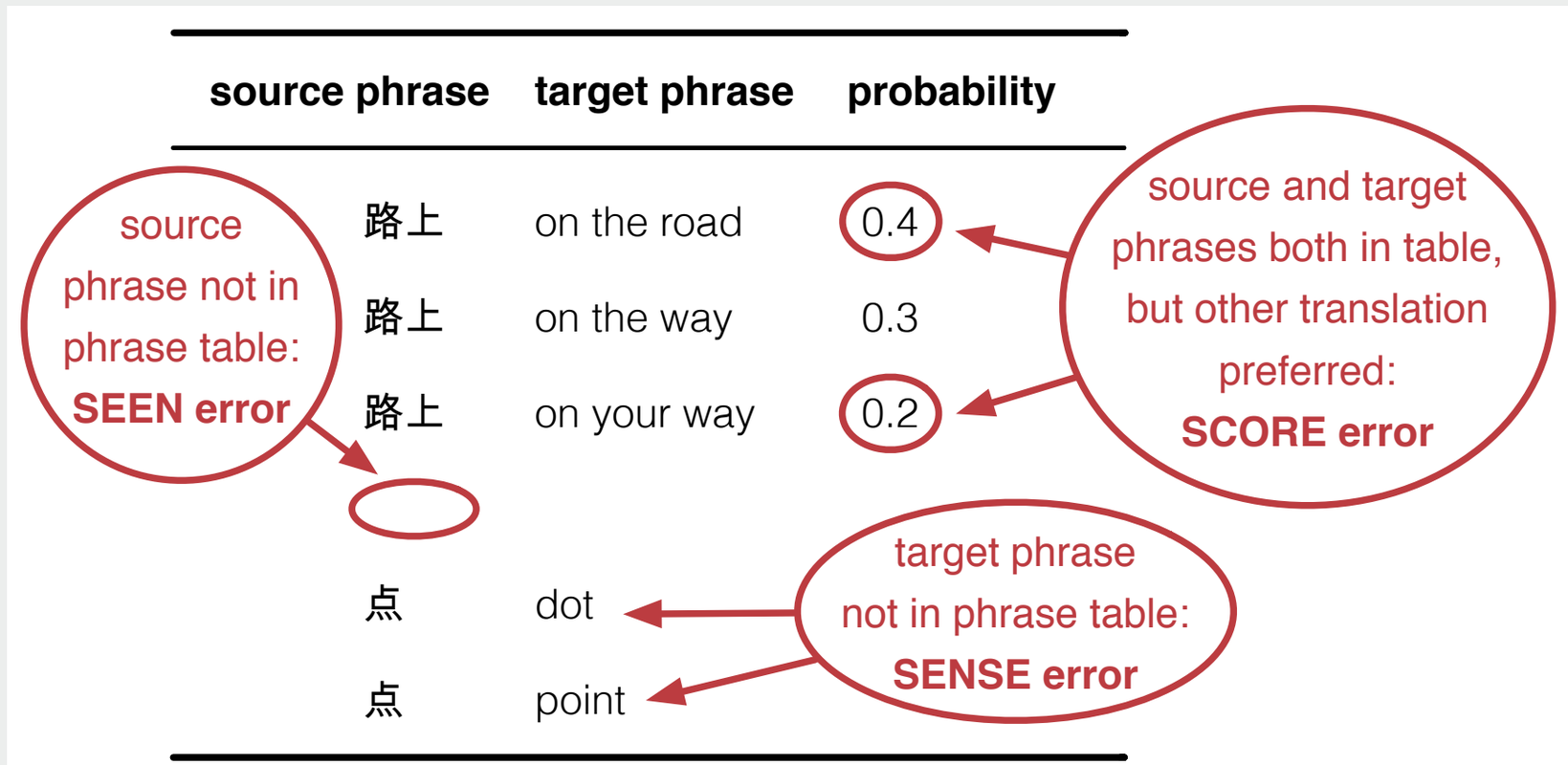
- What translation choices could have been made by the SMT system?

- Why did the SMT system make the choices that it made?



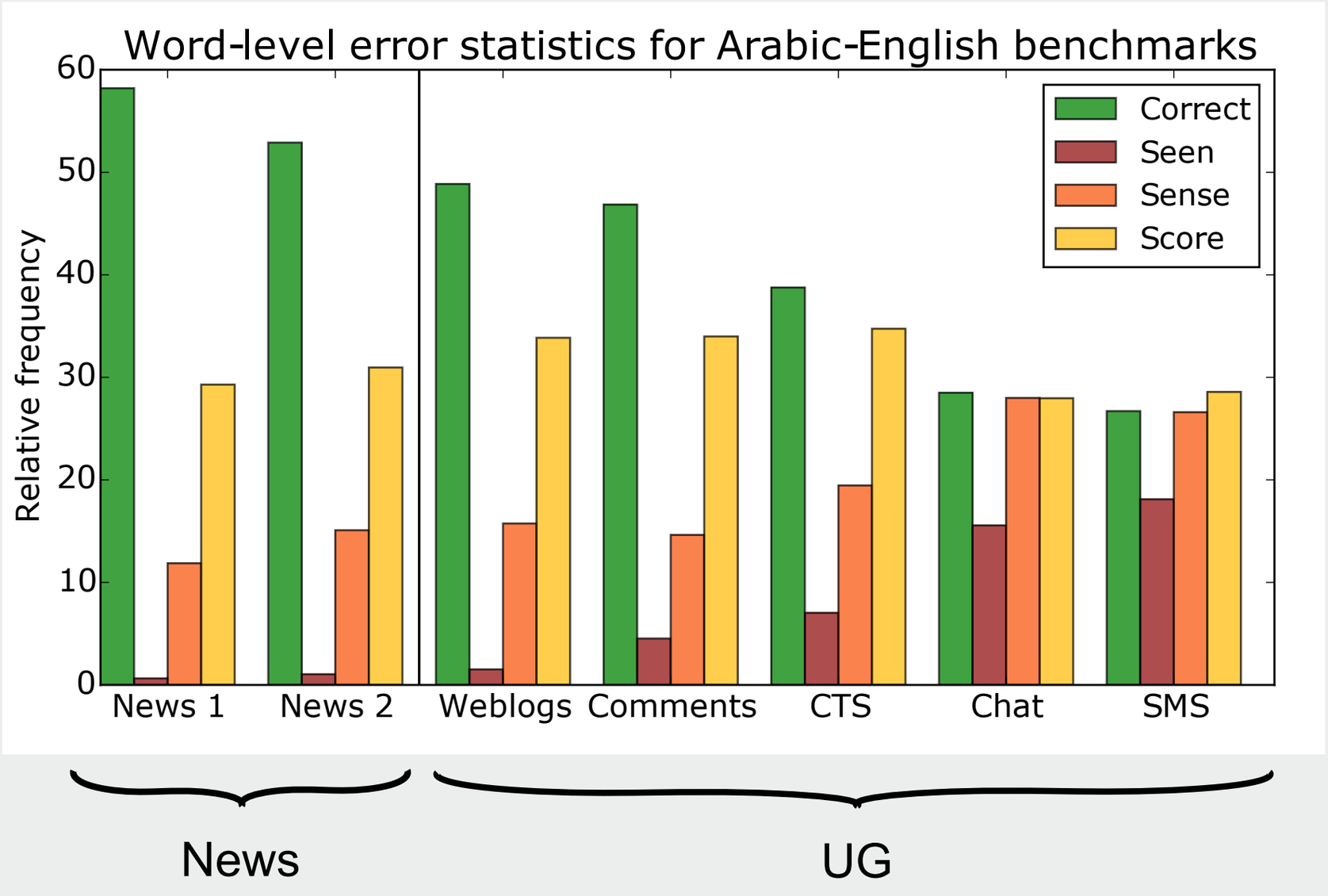
# Word Alignment Driven Evaluation: approach*

- For each word alignment link in the test (e.g. 你 — your) that is translated wrongly, determine:




* Approach adopted from Irvine et al., *Measuring Machine Translation Errors in New Domains*, 2013

# Word Alignment Driven Evaluation: results



# Word Alignment Driven Evaluation: findings

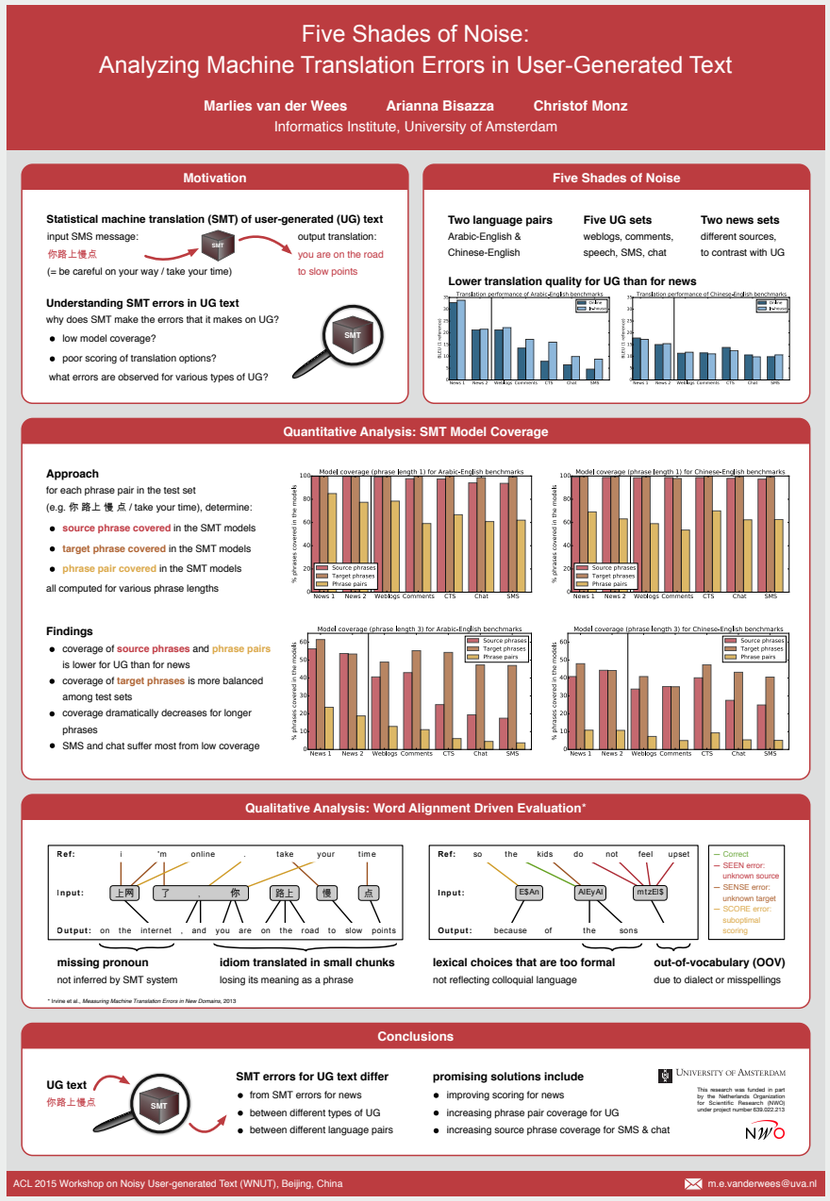
- SMT errors for UG text differ
  - from SMT errors for news
    - many SEEN and SENSE errors for UG
  - between different types of UG
    - SMS and chat messages are most affected
  - between different language pairs
    - differences in Chinese-English are more subtle than in Arabic-English

# Analyzing SMT errors in UG: what we learned

- Common errors in UG are due to:
  - misspellings or Arabic dialectal forms
  - formal lexical choices
  - idioms translated word by word
  - dropped pronouns in Chinese
- UG suffers from low model coverage
  - generate new translation candidates
  - normalize existing translation candidates

UNIVERSITY OF AMSTERDAM

# More Error Analysis?



Five Shades of Noise:
Analyzing Machine Translation Errors in User-Generated Text

Marlies van der Wees  Arianna Bisazza  Christof Monz
Informatics Institute, University of Amsterdam

**Motivation**

**Statistical machine translation (SMT) of user-generated (UG) text**

input SMS message: 你路上慢点 (= be careful on your way / take your time)

output translation: you are on the road to slow points

**Understanding SMT errors in UG text**

why does SMT make the errors that it makes on UG?

- low model coverage?
- poor scoring of translation options?

what errors are observed for various types of UG?

**Five Shades of Noise**

**Two language pairs** Arabic-English & Chinese-English

**Five UG sets** weblogs, comments, speech, SMS, chat

**Two news sets** different sources, to contrast with UG

**Lower translation quality for UG than for news**

**Quantitative Analysis: SMT Model Coverage**

**Approach**

for each phrase pair in the test set (e.g. 你 路上 慢 点 / take your time), determine:

- **source phrase covered** in the SMT models
- **target phrase covered** in the SMT models
- **phrase pair covered** in the SMT models

all computed for various phrase lengths

**Findings**

- coverage of **source phrases** and **phrase pairs** is lower for UG than for news
- coverage of **target phrases** is more balanced among test sets
- coverage dramatically decreases for longer phrases
- SMS and chat suffer most from low coverage

**Qualitative Analysis: Word Alignment Driven Evaluation***

Ref: i 'm online . take your time
Input: 上网 了 , 你 路上 慢 点
Output: on the internet , and you are on the road to slow points

**missing pronoun** not inferred by SMT system

**idiom translated in small chunks** losing its meaning as a phrase

Ref: so the kids do not feel upset
Input: EStAn AlEyAl mtzEl$
Output: because of the sons

**lexical choices that are too formal** not reflecting colloquial language

**out-of-vocabulary (OOV)** due to dialect or misspellings

* Irvine et al., *Measuring Machine Translation Errors in New Domains*, 2013

**Conclusions**

UG text 你路上慢点

**SMT errors for UG text differ**

- from SMT errors for news
- between different types of UG
- between different language pairs

**promising solutions include**

- improving scoring for news
- increasing phrase pair coverage for UG
- increasing source phrase coverage for SMS & chat

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.213

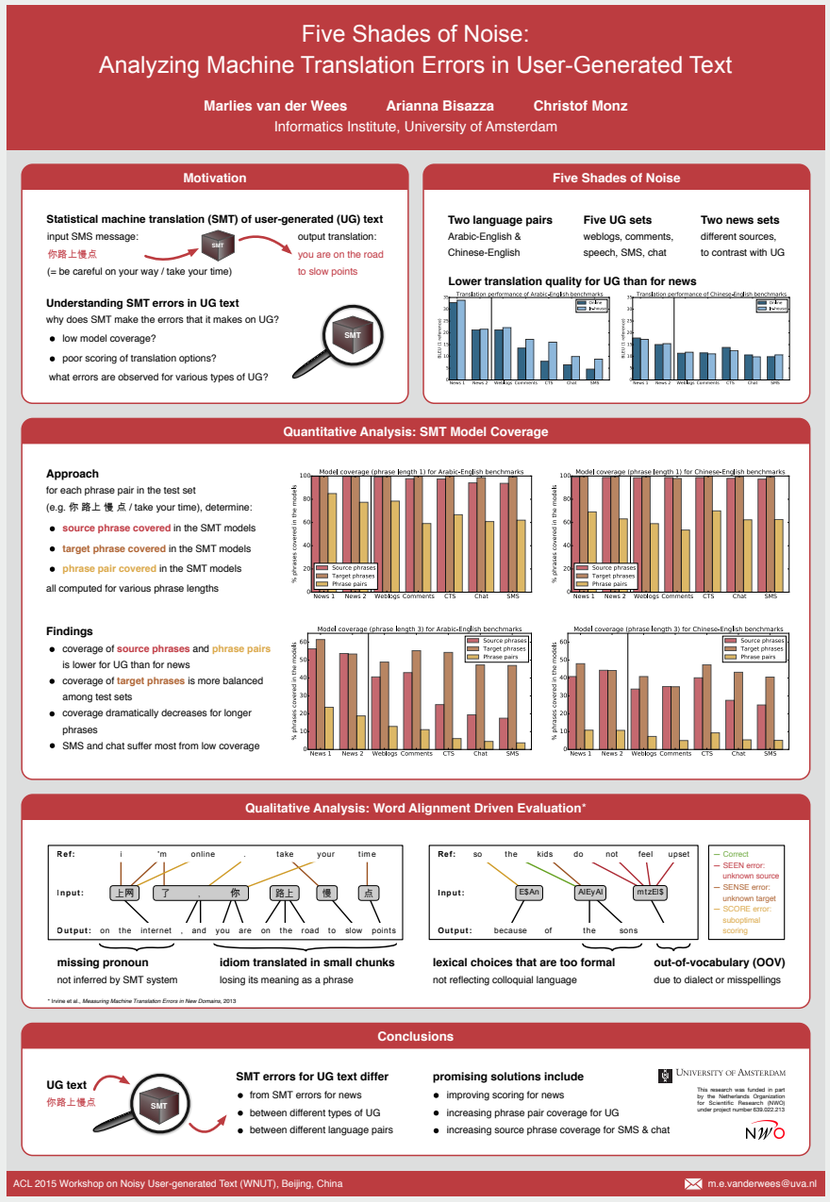
ACL 2015 Workshop on Noisy User-generated Text (WNUT), Beijing, China

m.e.vanderwees@uva.nl

- Visit the poster for:
  - Model coverage analysis
  - Arabic-English versus Chinese-English results
  - Qualitative Examples
- Read the paper for:
  - Phrase-length analysis
  - Detailed explanation and discussions

UNIVERSITY OF AMSTERDAM

# Thank you!



Five Shades of Noise:
Analyzing Machine Translation Errors in User-Generated Text

**Marlies van der Wees** **Arianna Bisazza** **Christof Monz**
Informatics Institute, University of Amsterdam

**Motivation**

**Statistical machine translation (SMT) of user-generated (UG) text**
input SMS message: 你路上慢点 (= be careful on your way / take your time)
output translation: you are on the road to slow points

**Understanding SMT errors in UG text**
why does SMT make the errors that it makes on UG?
- low model coverage?
- poor scoring of translation options?

what errors are observed for various types of UG?

**Five Shades of Noise**

**Two language pairs** Arabic-English & Chinese-English
**Five UG sets** weblogs, comments, speech, SMS, chat
**Two news sets** different sources, to contrast with UG

**Lower translation quality for UG than for news**

**Quantitative Analysis: SMT Model Coverage**

**Approach**
for each phrase pair in the test set
(e.g. 你 路上 慢 点 / take your time), determine:
- **source phrase covered** in the SMT models
- **target phrase covered** in the SMT models
- **phrase pair covered** in the SMT models

all computed for various phrase lengths

**Findings**
- coverage of **source phrases** and **phrase pairs** is lower for UG than for news
- coverage of **target phrases** is more balanced among test sets
- coverage dramatically decreases for longer phrases
- SMS and chat suffer most from low coverage

**Qualitative Analysis: Word Alignment Driven Evaluation***

Ref: i 'm online . take your time
Input: 上网 了 你 路上 慢 点
Output: on the internet , and you are on the road to slow points

**missing pronoun** not inferred by SMT system
**idiom translated in small chunks** losing its meaning as a phrase

Ref: so the kids do not feel upset
Input: ESAn AlEyAl m1zEl$
Output: because of the sons

**lexical choices that are too formal** not reflecting colloquial language
**out-of-vocabulary (OOV)** due to dialect or misspellings

* Irvine et al., *Measuring Machine Translation Errors in New Domains*, 2013

**Conclusions**

UG text 你路上慢点

**SMT errors for UG text differ**
- from SMT errors for news
- between different types of UG
- between different language pairs

**promising solutions include**
- improving scoring for news
- increasing phrase pair coverage for UG
- increasing source phrase coverage for SMS & chat

University of Amsterdam
This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.213
NWO

ACL 2015 Workshop on Noisy User-generated Text (WNUT), Beijing, China m.e.vanderwees@uva.nl


- Marlies van der Wees
- m.e.vanderwees@uva.nl