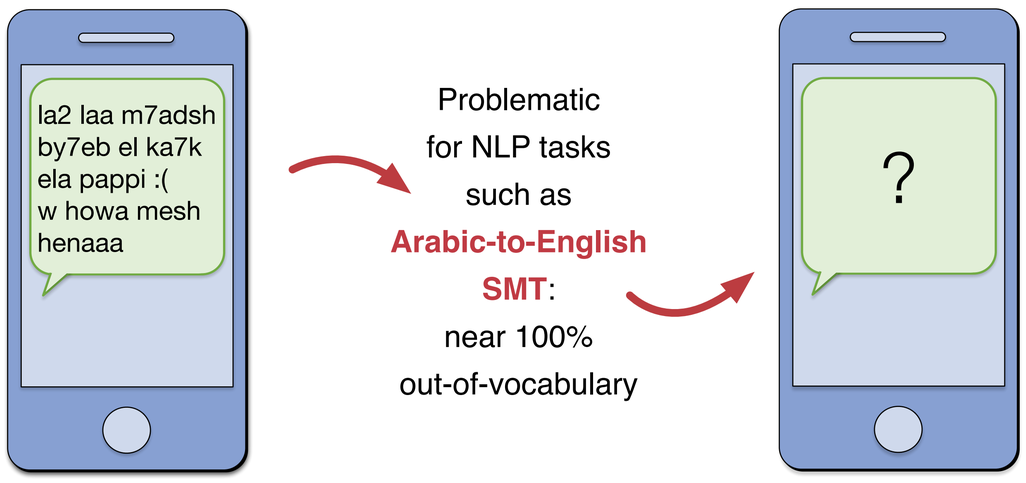
A Simple but Effective Approach to Improve Arabizi-to-English Statistical Machine Translation

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

Motivation

User-generated Arabic often written in Arabizi (romanized Arabic)



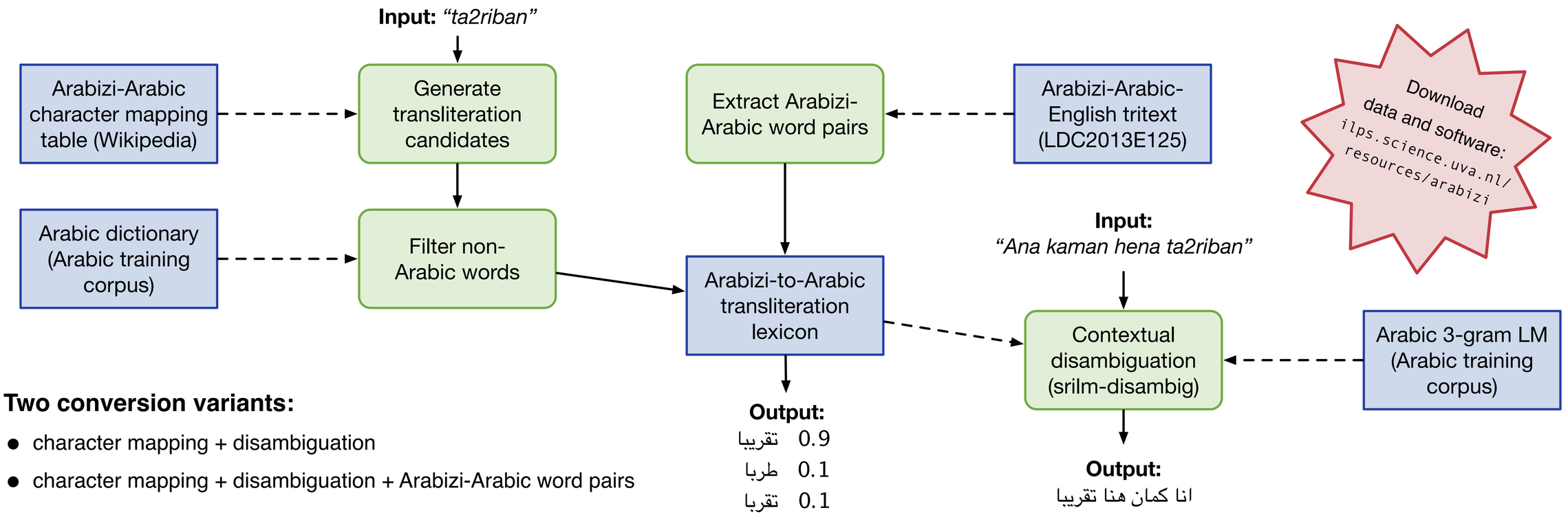
Problematic for NLP tasks such as **Arabic-to-English SMT**: near 100% out-of-vocabulary

What is Arabizi?

Arabic written in Latin script

- combination of numbers and Latin letters
- emerged due to practical reasons: lack of Arabic keyboard
- no standardization as in official transliteration schemes
- character mappings are:
  - based on orthographic similarities: ع **looks like 3**
  - based on phonetic similarities: ف **sounds like f**
  - highly ambiguous: ط **looks like 6** but **sounds like t**
- sensitive to dialects

Arabizi-to-Arabic Conversion Approach



Two conversion variants:

- character mapping + disambiguation
- character mapping + disambiguation + Arabizi-Arabic word pairs

Transliteration and Translation Results

Arabizi-to-Arabic transliteration

la2 laa m7adsh
by7eb el ka7k
ela pappi :(
w howa mesh
henaaa

Converted Arabizi:
لا لا محدش بيحب ال كحك الا ببب :(و هو مش هنا

Arabic (human):
لا لا محدش بيحب ال كحك الا بابا : (و هو مش هنا

Word-level transliteration error rates:

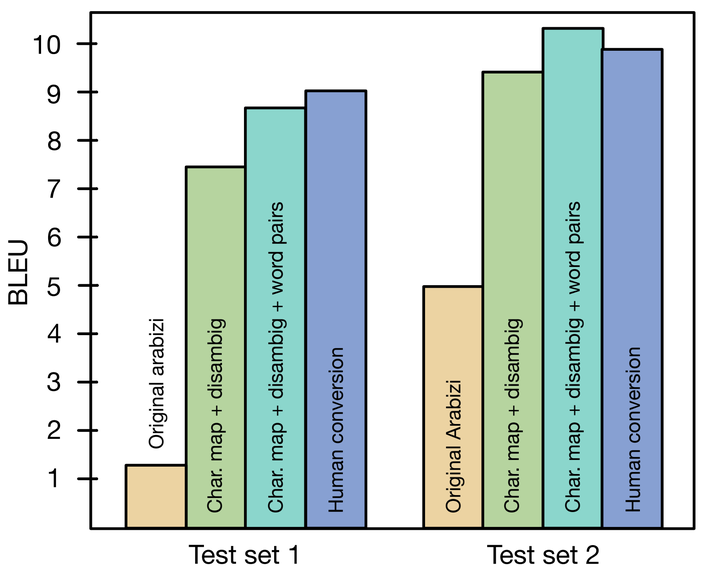
- char. mapping + disambig: **~50%**
- char. mapping + disambig + Arabizi-Arabic word pairs: **~25%**

Arabizi-to-English translation

English (SMT after conversion):
no, no, no one loves you talk, but ببب :(, and he is not here

English (human):
no, no one likes cookies except my father :(and he's not here

If no transliteration found:
pass to SMT system which uses
fill-up with small Arabizi-English bitext



Conclusions

la2 laa m7adsh
by7eb el ka7k
ela pappi :(
w howa mesh
henaaa

Arabizi

- Romanized Arabic in UGC
- not standardized or regulated
- problematic for Arabic-to-English machine translation: ~100% OOV

Our solution

- simple transliteration pipeline
- no expert knowledge needed
- use of only publicly available data and tools

Results

- word-level error rates 25-50%
- SMT performance meets performance after human transliteration
- pipeline and data available for download

UNIVERSITY OF AMSTERDAM

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

NWO

m.e.vanderwees@uva.nl