What's in a Domain?
Analyzing Genre and Topic Differences in Statistical Machine Translation

Marlies van der Wees **Arianna Bisazza** **Wouter Weerkamp*** **Christof Monz**

Informatics Institute, University of Amsterdam *904Labs, Amsterdam

What's in a Domain?

Domain ambiguous concept in adaptation for SMT

- **Topic** general subject
- **Genre** function / style / type, orthogonal to topic

Genre ≠ Topic

but difference is often neglected in adaptation for SMT

Our contributions

- disentangling the concepts of **genre** and **topic**
- introducing the Gen&Topic evaluation set
- quantifying the impact of **genre** and **topic** on SMT
- analyzing OOV types for **genres** and **topics**

The Gen&Topic Evaluation Set

Arabic-English
web-crawled
manual translations

Two genres
newswire (NW) &
comments (UG)

Five topics
culture, health, politics,
economy & security

Controlled genre-topic distributions
similar-sized NW and UG fragments
from all NW-UG document pairs
that discuss the same article:

Download: ilps.science.uva.nl/resources/gen-topic

	News (NW)	Comment (UG)
Culture	The 12 contestants competed during a May 3rd Prime before a panel of judges and millions of viewers across the Arab world.	Your program's name is 'Arab Idol', which is in English, and you allowed Barwas to participate and represent Iraq while she sings in Kurdish!!!
Economy	Yemen is mulling the establishment of 13 industrial zones across its six planned administrative regions in a bid to stimulate development and create job opportunities.	What development in Yemen are you talking about? We will continue to call for freedom until independence and liberation and the routing of the northern occupation from our lands.

Impact of Genre and Topic Differences on SMT

Translation model
balanced **genres**: 50% NW, 50% UG
wide range of **topics**

Language model
linear interpolation
covers all **genres** & **topics**

- **Translation quality** (BLEU)

NW	19.9
UG	16.0

Δ: **3.9** > Δ: **0.1-1.1**

19.3	Culture
18.9	Economy
18.8	Health
18.5	Security
18.2	Politics

- **Average source-side phrase length** (#tokens)

NW	1.45
UG	1.38

Δ: **0.07** > Δ: **0.0-0.03**

1.42	Economy
1.42	Security
1.41	Health
1.41	Politics
1.39	Culture

- **Phrase pair model coverage** (% of test-set phrase pairs covered by translation model)

NW	28.5
UG	24.0

Δ: **4.5** > Δ: **0.1-1.9**

26.7	Security
26.6	Politics
26.2	Economy
25.8	Culture
24.8	Health

Manual OOV Analysis

Five out-of-vocabulary (OOV) classes
rare (rare words), morph (morphological variants),
dial (dialectal forms), spell (spelling errors), coll (colloquialisms)

Differences between genres
rare words dominate NW
spelling errors dominate UG

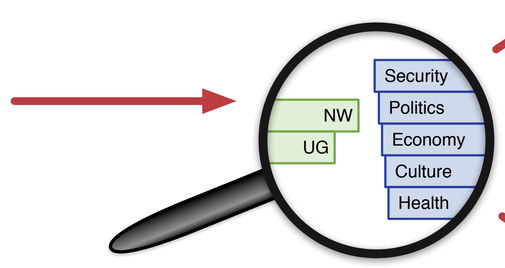
Differences between topics
more subtle than genre differences,
dial & morph least uniformly distributed

Examples:

class	Arabic OOV	English translation	explanation of OOV
rare	داعش	ISIL	new proper noun
dial	هينسوا	(they) will forget	dialectal future tense
morph	يقدسون	(they) revere	third person plural
spell	توفيرالوظائف	creationofjobs	missing blank
coll	المتطوعيين	volunteeeeers	repeated characters

Conclusions

Gen&Topic data set



Genre differences have larger impact on SMT than **topic** differences

Advice for **topic** adaptation: improve lexical selection

Advice for **genre** adaptation: increase model coverage

University of Amsterdam
904 Labs
NWO Netherlands Organisation for Scientific Research

m.e.vanderwees@uva.nl