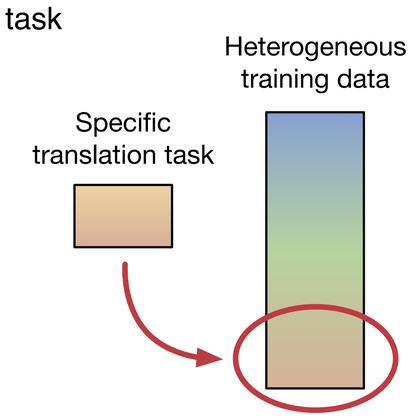
Translation Model Adaptation Using Genre-Revealing Text Features

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

General Task

adaptation for SMT

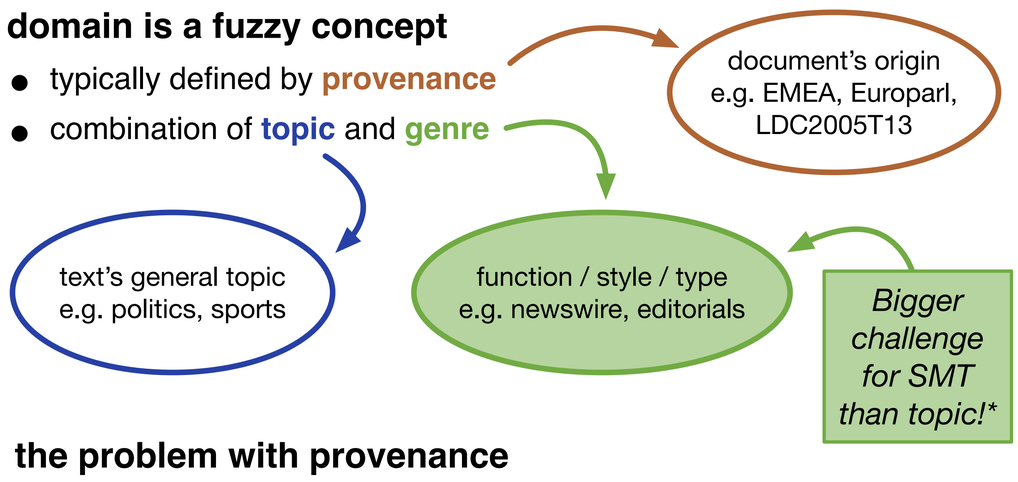
prioritize translation candidates that are most relevant to current task



Problem

domain is a fuzzy concept

- typically defined by **provenance**
- combination of **topic** and **genre**



the problem with provenance

- not an intrinsic text property, requires document's meta-information
- if not available: manual labeling is labor-intensive and can be arbitrary

* Van der Wees et al., *What's in a Domain? Analyzing Genre and Topic in Statistical Machine Translation*, 2015

Contribution

genre adaptation for SMT

- adapt system to multi-genre translation task
- exploit document-level genre-revealing text features inspired by classification literature
- replace dependency on manual domain labels with automatic measures of genre

Approach

adaptation scenario

- Arabic-English phrase-based SMT
- translation model adaptation with a vector space modeling (VSM) approach
- two test sets with two genres:
  Gen&Topic*:
  - newswire (NW)
  - user-generated comments (UG)

  NIST 2008+2009:
  - newswire (NW)
  - user-generated weblogs (UG)

vector space modeling (VSM) adaptation approach**

new decoder feature: similarity between each phrase pair and dev set

source	target	p(f|e)	p(e|f)	…	phrase vector	similarity score
الحمد ل	praise be to	0.1	0.2	…	$< w_1 \dots w_N >$	0.1
الحمد ل	praise for	0.2	0.2	…	$< w_1 \dots w_N >$	0.2
الحمد ل	thank	0.1	0.2	…	$< w_1 \dots w_N >$	0.4
حبيبتي ي	my dear	0.2	0.1	…	$< w_1 \dots w_N >$	0.3
حبيبتي ي	my love	0.2	0.1	…	$< w_1 \dots w_N >$	0.4
حبيبتي ي	my sweetheart	0.1	0.1	…	$< w_1 \dots w_N >$	0.1

Vector for development set: $< w_1(\text{dev}) \dots w_N(\text{dev}) >$

vectors can be constructed from

- **provenance** features:
  manually grouped subcorpus labels
- **topic** features:
  LDA-inferred topics
- **genre** features:
  `counts of exclamation marks, question marks, repeating punctuation, emoticons, numbers, first & second person pronouns`

* Van der Wees et al., *What's in a Domain? Analyzing Genre and Topic in Statistical Machine Translation*, 2015

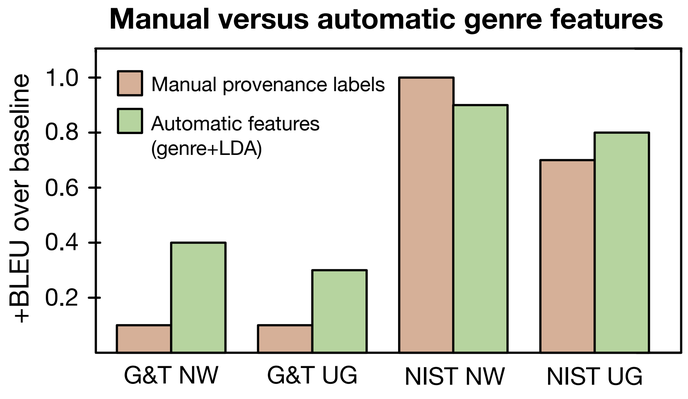
** Following Chen et al., *Vector Space Model for Adaptation in Statistical Machine Translation*, 2013

Results

improved translation performance

- automatic indicators of genre can replace manual sub-corpus labels
- best system with automatic features: genre+LDA

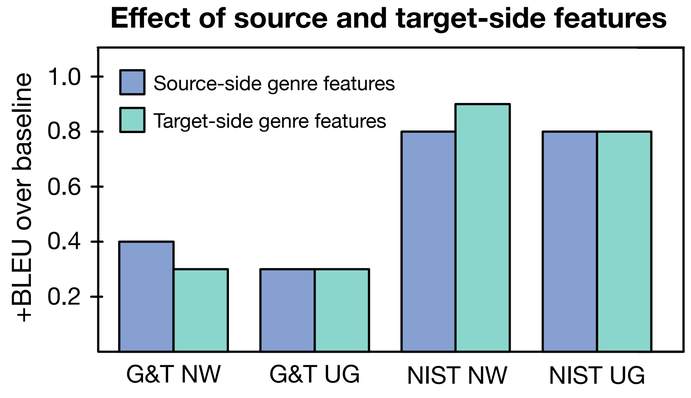
Manual versus automatic genre features



projection across languages

- similar performance for feature values computed on Arabic or English side of the bitext

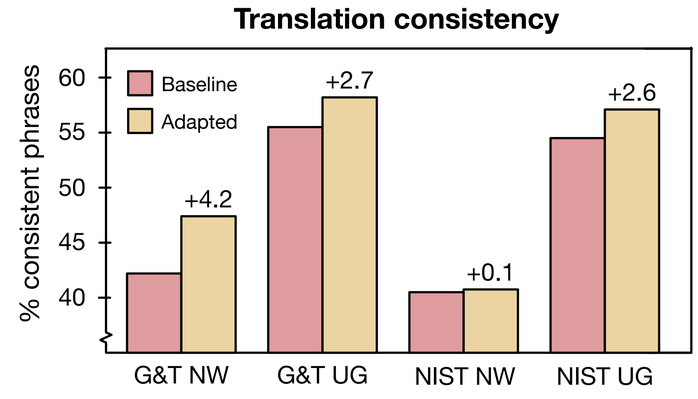
Effect of source and target-side features



increased translation consistency

- phrases with identical translations for each occurrence in a single document
- higher consistency for genre-adapted system

Translation consistency



Conclusions

we address genre adaptation by

- distinguishing genre from provenance and topic
- using genre-revealing text features for translation model adaptation
- eliminating the need for manual sub-corpus labels

the proposed method

- improves translation quality over a competitive baseline
- exploits features that can be projected across languages
- increases document-level translation consistency

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