Polylingual Topic Models

Hanna M. Wallach

University of Massachusetts Amherst
wallach@cs.umass.edu

August 7, 2009

Joint work with D. Mimno, J. Naradowsky, D.A. Smith and A. McCallum

Statistical Topic Models

- Useful for analyzing large, unstructured text collections

bounds	units	policy	data	neurons
bound	hidden	action	space	neuron
loss	network	reinforcement	clustering	spike
functions	layer	learning	points	synaptic
error	unit	actions	distance	firing

- Topic-based search interfaces (`http://rexa.info`)
- Analysis of scientific progress over time (Blei & Lafferty, '07)
- Information retrieval (Wei & Croft, '06)

Automated Analysis of Text

- Previously: analyzing trends in text collections (Hall et al., '08)
- Monolingual models often work well: collections in English only
- Multilingual text collections are increasingly common
- Automated tools are most important for multilingual collections:
  - Don't know the language → cannot eyeball the data
  - New documents will appear in other languages
  - People typically only know a few languages
- Simultaneously analyze document content in many languages

Multiple Languages

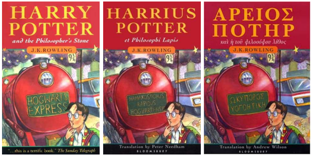
- Why model multiple languages explicitly?
- Most statistical topic models are language-agnostic

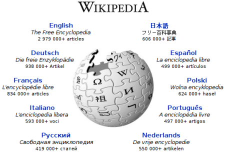
graph	problem	rendering	algebra	und	la
graphs	problems	graphics	algebras	von	des
edge	optimization	image	ring	die	le
vertices	algorithm	texture	rings	der	du
edges	programming	scene	modules	im	les

- Hodgepodge of English, German, French topics
- Imbalanced corpus: maybe only one or two French topics

Parallel vs. Comparable Corpora

- A set of aligned documents is a “document tuple”

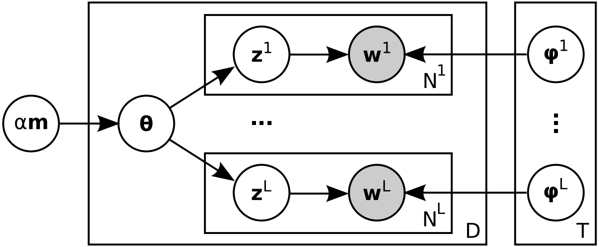

HARRY POTTER
and the Philosopher's Stone
J.K.ROWLING
HOGWARTS EXPRESS
"...this is a terrific book." The Sunday Telegraph
HARRIUS POTTER
et Philosophi Lapis
J.K.ROWLING
Translation by Peter Needham
BLOOMSBURY
ΑΡΕΙΟΣ ΠΟΤΗΡ
καὶ ἡ τοῦ φιλοσόφου λίθος
J.K.ROWLING
Translation by Andrew Wilson
BLOOMSBURY



- Fully parallel corpora: documents are direct translations
- Corpora with a few parallel “glue” document tuples
- Comparable corpora: documents have similar semantic content

Polylingual Topic Model

- Generates a document tuple $\mathbf{w} = \mathbf{w}^1, \ldots, \mathbf{w}^L$ by drawing...

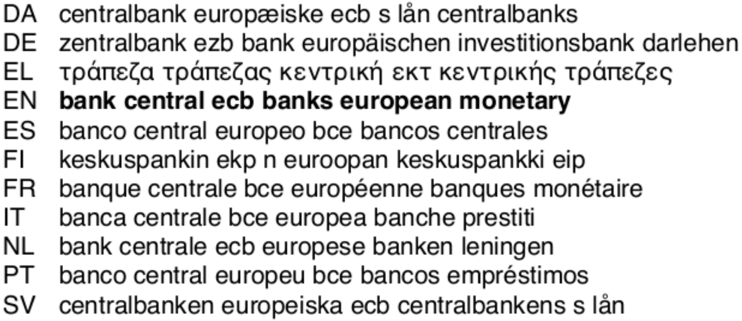


- For real-world data, only the word tokens are observed

Key Characteristics

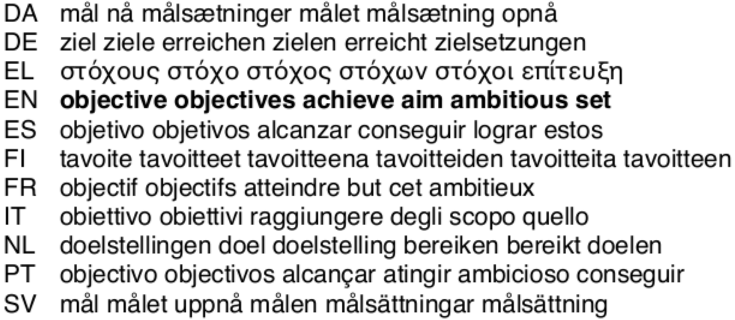
- Learning a model of *all* languages simultaneously
- A topic is a *set* of distributions over words, e.g., $\phi_t = \phi_t^1, \ldots, \phi_t^L$
- Works on tuples of aligned documents, rather than documents, but each tuple can be comprised of only a subset of languages
- Tuple-specific topic distributions ensure cross-language consistency: e.g., topic 13 in French is semantically similar to topic 13 in English
- Simple, Gibbs sampling inference algorithm
  - Inference is linear in # of languages, not # of language pairs

EuroParl: Example Topics ($T = 400$)



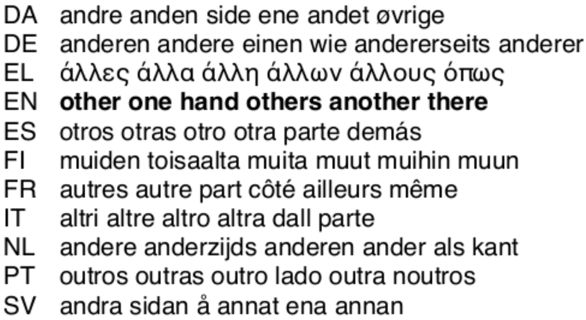
DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

EuroParl: Example Topics ($T = 400$)



DA	mål nå målsætninger målet målsætning opnå
DE	ziel ziele erreichen zielen erreicht zielsetzungen
EL	στόχους στόχο στόχος στόχων στόχοι επίτευξη
EN	**objective objectives achieve aim ambitious set**
ES	objetivo objetivos alcanzar conseguir lograr estos
FI	tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen
FR	objectif objectifs atteindre but cet ambitieux
IT	obiettivo obiettivi raggiungere degli scopo quello
NL	doelstellingen doel doelstelling bereiken bereikt doelen
PT	objectivo objectivos alcançar atingir ambicioso conseguir
SV	mål målet uppnå målen målsättningar målsättning

EuroParl: Example Topics ($T = 400$)



DA	andre anden side ene andet øvrige
DE	anderen andere einen wie andererseits anderer
EL	άλλες άλλα άλλη άλλων άλλους όπως
EN	**other one hand others another there**
ES	otros otras otro otra parte demás
FI	muiden toisaalta muita muut muihin muun
FR	autres autre part côté ailleurs même
IT	altri altre altro altra dall parte
NL	andere anderzijds anderen ander als kant
PT	outros outras outro lado outra noutros
SV	andra sidan å annat ena annan

Parallel Corpora: “Glue” Tuples

- How many aligned documents are needed to get aligned topics?

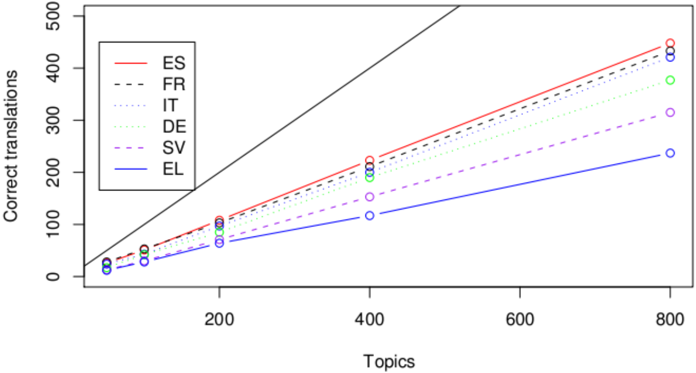
	1% “glue” document tuples
DE	rußland russland russischen tschetschenien sicherheit
EN	china rights human country s burma
IT	ho presidente mi perché relazione votato

	25% “glue” document tuples
DE	rußland russland russischen tschetschenien ukraine
EN	russia russian chechnya cooperation region belarus
IT	russia unione cooperazione cecenia regione russa

Generating Bilingual Lexica

- Bilingual lexicon: word pairs (e.g., English word, translation)
- High probability words in different languages for a topic are likely to include translations – can use these to generate lexica
- Advantages: unsupervised; all kinds of words, not just nouns
- Form candidate translations: Cartesian product of most probable K words in English and in each translation language
- Count # of lexicon pairs that are in the candidate set
- No morphological variants: e.g., rules/vorschriften, rule/vorschrift

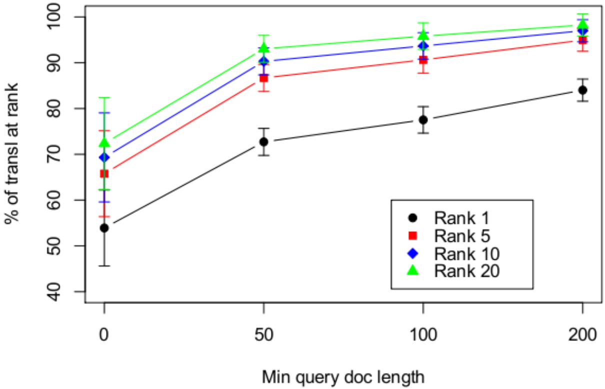
Generating Bilingual Lexica ($K = 1$)



Finding Translations

- Train model on aligned document tuples
- Output: set of polylingual topics, e.g., $\phi_t = \phi_t^1, \ldots, \phi_t^L$
- Map each test document to the low-dimensional space defined by the polylingual topics → document-topic distributions
- For each query/target language pair:
    - Compute similarities for all query/target document pairs
    - For each query document, rank target documents by similarity
- Jensen-Shannon divergence, cosine distance

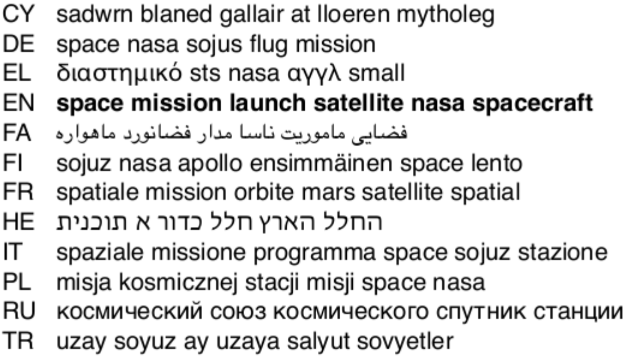
Finding Translations (Jensen-Shannon)



Comparable Corpora

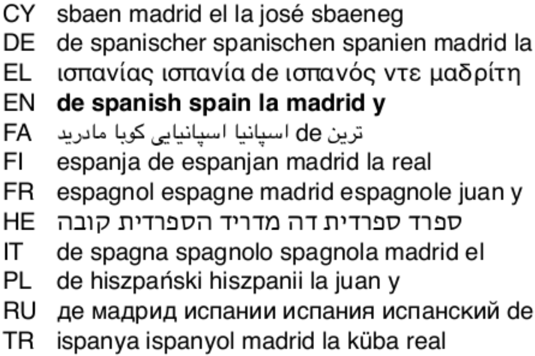
- Directly parallel translations are rare, expensive to produce
- Comparable corpora more common: e.g., Wikipedia, web pages
    - Our data set: all Wikipedia articles in English, Farsi, Finnish, French, German, Greek, Hebrew, Italian, Polish, Russian, Turkish, Welsh
- Documents are topically similar but not direct translations
- More interesting questions, more real-world applications:
    - Do comparable document tuples support alignment of topics?
    - Do different languages have different perspectives?
    - Which topics do particular languages focus on?

Wikipedia: Example Topics ($T = 400$)



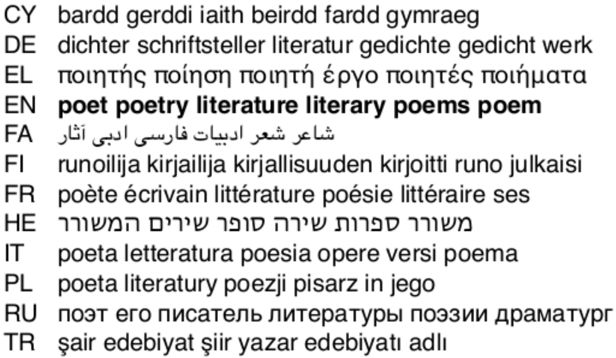
CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضايی ماموريت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

Wikipedia: Example Topics ($T = 400$)



CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

Wikipedia: Example Topics ($T = 400$)

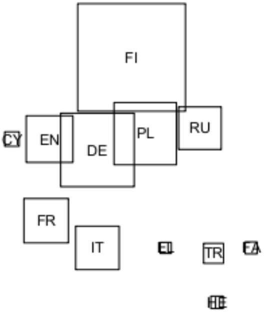


CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

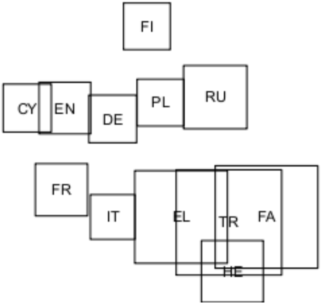
Topic Divergence Between Languages

- Estimate document-specific distributions over topics
- Compute Jensen-Shannon divergence between documents in a tuple
- Average document-document divergences for each language pair:
  - "Disagreement" score for each language pair
- Almost all pairs have divergences consistent with EuroParl, even languages that have historically been in conflict
- Although individual articles may have high between-language divergence, Wikipedia is on average consistent between languages

Differences in Topic Emphasis



world ski km won...



ottoman empire khan byzantine...

Conclusions

- Can discover topics aligned across multiple languages
- A small number of aligned documents is sufficient to align topics
- Can use the model to create bilingual lexica and find translations
- For comparable corpora, e.g., Wikipedia, someone who speaks *any one* language can perform data-driven analysis of topic trends, similarities and differences in *all* available languages
- Future work: adapting machine translation and cross-language information retrieval systems to new domains

Questions?

wallach@cs.umass.edu

http://www.cs.umass.edu/~wallach/