Opportunities and Challenges in Working with Low-Resource Languages

Yulia Tsvetkov

Language Technologies Institute
Carnegie Mellon University

June 22, 2017

Carnegie Mellon University

Plan for Part 1

1. What is low-resource NLP?
2. Why low-resource NLP is hard?
3. Why care about low-resource languages?
4. Why standard techniques used in NLP cannot simply be applied to low-resource languages?
5. Approaches to low-resource NLP

What does an NLP system need to "know"?

- Language consists of many levels of structure
- Humans fluently integrate all of these in producing/understanding language
- Ideally, so would a computer!

Sounds

SOUNDS Th i a si e n

Words

WORDS This is a simple sentence

Example from Nathan Schneider

Morphology

WORDS This is a simple sentence

MORPHOLOGY

be
3sg
present

Example from Nathan Schneider

Parts of Speech

PART OF SPEECH	DT	VBZ	DT	JJ	NN
WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

Example from Nathan Schneider

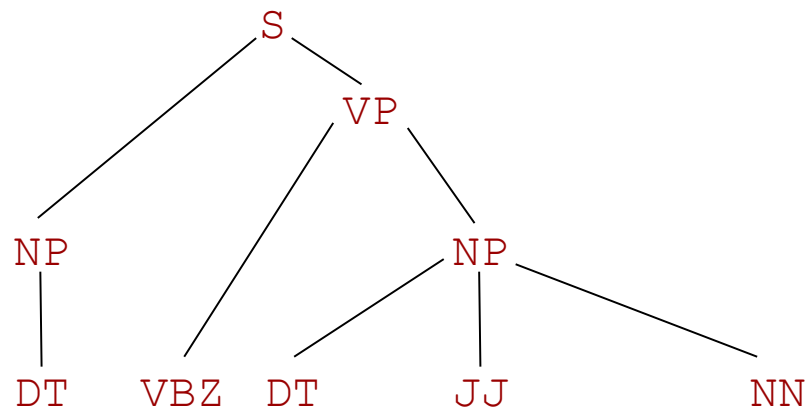
Syntax

SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY



Example from Nathan Schneider

Semantics

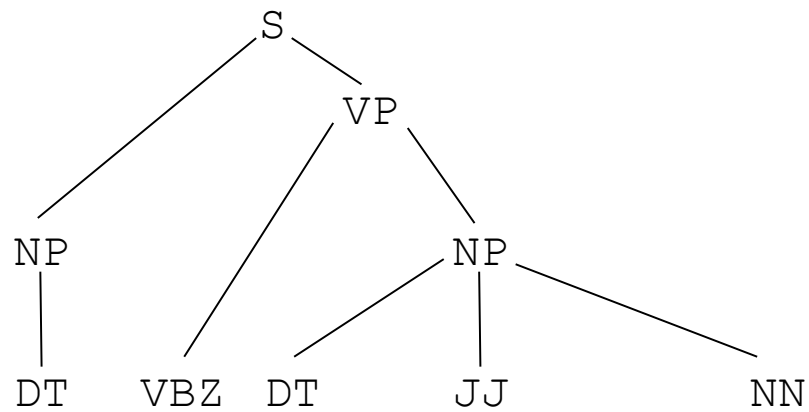
SYNTAX

PART OF SPEECH

WORDS

MORPHOLOGY

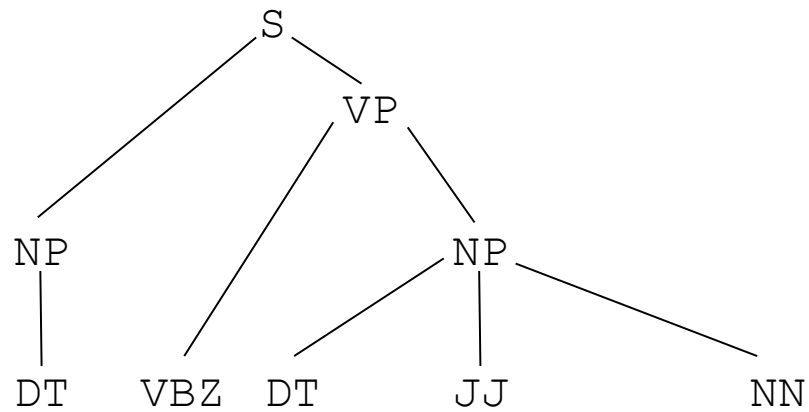
SEMANTICS



Example from Nathan Schneider

Discourse

SYNTAX



PART OF SPEECH

WORDS This is a simple sentence

MORPHOLOGY

be
3sg
present

SIMPLE1
having
few
parts

SENTENCE1
string of words
satisfying the
grammatical rules
of a language

SEMANTICS

CONTRAST

DISCOURSE But it is an instructive one.

Example from Nathan Schneider

Natural Language Processing Tasks

Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- ...

Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Why NLP is hard?

1. **Ambiguily** at many levels:
   - Word senses: bank (finance or river?)
   - Part of speech: chair (noun or verb?)
   - Syntactic structure: I saw a man with a telescope
   - Quantifier scope: Every child loves some movie
   - Multiple: I saw her duck

   $\Rightarrow$ NLP algorithms model ambiguity, and choose the correct analysis in context
2. Linguistic diversity

Why NLP is hard?

1. **Ambiguily** at many levels:
   - Word senses: bank (finance or river?)
   - Part of speech: chair (noun or verb?)
   - Syntactic structure: I saw a man with a telescope
   - Quantifier scope: Every child loves some movie
   - Multiple: I saw her duck

   $\Rightarrow$ NLP algorithms model ambiguity, and choose the correct analysis in context
2. **Linguistic diversity**

Linguistic Diversity: 6-7K World Languages

Linguistic Diversity: Words

这是一个简单的句子

WORDS This is a simple sentence

זה משפט פשוט

Linguistic Diversity: Hebrew Words

in tea her daughter	בתה

- most of the vowels unspecified

Linguistic Diversity: Words

in tea	בתה
in the tea	בהתה
that in tea	שבתה
that in the tea	שבהתה
and that in the tea	ושבהתה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them

Linguistic Diversity: Words

ושבתה

and her saturday	ו+שבת+ה
and that in tea	ו+ש+ב+תה
and that her daughter	ו+ש+בת+ה

- most of the vowels unspecified
- particles, prepositions, the definite article, conjunctions attach to the words which follow them
- tokenization is highly ambiguous

Linguistic Diversity:Morphology

WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

Linguistic Diversity: Quechua Morphology

Much'ananayakapushasqakupuniñataqsunamá

Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

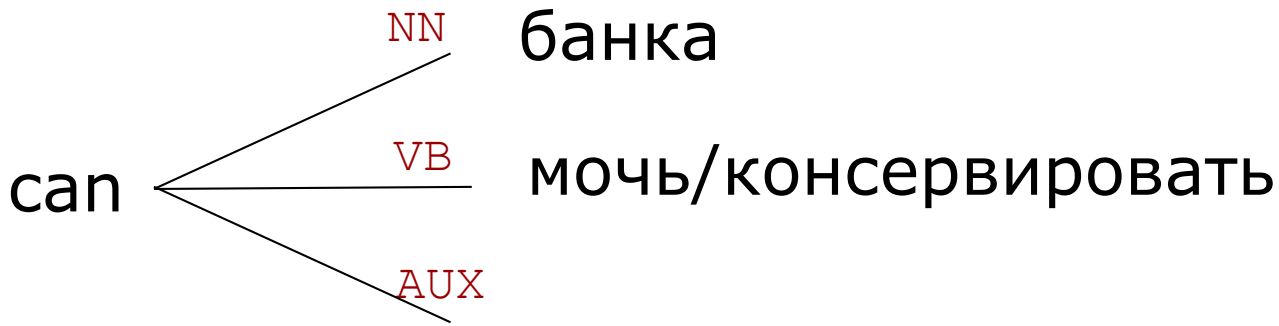
"So they really always have been kissing each other then"

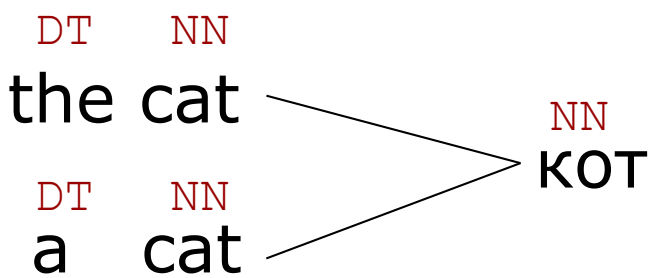
```
Much'a  to kiss
-na     expresses obligation, lost in translation
-naya   expresses desire
-ka     diminutive
-pu     reflexive (kiss *eachother*)
-sha    progressive (kiss*ing*)
-sqa    declaring something the speaker has not personally witnessed
-ku     3rd person plural (they kiss)
-puni   definitive (really*)
-ña     always
-taq    statement of contrast (...then)
-suna   expressing uncertainty (So...)
-má     expressing that the speaker is surprised
```

Linguistic Diversity: Russian Morphology

	Singular+neut	Plural+neut	
Nominative	предложение	предложения	sentence (s)
Genitive	предложения	предложений	(of) sentence (s)
Dative	предложению	предложениям	(to) sentence (s)
Accusative	предложение	предложения	sentence (s)
Instrumental	предложением	предложениями	(by) sentence (s)
Prepositional	предложении	предложениях	(in/at) sentence (s)

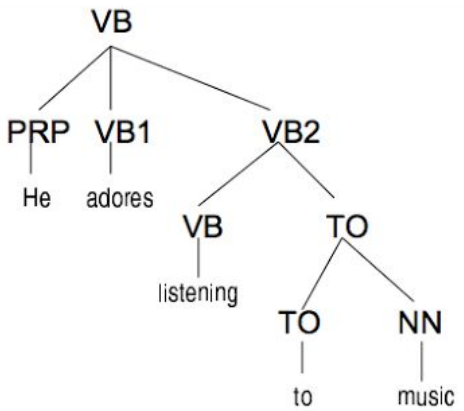
Linguistic Diversity: Parts of Speech



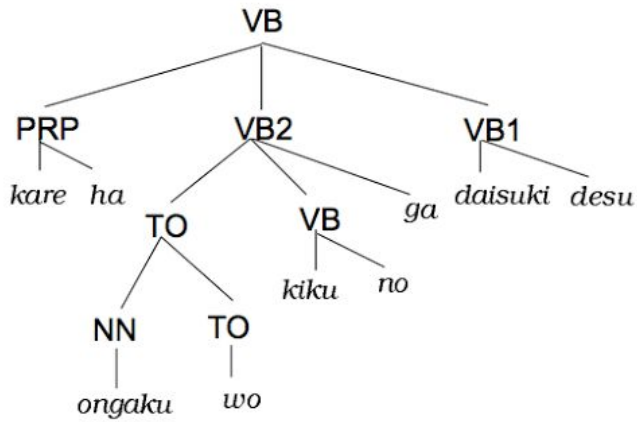


Linguistic Diversity: Japanese Syntax

SVO

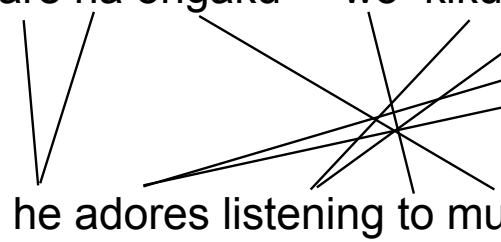


SOV



he adores listening to music

かれ は おんがく を きく の が だいすき です
kare ha ongaku wo kiku no ga daisuki desu



(Yamada & Knight '02)

Linguistic Diversity: Semantics

Every language describes the world in a different way,
for example, it depends on culture or historical conditions.

- Russian has relatively few names for colors; Japanese has hundreds
- Multiword expressions, e.g. it's raining cats and dogs or wake up and metaphors, e.g. Love is a journey are very different across languages

Sapir-Whorf Hypothesis:
the language we speak both affects and reflects our view of the world

Linguistic Diversity: Language Families

www.ethnologue.com

1. Niger–Congo (1,538 languages) (20.6%)
2. Austronesian (1,257 languages) (16.8%)
3. Trans–New Guinea (480 languages) (6.4%)
4. Sino-Tibetan (457 languages) (6.1%)
5. Indo-European (444 languages) (5.9%)
6. Australian (378 languages) (5.1%)
7. Afro-Asiatic (375 languages) (5.0%)
8. Nilo-Saharan (205 languages) (2.7%)
9. Oto-Manguean (177 languages) (2.4%)
10. Austroasiatic (169 languages) (2.3%)
11. Volta Congo (108 languages) (1.5%)
12. Tai–Kadai (95 languages) (1.3%)
13. Dravidian (85 languages) (1.1%)
14. Tupian (76 languages) (1.0%)

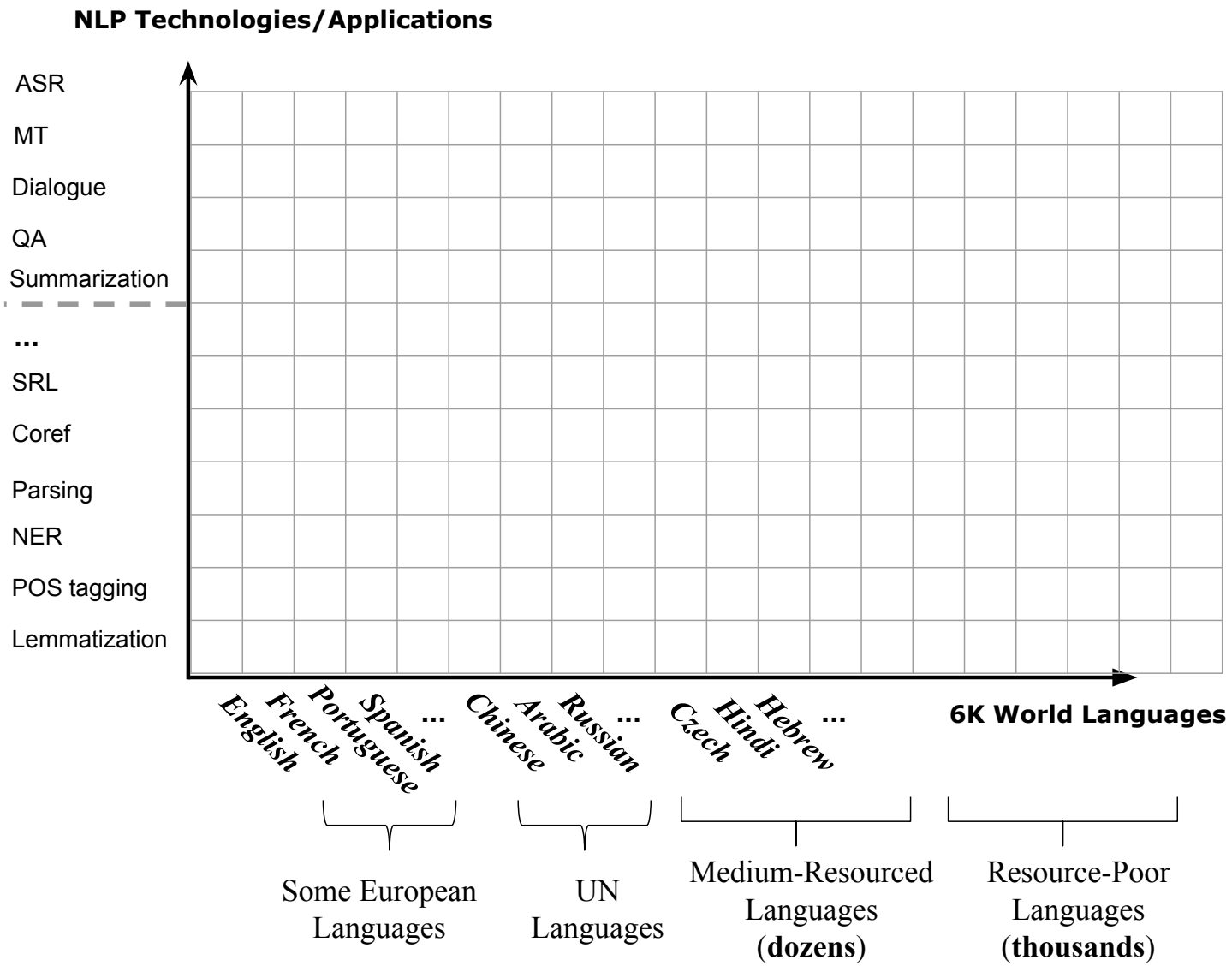
Why NLP is hard?

1. **Ambiguily** at many levels:
   - Word senses: bank (finance or river?)
   - Part of speech: blue (noun or verb?)
   - Syntactic structure: I saw a man with a telescope
   - Quantifier scope: Every child loves some movie
   - Multiple: I saw her duck

   $\Rightarrow$ NLP algorithms model ambiguity, and choose the correct analysis in context
2. **Linguistic diversity**
   - 6–7K languages in the world, $>$ 14 language families
   - Languages diverge across all levels of linguistic structure
     $\Rightarrow$ no generic solution for a particular NLP task
   - Most of the languages do not have sufficient resources to build statistical NLP models

Low-resource languages – languages lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications

What NLP Technologies are Resource-Rich?



Performance of Resource-Rich vs. Resource-Poor NLP

Machine Translation

- Parallel corpus

Nenhum deles reparou na janela , através da qual teria podido ver uma enorme coruja amarelada , esvoaçando em grande alvoroço .	None of them noticed a large , tawny owl flutter past the window .
assim , não viu as corujas descendo rapidamente em plena luz do dia , apesar de todos os transeuntes apontarem estarrecidos e de boca aberta enquanto coruja após coruja lhes passavam A grande velocidade sobre as cabeças .	He didn ' t see the owls swoop ing past in broad daylight , though people down in the street did ; They pointed and gazed open-mouthed as owl after owl sped overhead .
Queira enviar-nos A sua coruja até dia 31 de Julho , sem falta .	We await your owl by no later than July 31 .
- O que é que quer dizer esperarem A minha coruja ?	after a few minutes He stammered , " what does it mean , They await My owl ?
Hagrid Hagrid enrolou A nota , deu-a à coruja que A agarrou com O bico e , dirigindo-se à porta , soltou A ave no meio da tempestade .	Hagrid Hagrid rolled up the note , gave it to the owl , which clamped it in its beak , went to the door , and threw the owl out into the Storm .
O próprio Hagrid adormecera no sofá totalmente destruído e , bicando no vidro da janela , estava uma coruja que segurava um jornal .	the hut was full of sunlight , the Storm was over , Hagrid himself was asleep on the collapsed sofa , and there was an owl rapping its claw on the window , a newspaper held in its beak .
A coruja entrou e depôs O jornal em cima de Hagrid	the owl swooped in and dropped the newspaper on top of Hagrid , who didn ' t

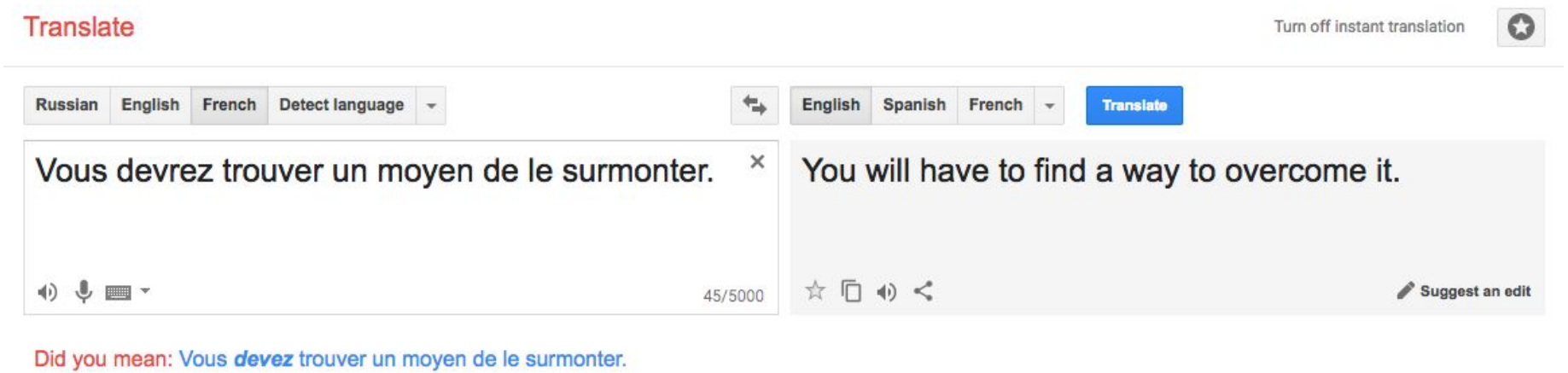
- Resource-rich: millions of parallel sentences
- Resource-poor: few thousands of parallel sentences

Performance of Resource-Rich vs. Resource-Poor NLP

Machine Translation

English → French

French → English

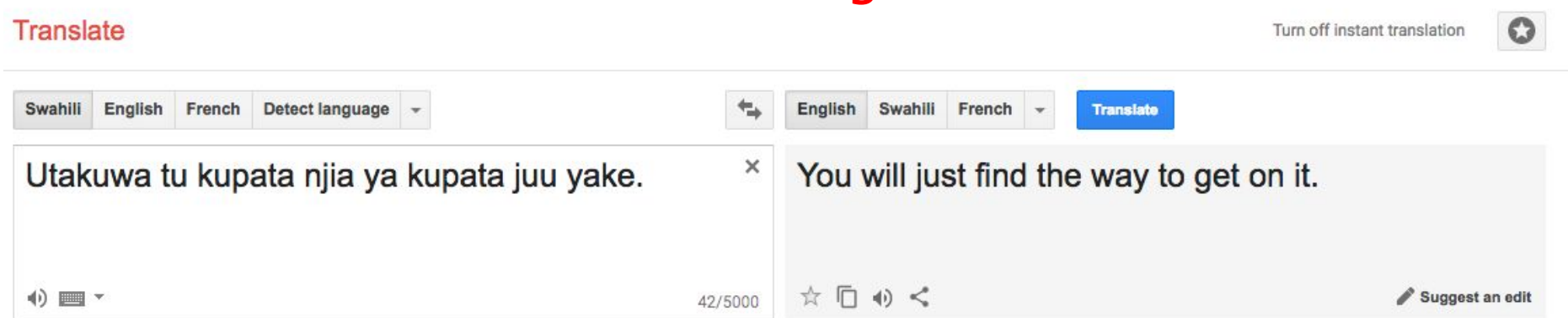


Performance of Resource-Rich vs. Resource-Poor NLP

Machine Translation

English → Swahili

Swahili → English

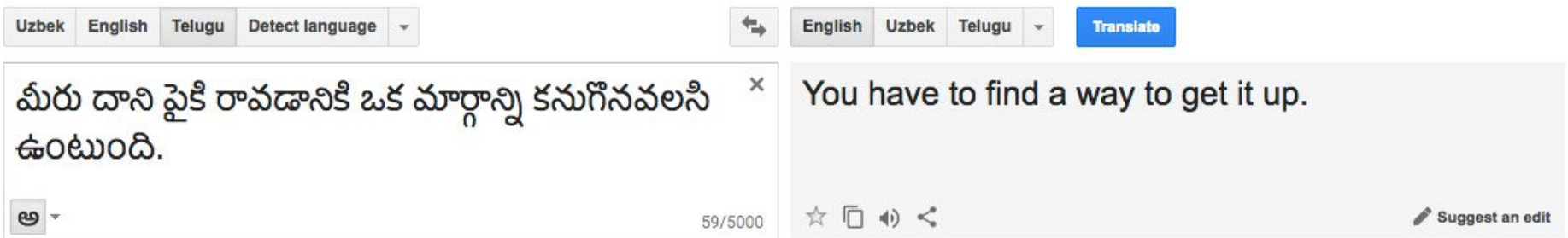


Performance of Resource-Rich vs. Resource-Poor NLP

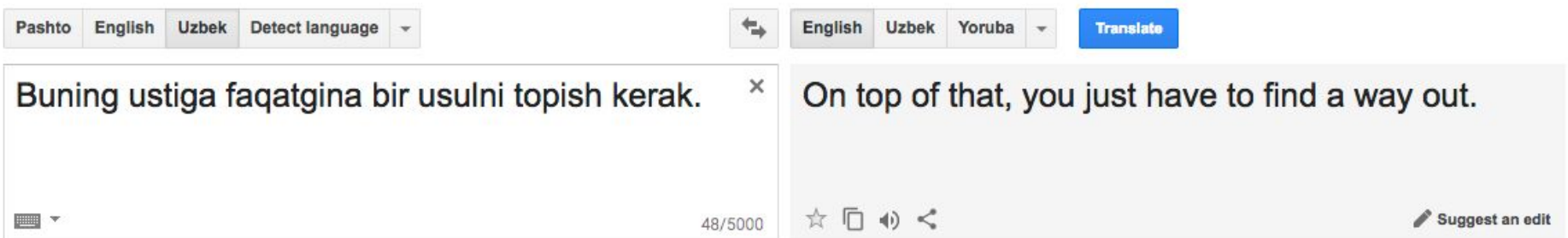
Machine Translation

English → Hindi → English

English → Telugu → English



English → Uzbek → English

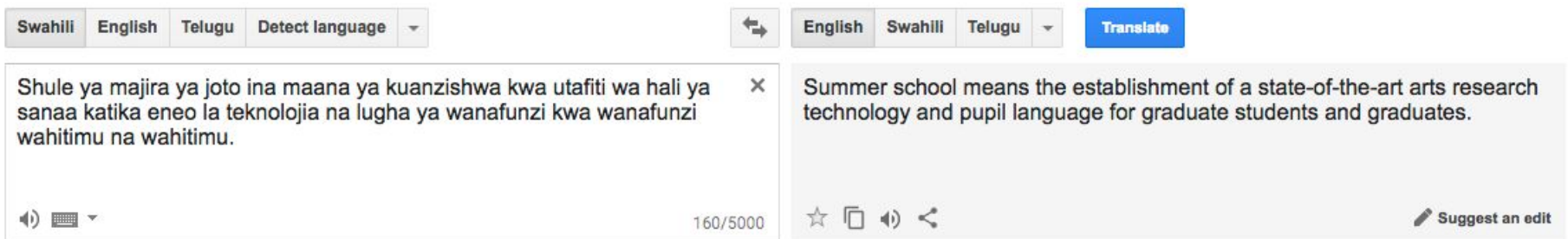


Performance of Resource-Rich vs. Resource-Poor NLP

Machine Translation

English → Swahili

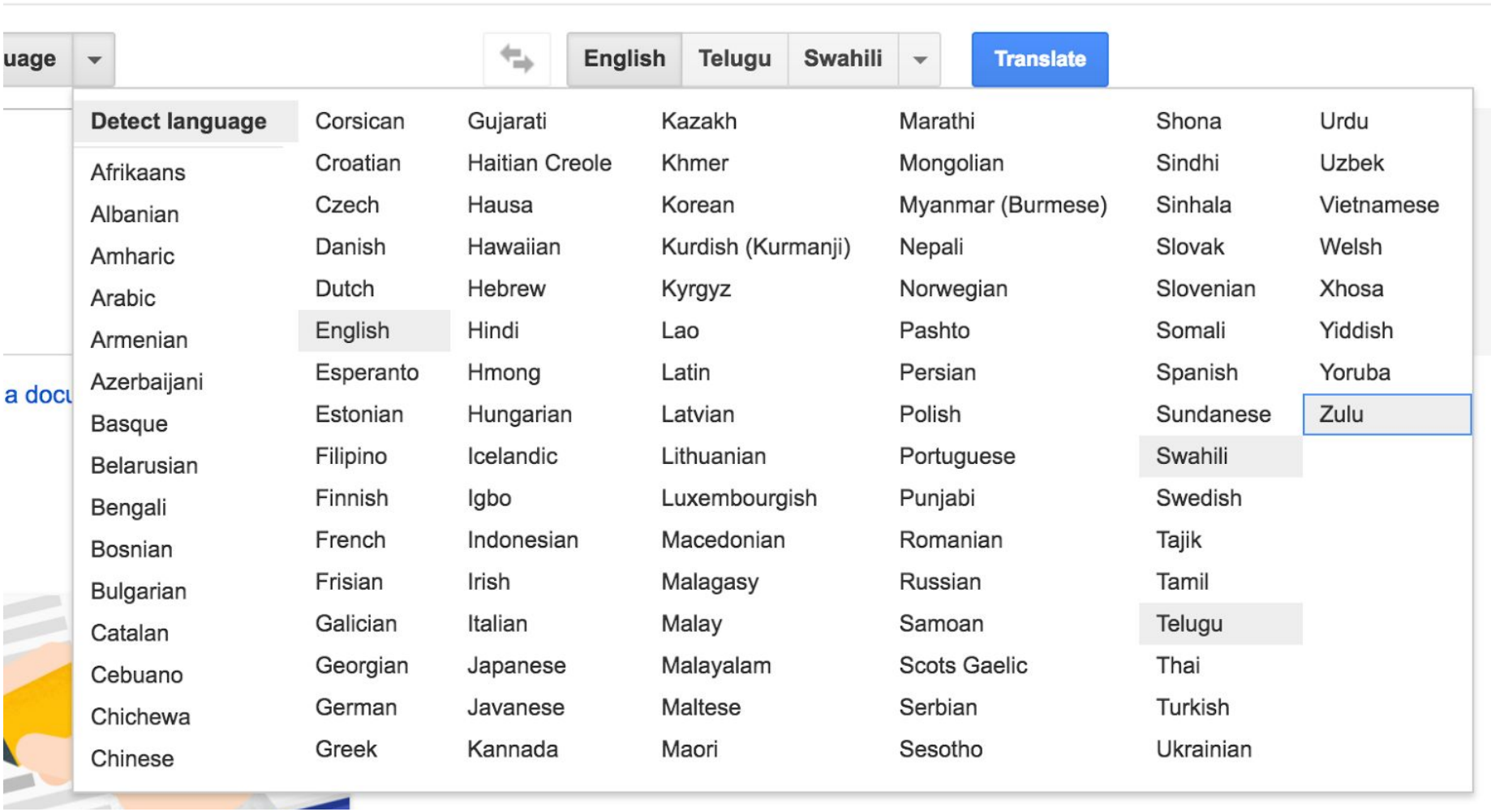
Swahili → English



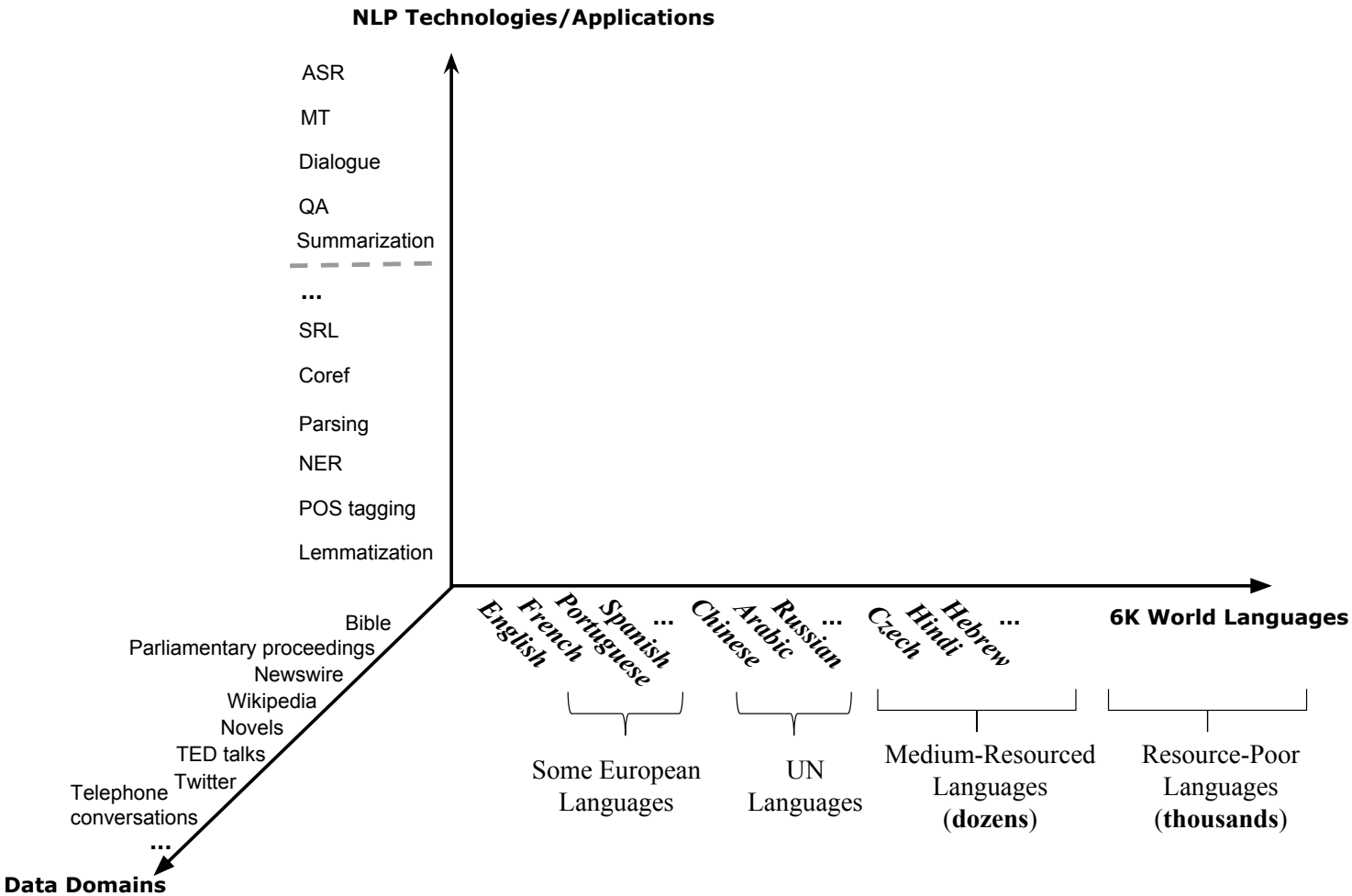
Performance of Resource-Rich vs. Resource-Poor NLP

Machine Translation

- About 100 out of 6K languages

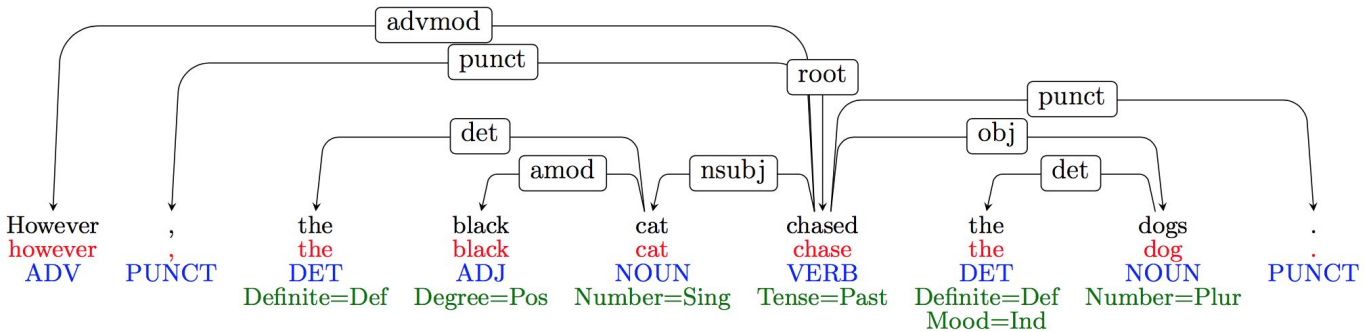


Low-Resource NLP is Not Only About Multilinguality



Low-Resource NLP is Not Only About Multilinguality

- Parsing models trained using WSJ treebanks do not work for spoken language domain



- Spoken language is riddled with verbal disfluencies that interrupt the flow of speech, including long pauses, repeated words or phrases, restarts, and revisions of content:

Um, the black the black cat ch- chased the dogs.

Low-Resource NLP is Not Only About Multilinguality

- Twitter processing is hard

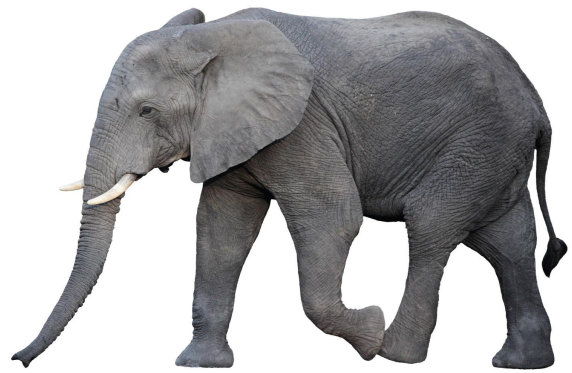
Low-Resource NLP is Not Only About Multilinguality

- Much of the world knowledge is not in text, corpora contain what people said, but not what they meant, or how they understood things, or what they did in response to the language

This is milk

?

<



Plan for Part 1

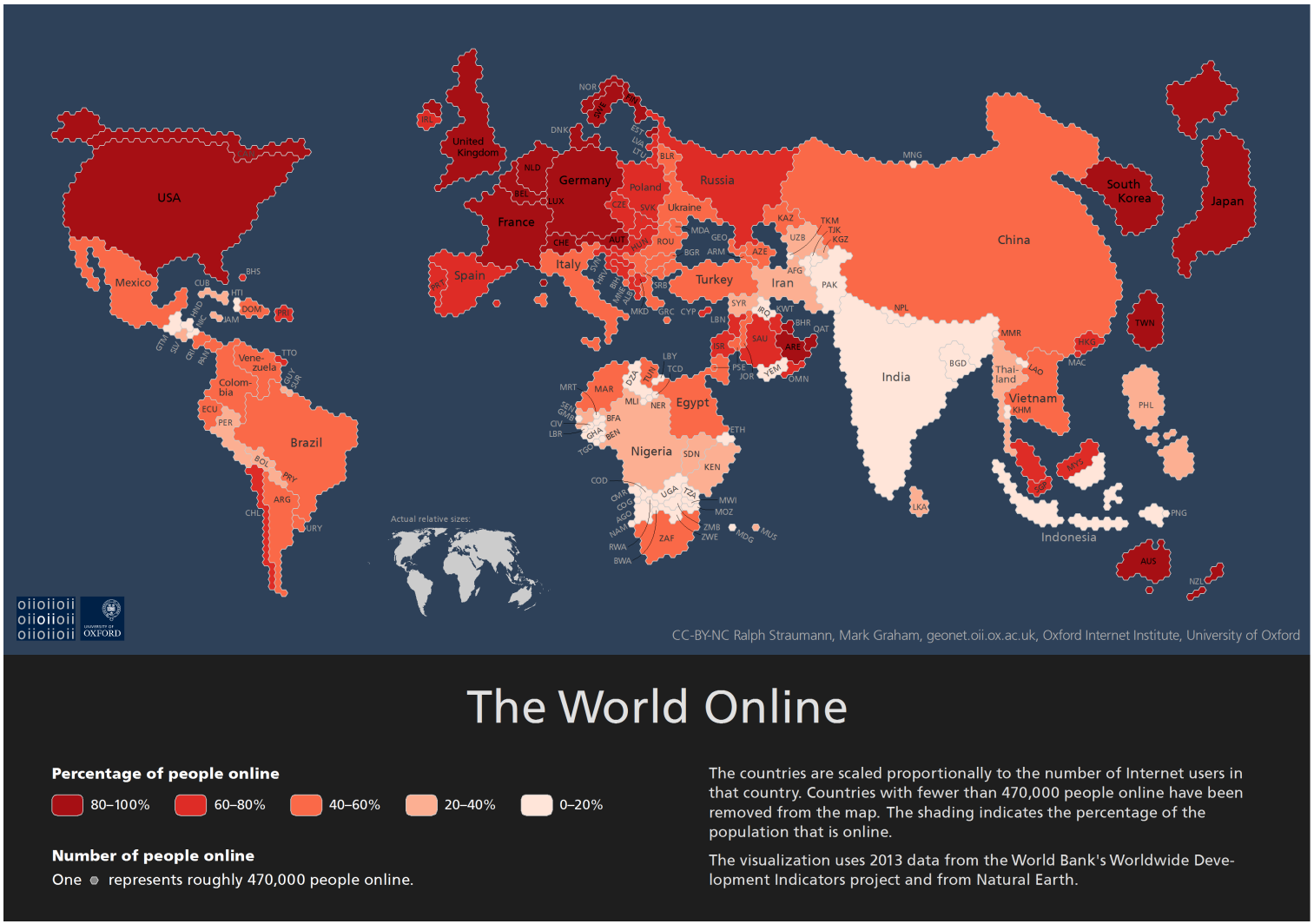
- ✓ What is low-resource NLP?
- ✓ Why low-resource NLP is hard?
- Why care about low-resource languages?
- Why standard techniques used in NLP cannot simply be applied to low-resource languages?
- Approaches to low-resource NLP

Why Care About Low-Resource NLP?

1. Commercial value
2. Social-good reasons

Why Care About Low-Resource NLP?

Commercial value



Why Care About Low-Resource NLP?

Africa

Africa is a continent with a very high linguistic diversity:
there are an estimated 1.5-2K African languages from 6 language families.

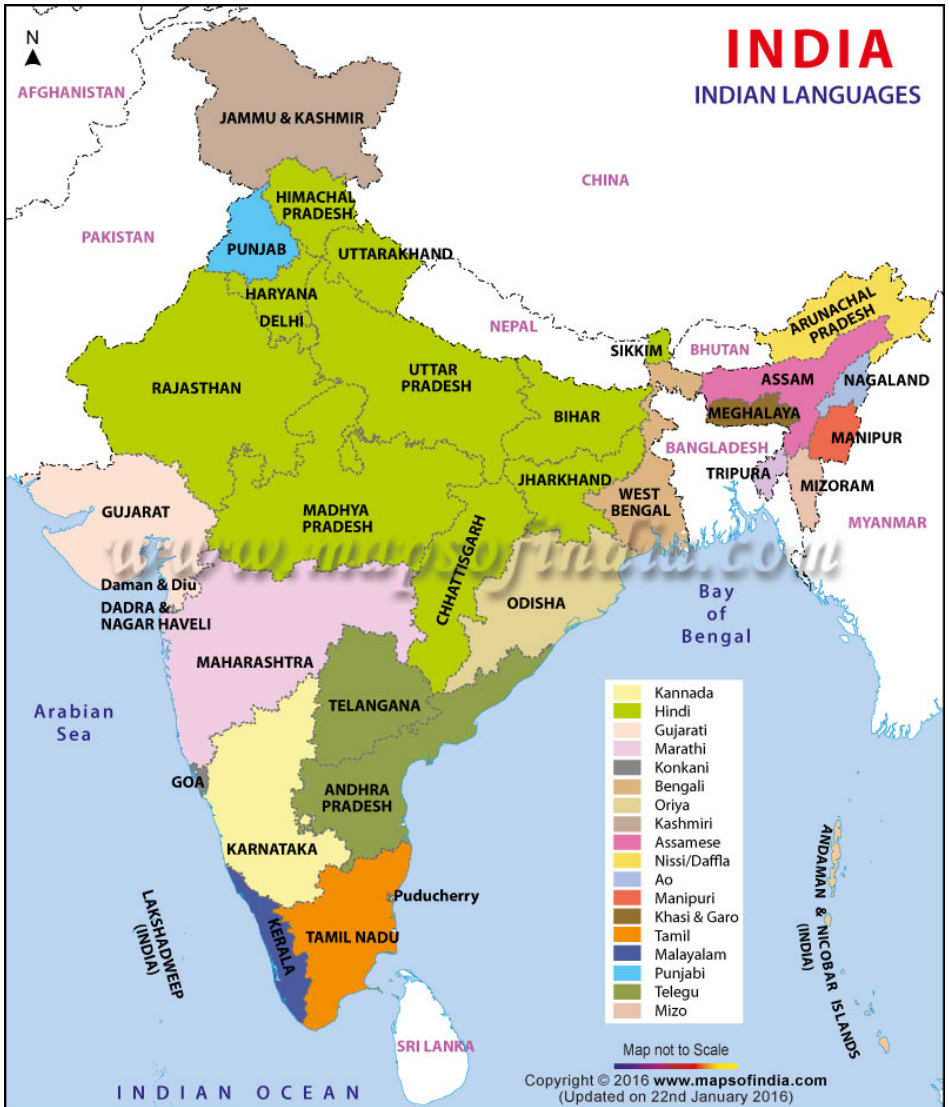
- 1.2 billion people

Why Care About Low-Resource NLP?

India

There are about 460 languages in India.

- 1.3 billion people



Why Care About Low-Resource NLP?

Social good reasons

- Translation systems
- Speech interfaces
- Dialogue systems
- Educational applications
- Emergency response applications
- Monitoring democratic processes

Why Care About Low-Resource NLP?

Social good reasons: Emergency response

- About 3 million people were affected by the quake

* Slide by Rob Munro `http://web.stanford.edu/class/cs124`

Why Care About Low-Resource NLP?

Social good reasons: Emergency response

Messages start streaming in

- Fanmi mwen nan Kafou, 24 Cote Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31

IDIBON

* Slide by Rob Munro `http://web.stanford.edu/class/cs124`

Why Care About Low-Resource NLP?

Social good reasons: Emergency response

Messages start streaming in

- Fanmi mwen nan Kafou, 24 Cote Plage, 41A bezwen manje ak dlo
- Moun kwense nan Sakre Kè nan Pòtoprens
- Ti ekipman Lopital General genyen yo paka minm fè 24 è
- Fanm gen tranche pou fè yon pitit nan Delmas 31
- My family in Carrefour, 24 Cote Plage,41A needs food and water
- People trapped in Sacred Heart Church, PauP
- General Hospital has less than 24 hrs. supplies
- Undergoing children delivery Delmas 31

IDIBON

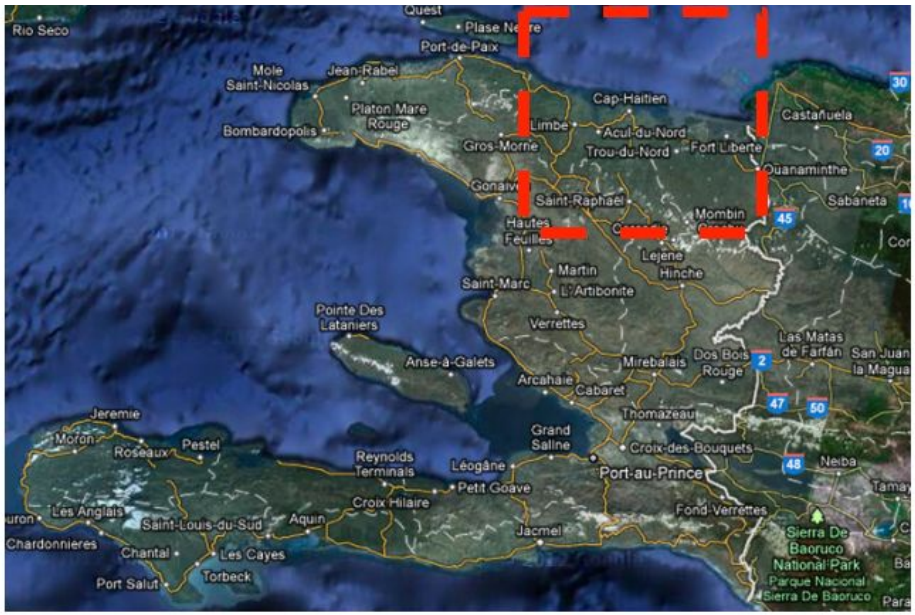
* Slide by Rob Munro `http://web.stanford.edu/class/cs124`

Why Care About Low-Resource NLP?

Social good reasons: Emergency response

Lopital Sacre-Coeur ki nan vil Okap, pre pou li resevwa moun malad e lap mande pou moun ki malad yo ale la.

"Sacre-Coeur Hospital which located in this village of Okap is ready to receive those who are injured. Therefore, we are asking those who are sick to report to that hospital."



IDIBON

* Slide by Rob Munro `http://web.stanford.edu/class/cs124`

Why Care About Low-Resource NLP?

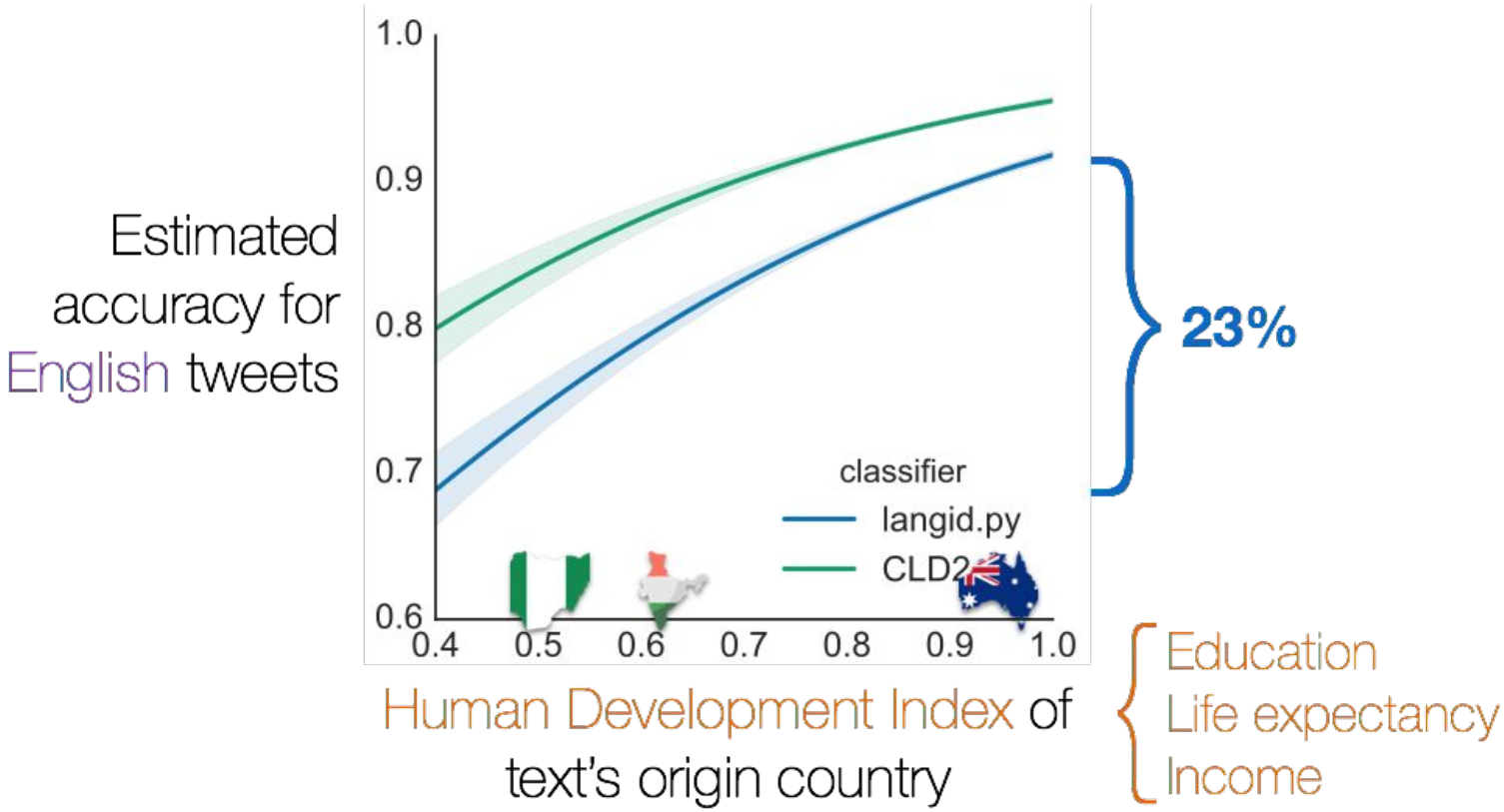
Social good reasons: Identifying outbreaks of diseases

Why Care About Low-Resource NLP?

Social good reasons: Identifying outbreaks of diseases

Why Care About Low-Resource NLP?

Social good reasons: Identifying outbreaks of diseases

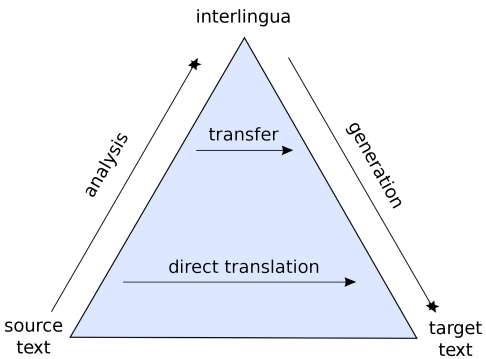


Plan for Part 1

- ✓ What is low-resource NLP?
- ✓ Why low-resource NLP is hard?
- ✓ Why care about low-resource languages?
- Why standard techniques used in NLP cannot simply be applied to low-resource languages?
- Approaches to low-resource NLP

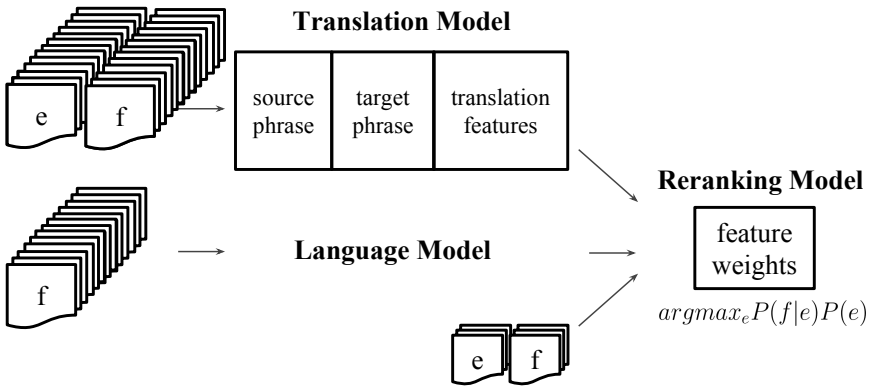
Paradigm Shifts in NLP

Logic-based/Rule-based NLP



~90s

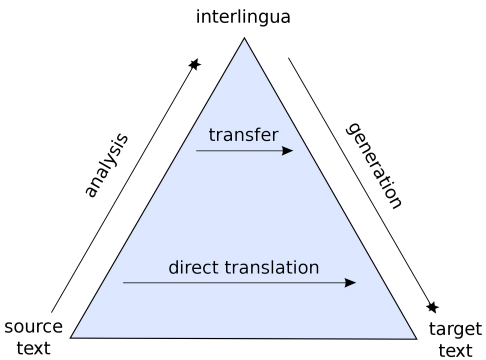
Statistical NLP



- Rule-based models: high precision but very low recall

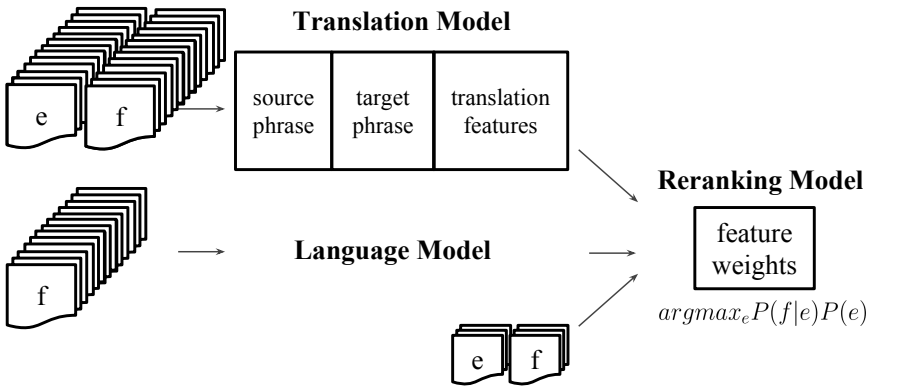
Paradigm Shifts in NLP

Logic-based/Rule-based NLP



<

Statistical NLP

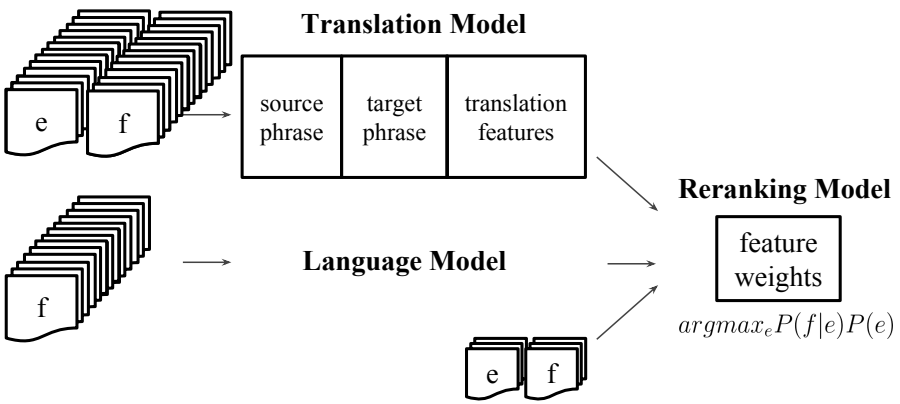


* In resource-rich settings

- Statistical models: robust in the face of real-world data
- Better performance
- Less engineering of hand-crafted rules/knowledge

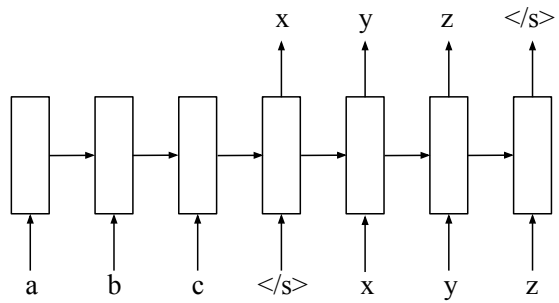
Paradigm Shifts in NLP

Statistical NLP



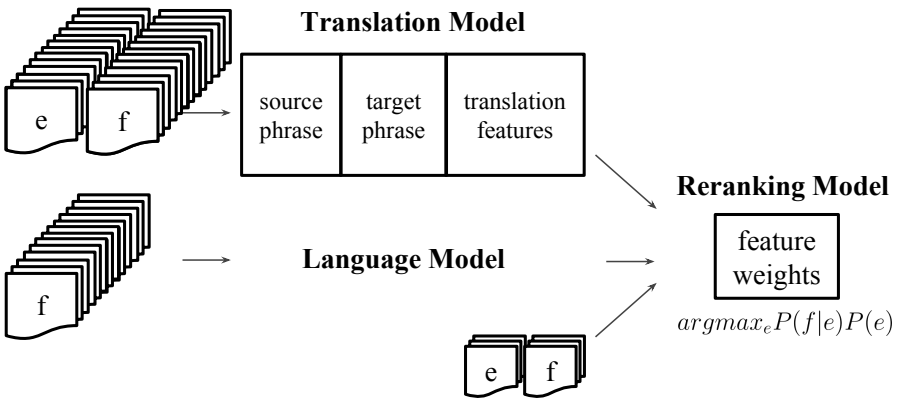
~mid 2010s

Statistical Neural NLP



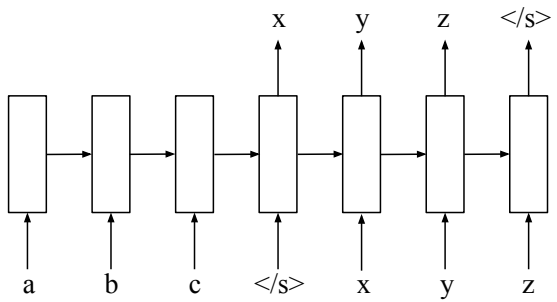
Paradigm Shifts in NLP

Statistical NLP



<

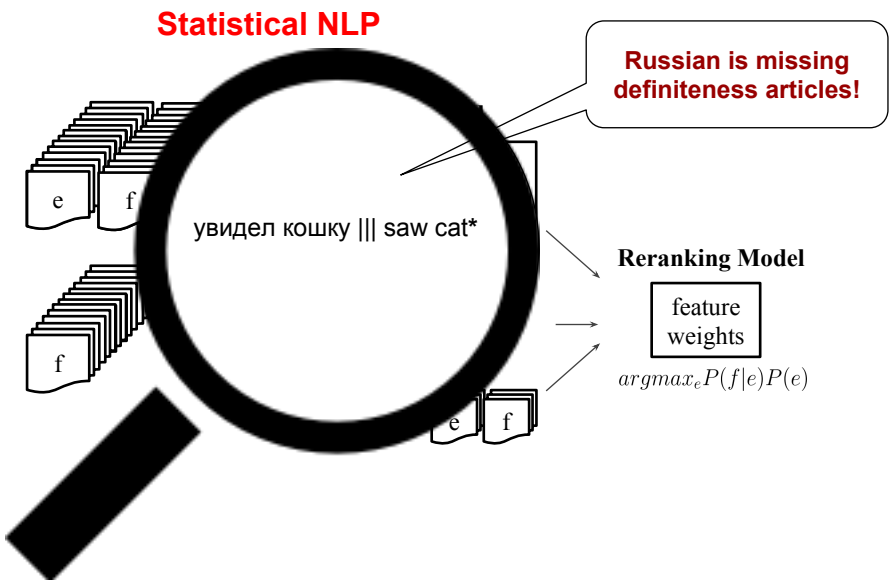
Statistical Neural NLP

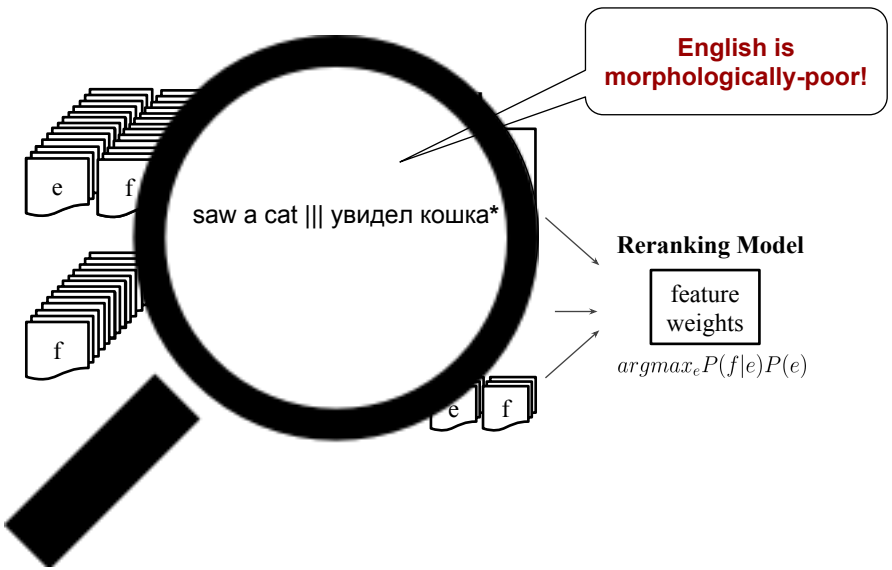


* In resource-rich settings

- Robustness in the face of real-world data
- Better performance
- Less engineering of hand-crafted rules/knowledge

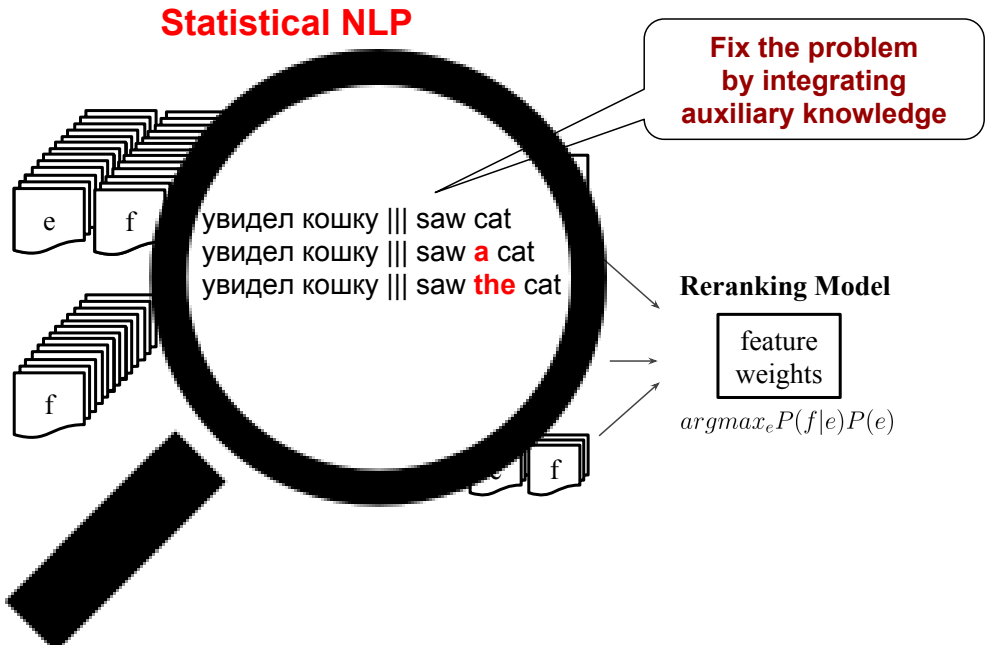
Price $$$





- But statistical models perform poorly in low-resource settings
- Models learn also linguistically-implausible generalizations

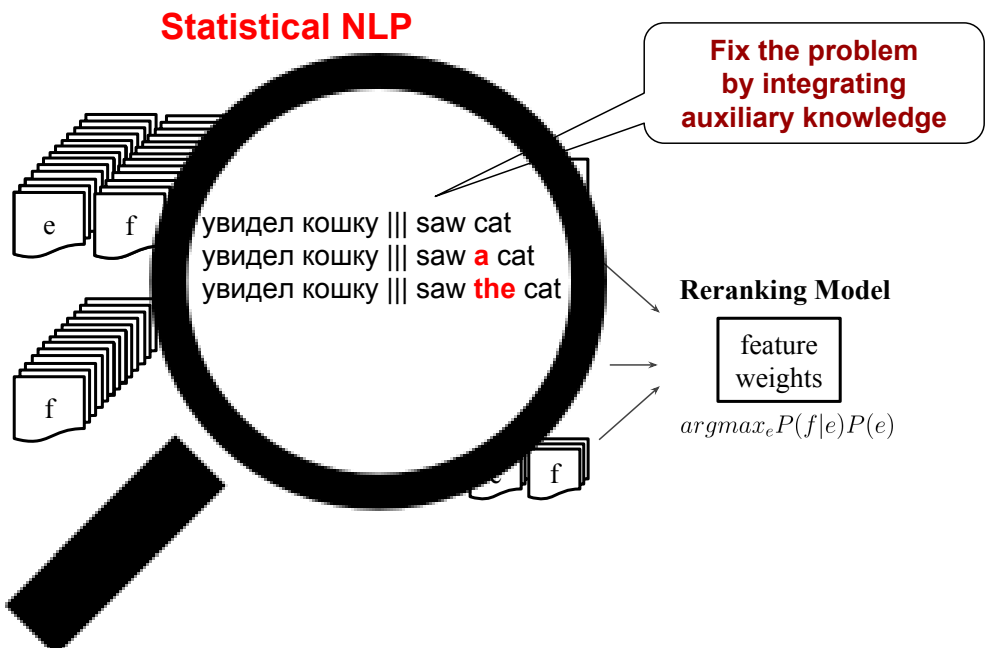
Hybrid Statistical NLP Models



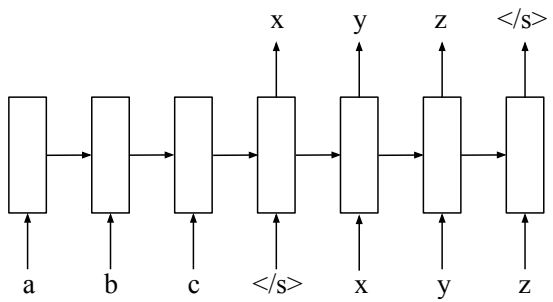
- To train high-quality models we need large amounts of training data
- We can partially compensate data scarcity with more sophisticated models that combine statistical learning with linguistic knowledge

Building Blocks in Conventional Statistical NLP Models

Words, phrases ⇒ easier analysis, easier adaptation

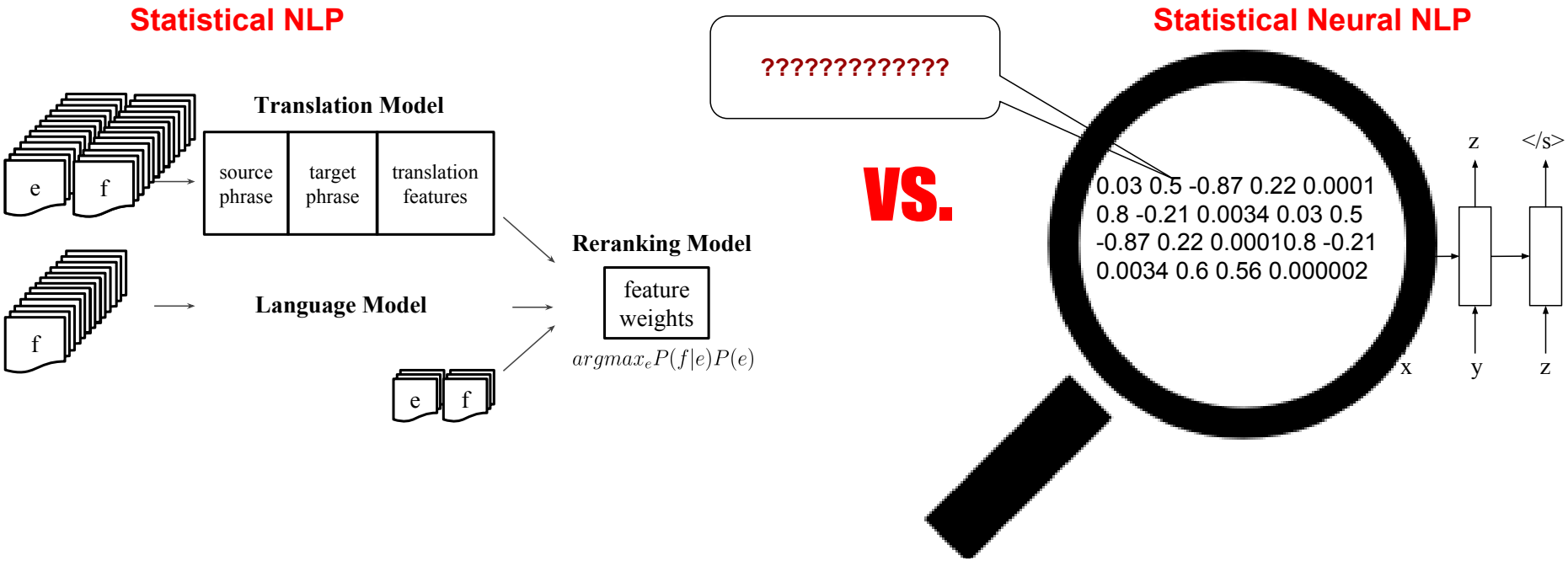


VS.



Building Blocks in Neural NLP Models

Vectors, matrices ⇒ not clear yet how to interpret



- How to interpret continuous representations?
- How to integrate auxiliary knowledge into neural network architectures?

Why standard techniques used in NLP cannot simply be applied to low-resource languages?

- State-of-the-art NLP models require large amounts of training data and/or sophisticated language-specific engineering
- Large amounts of training data are unavailable for most languages
  - an extreme case is languages that don't have a written form, e.g. Shanghainese spoken by 14 million people
  - or languages that just don't have online presence, e.g. Chichewa, a Bantu language spoken by 12 million people
- Language-specific engineering is expensive, requires linguistically trained speakers of the language

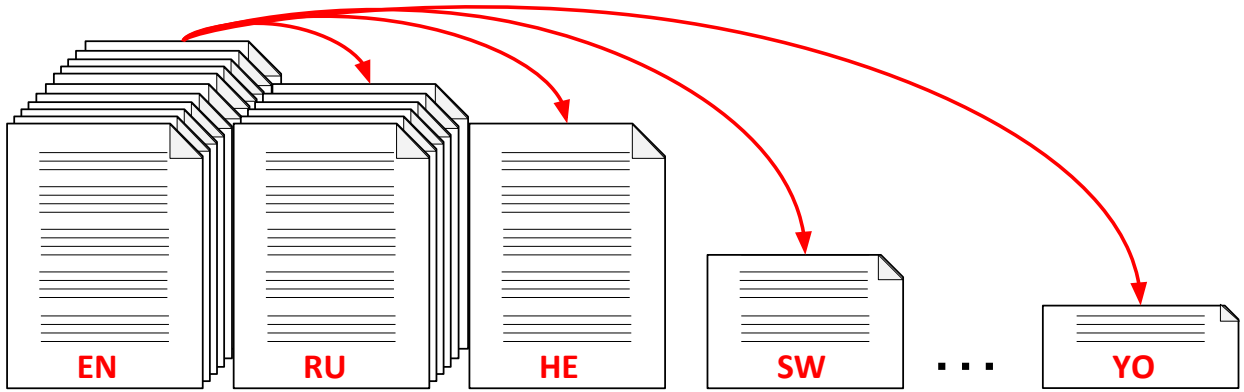
Plan for Part 1

- ✓ What is low-resource NLP?
- ✓ Why low-resource NLP is hard?
- ✓ Why standard techniques used in NLP cannot simply be applied to low-resource languages?
- ✓ Why care about low-resource languages?
- Approaches to low-resource NLP

Unsupervised Learning

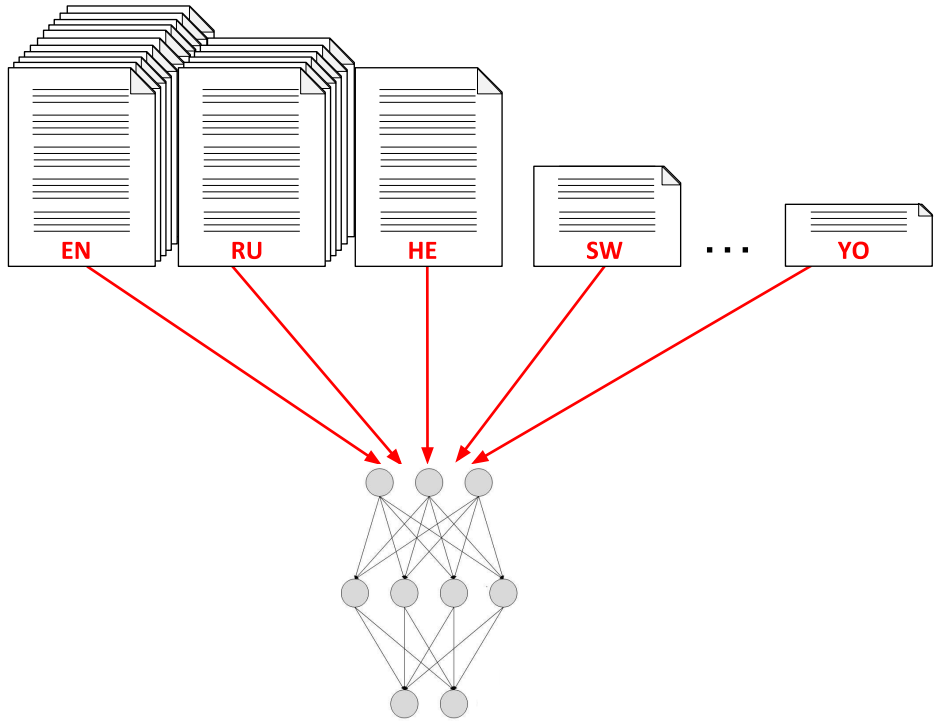
- Unsupervised feature induction: Brown clustering, Word vectors
- Unsupervised POS tagging
- Unsupervised dependency parsing
- ...

Transfer Learning or "Zero-Shot" Learning



- Cross-lingual transfer learning – transfer of resources and models from resource-rich source to resource-poor target languages
  - Transfer of annotations (e.g., POS tags, syntactic or semantic features) via cross-lingual bridges (e.g., word or phrase alignments)
  - Transfer of models – train a model in a resource-rich language and apply it in a resource-poor language
- Zero-shot learning – train a model in one domains and assume it generalizes more or less out-of-the-box in a low-resource domain
- One-shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it

Joint Multilingual or "Polyglot" Learning



- Joint resource-rich and resource-poor learning using a language-universal representation.
  - Convert data in all languages to a shared representation (e.g., phones or multilingual word vectors)
  - Train a single model on a mix of datasets in all languages, to enable parameter sharing where possible

Plan for Part 2

Case studies in low-resource NLP:

- Cross-lingual bridging via language lexicons
- Transfer learning
  - Projection of translations
  - Projection of syntactic annotations
- Polyglot models
  - Language modeling

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
barka
welalin
tack
spas
kia ora
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
dhanyavad
nanni
nandri
mauruuru
köszönöm
thank you
gracias
хвала
asante
manana
obrigada
murakoze
tenki
bedankt
bayarlalaa
gràcie
hvala
dankon
aciú
akun
djiere dieuf
tau
mochchakkeram
enkosi
dziękuję
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
obrigado
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
najis tuke
kam sah hamnida
terima kasih
tanemirt
rahmet
grazie
rahmat
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercé
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci