Ethics in Natural Language Processing – SS 2022

Lecture 7
Low-Resource NLP

TECHNISCHE UNIVERSITÄT DARMSTADT

Dr. Thomas Arnold
Aniket Pramanik


Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Slides and material from Yulia Tsvetkov

Carnegie Mellon University
Language Technologies Institute

Syllabus (tentative)

Nr.	Lecture
01	Introduction, Foundations I
02	Foundations II
03	Bias I
04	Bias II
05	Incivility and Hate Speech I
06	NO LECTURE – Christi Himmelfahrt
07	Incivility and Hate Speech II
08	Low-Resource NLP
09	NO LECTURE - Fronleichnam
10	Privacy and Security I
11	Privacy and Security II
12	Language of Manipulation I
13	Language of Manipulation II

Learning Goals

After hearing this lecture, you should be able to…

- **Explain why traditional NLP techniques fail on other languages**
- **Explain syntactic and semantic ambiguities**
- **Discuss the advantages and disadvantages of statistical neural NLP methods in Low-Resourse scenarios**

Outline

Low-Resource NLP: Introduction

Approaches to Low-Resource NLP

Low-Resource NLP: Lorelei Program

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

Morphology

TECHNISCHE UNIVERSITÄT DARMSTADT

WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

Parts of Speech

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

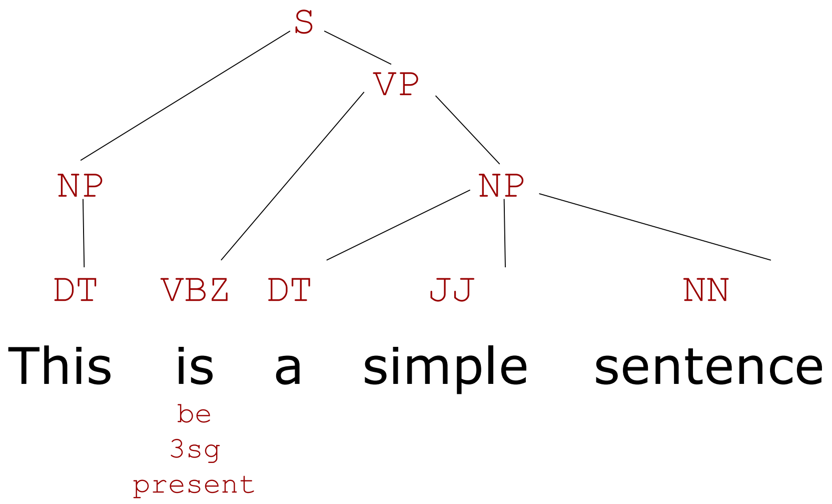
Syntax

SYNTAX

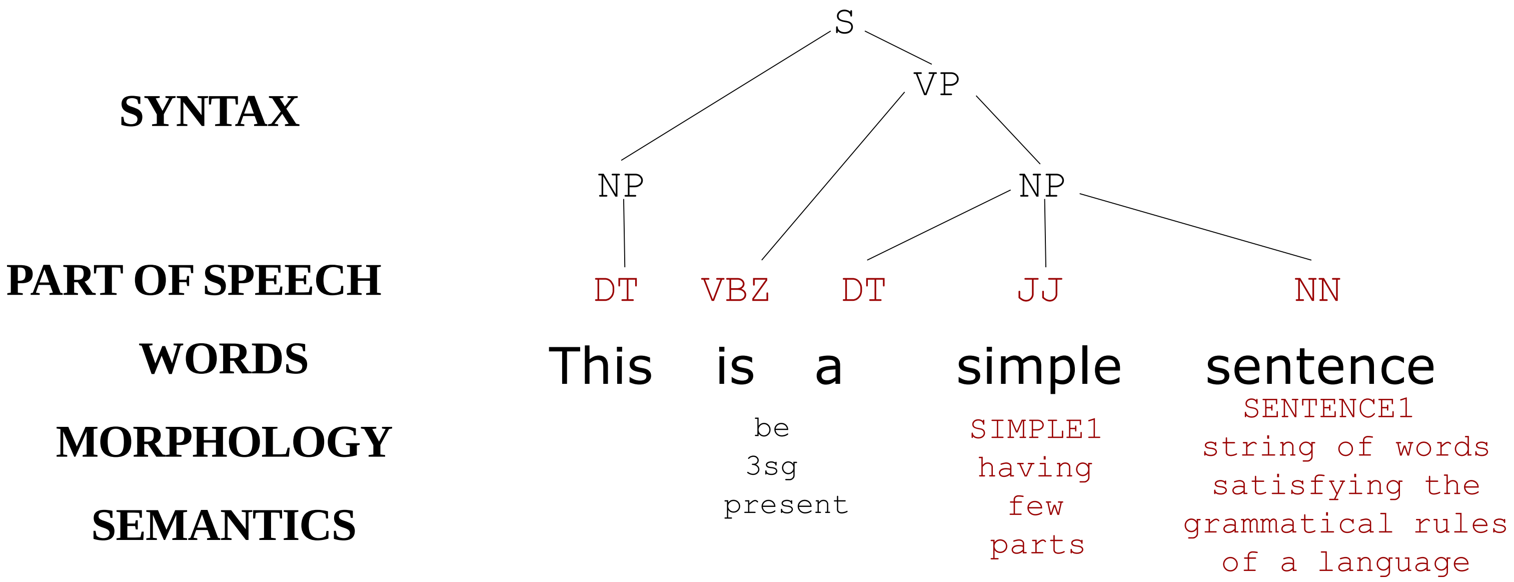
PART OF SPEECH

WORDS

MORPHOLOGY

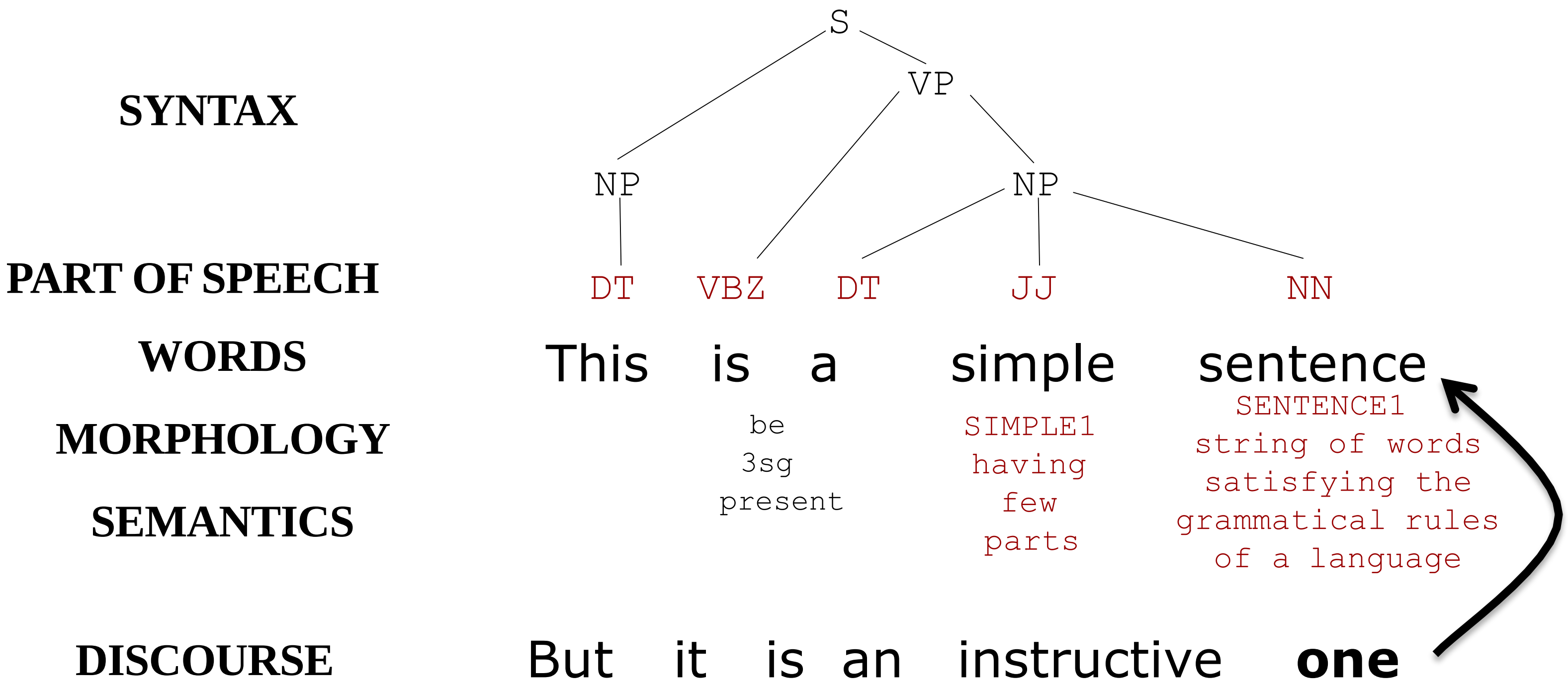


Semantics



SYNTAX
PART OF SPEECH
WORDS
MORPHOLOGY
SEMANTICS
S
VP
NP
NP
DT
VBZ
DT
JJ
NN
This
is
a
simple
sentence
be
3sg
present
SIMPLE1
having
few
parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a language

Discourse


S
VP
SYNTAX
NP
NP
PART OF SPEECH
DT
VBZ
DT
JJ
NN
WORDS
This
is
a
simple
sentence
MORPHOLOGY
be
3sg
present
SEMANTICS
SIMPLE1
having
few
parts
SENTENCE1
string of words
satisfying the
grammatical rules
of a language
DISCOURSE
But it is an instructive **one**

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?

TECHNISCHE
UNIVERSITÄT
DARMSTADT

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

   ⇒ 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

   ⇒ NLP algorithms model ambiguity, and choose the correct analysis in context

2. **Linguistic diversity**

6 – 7k World Languages

Linguistic Diversity: Words

TECHNISCHE UNIVERSITÄT DARMSTADT

这是一个简单的句子

WORDS This is a simple sentence

זה משפט פשוט

Linguistic Diversity: Hebrew Words

in tea   בתה
her daughter

- most of the vowels unspecified

Linguistic Diversity: Words

ושבתה

and her saturday	ו +שבת+ה
and that in tea	ו +ש+ב+התה
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

TECHNISCHE UNIVERSITÄT DARMSTADT

WORDS	This	is	a	simple	sentence
MORPHOLOGY		be 3sg present			

Linguistic Diversity: Quechua Morphology

TECHNISCHE UNIVERSITÄT DARMSTADT

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

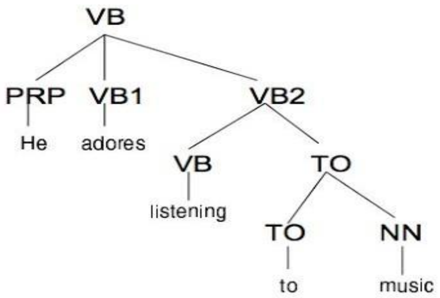
TECHNISCHE UNIVERSITÄT DARMSTADT

	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: Japanese Syntax

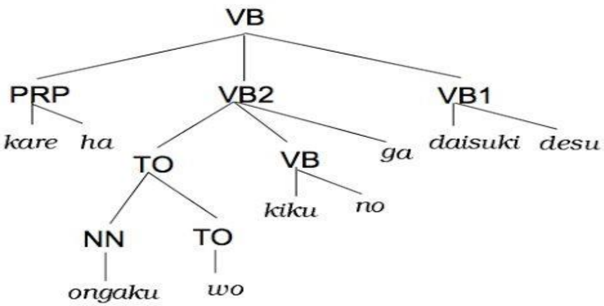
TECHNISCHE UNIVERSITÄT DARMSTADT

SVO



he adores listening to music

SOV

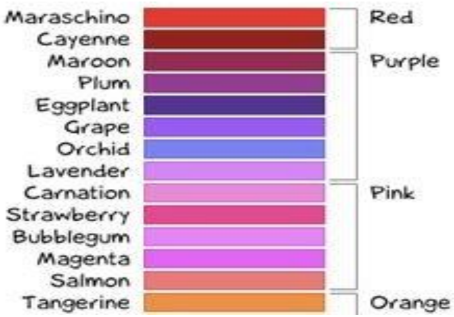


かれ は おんがく を きく の が だいすき です
kare ha ongaku wo kiku no ga daisuki desu

he adores listening to music

(Yamada & Knight ‘02)

Linguistic Diversity: Semantics



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

Why NLP is hard?

TECHNISCHE UNIVERSITÄT DARMSTADT

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

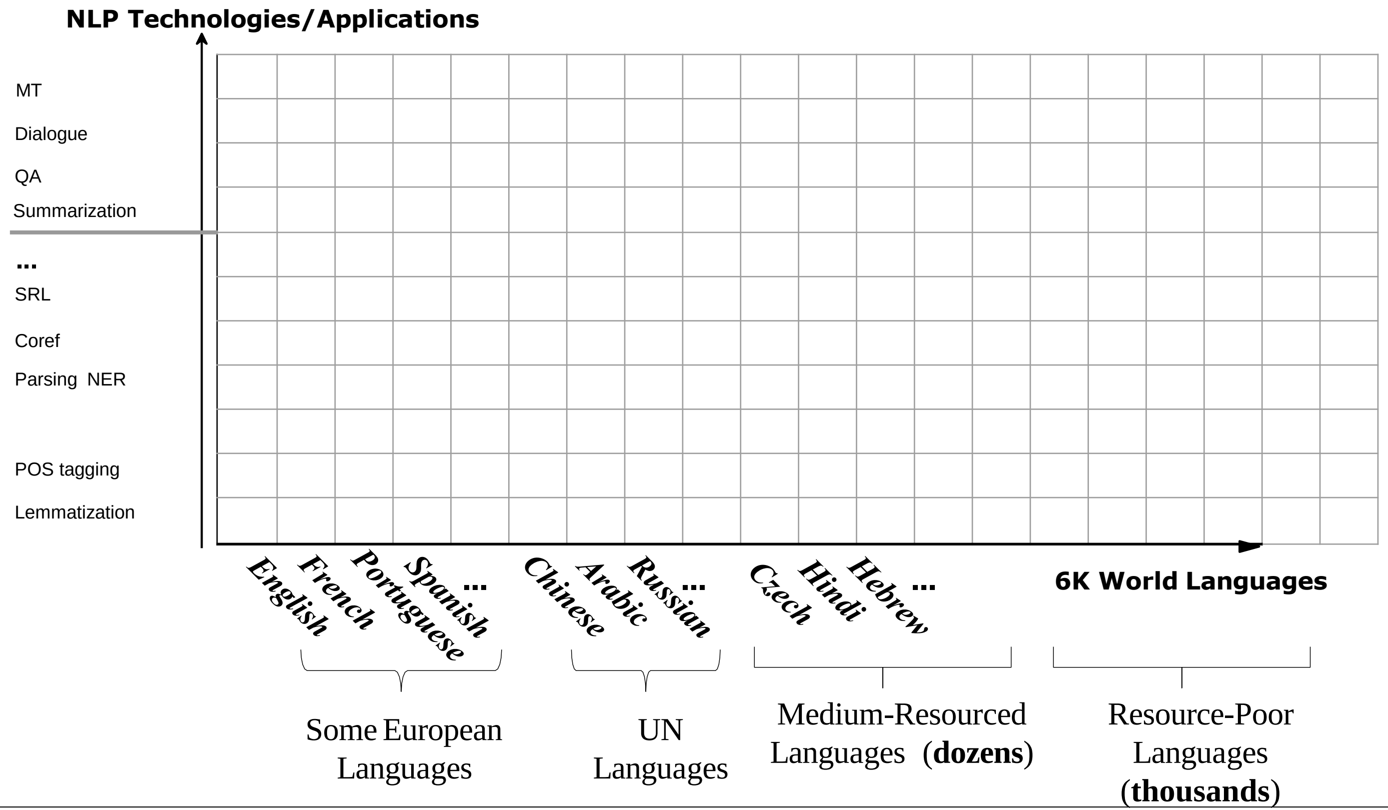
⇒ NLP algorithms model ambiguity, and choose the correct analysis in context

2. **Linguistic diversity**
   - 6–7K languages in the world
   - Languages diverge across all levels of linguistic structure

     ⇒ 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

TECHNISCHE UNIVERSITÄT DARMSTADT

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

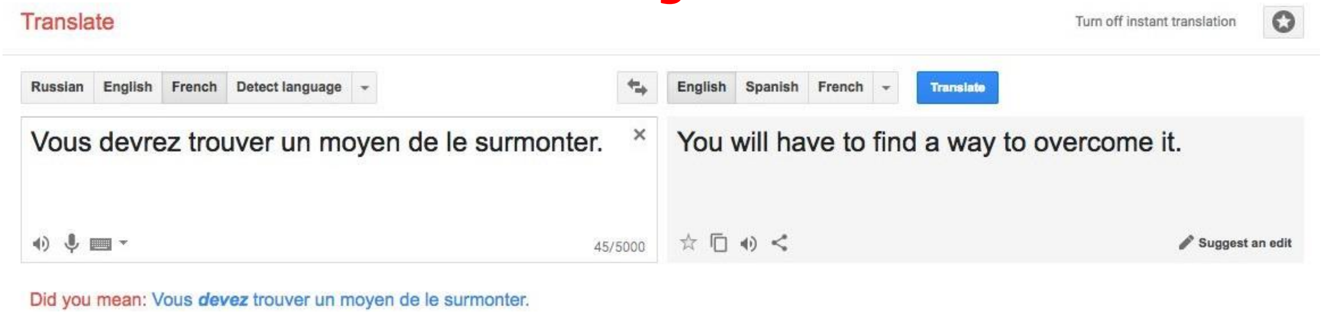
Resource-poor: few thousands of parallel sentences

Performance of Resource-Rich vs. Resource-Poor NLP Machine Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

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

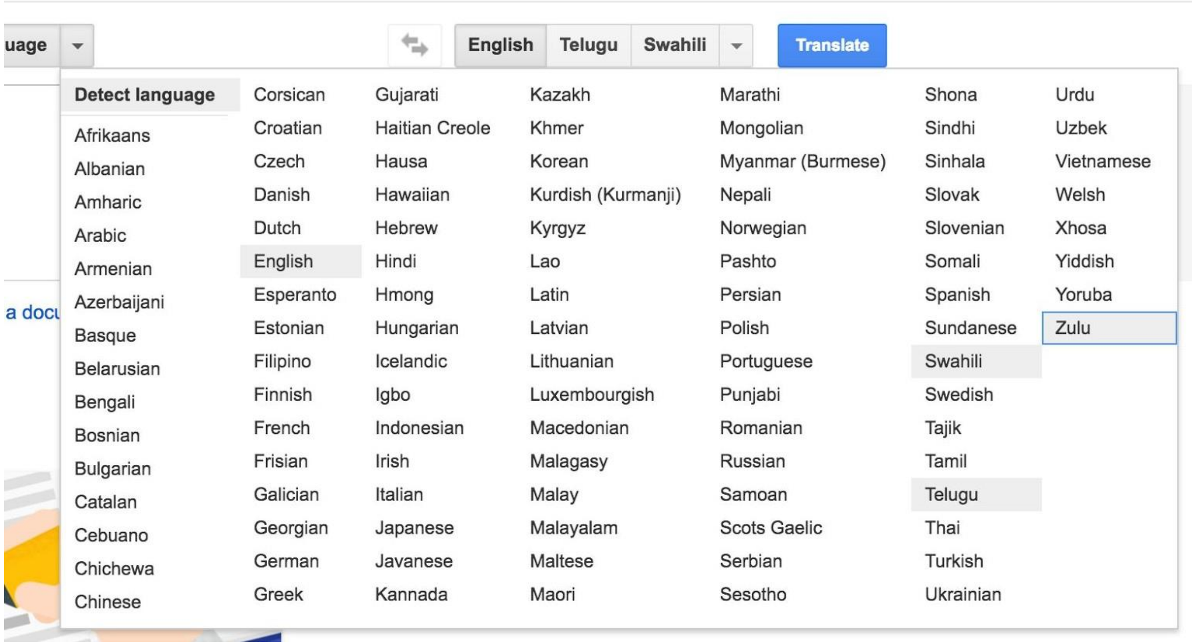
English → Swahili

Swahili → English

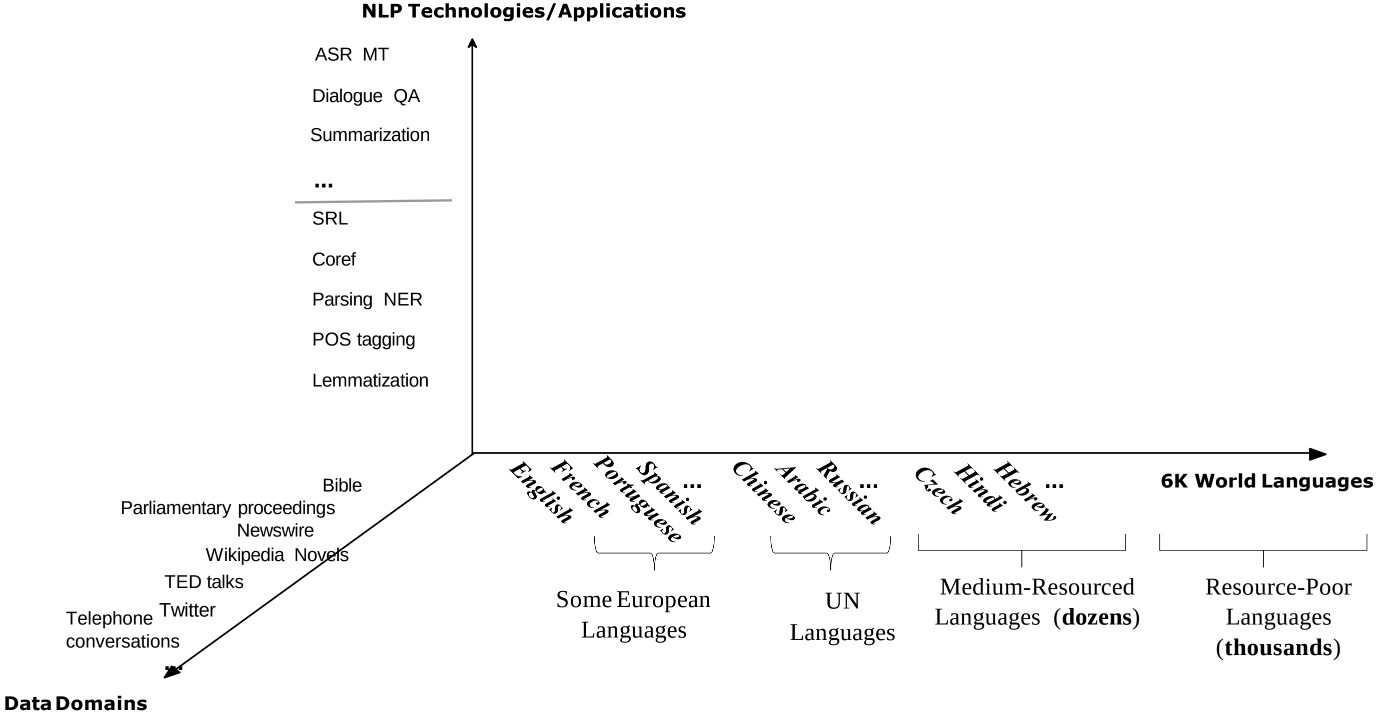
Performance of Resource-Rich vs. Resource-Poor NLP Machine Translation

TECHNISCHE UNIVERSITÄT DARMSTADT

About 130 out of 6K languages (as of 2022)

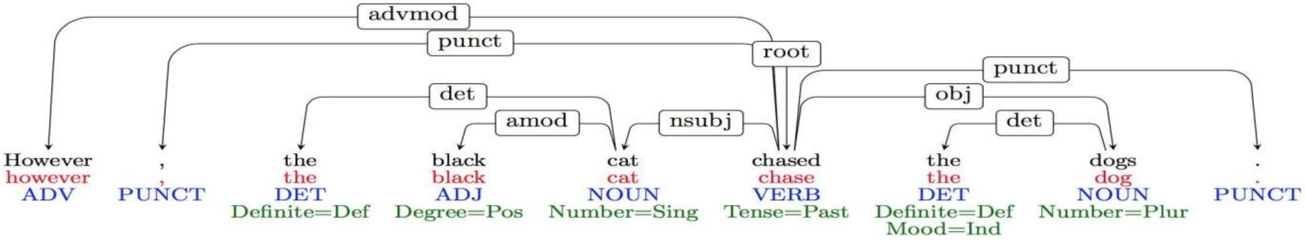


Low-Resource NLP is Not Only About Multilinguality



Low-Resource NLP is Not Only About Multilinguality

Parsing models trained using Wall Street Journals 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

TECHNISCHE UNIVERSITÄT DARMSTADT

- 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

$\overset{?}{<}$

Outline

Low-Resource NLP: Introduction

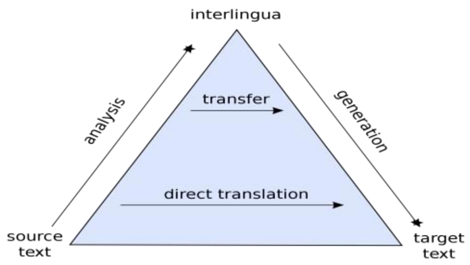
Approaches to Low-Resource NLP

Low-Resource NLP: Lorelei Program

Paradigm Shifts in NLP

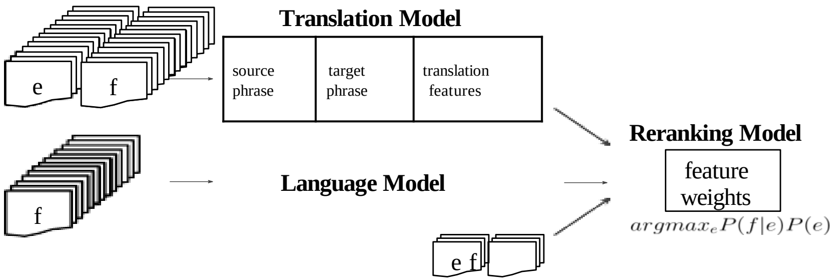
TECHNISCHE UNIVERSITÄT DARMSTADT

Logic-based/Rule-based NLP



~ 90s

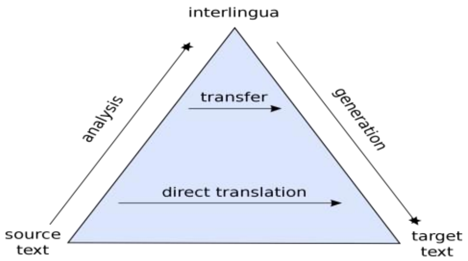
Statistical NLP



Rule-based models: high precision but very low recall

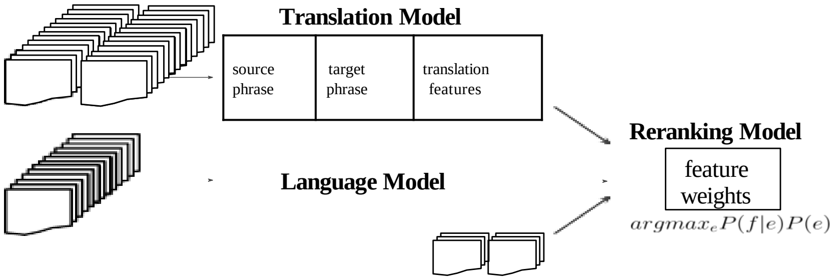
Paradigm Shifts in NLP

Logic-based/Rule-based NLP



* In resource-rich settings

Statistical NLP



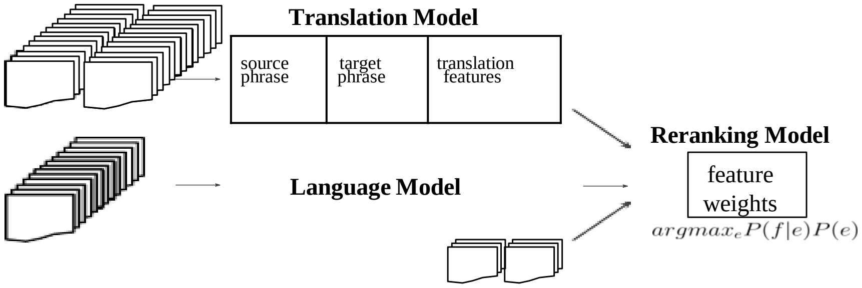
Statistical models: robust in the face of real-world data

Better performance

Less engineering of hand-crafted rules/knowledge

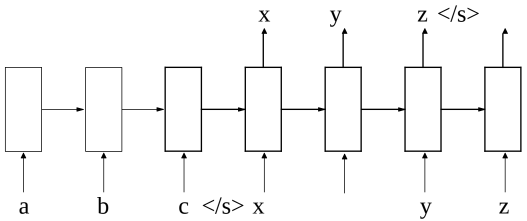
Paradigm Shifts in NLP

Statistical NLP



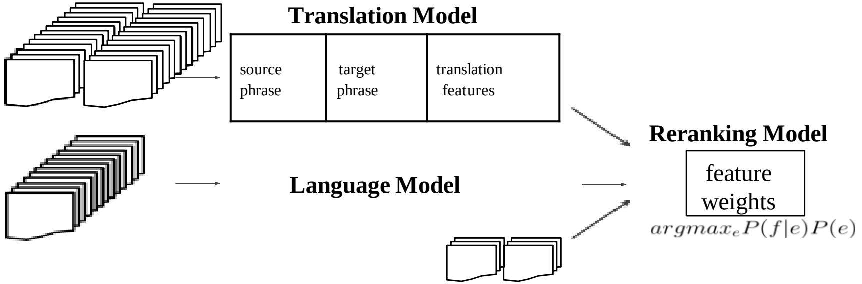
~mid 2010s

Statistical Neural NLP



Paradigm Shifts in NLP

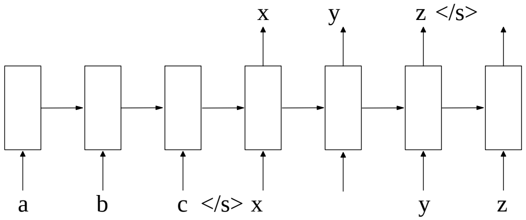
Statistical NLP



<

* In resource-rich settings

Statistical Neural NLP



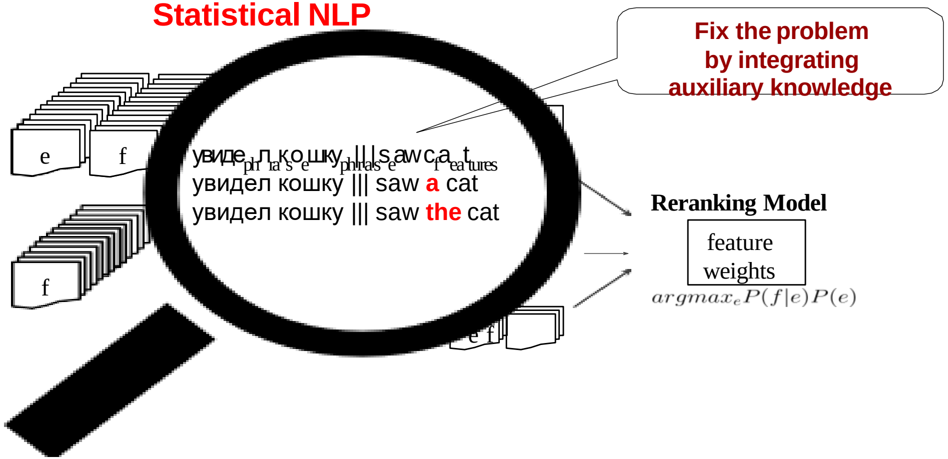
Robustness in the face of real-world data

Better performance

Less engineering of hand-crafted rules/knowledge

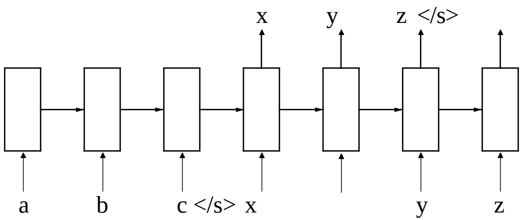
Building Blocks in Conventional Statistical NLP Models
Words, phrases ⇒ easier analysis, easier adaptation

Statistical NLP



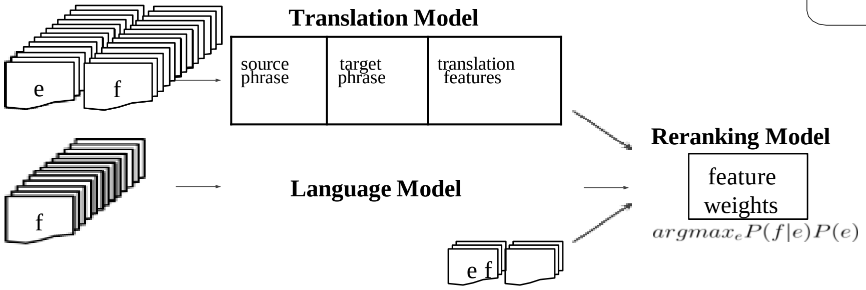
VS.

Statistical Neural NLP



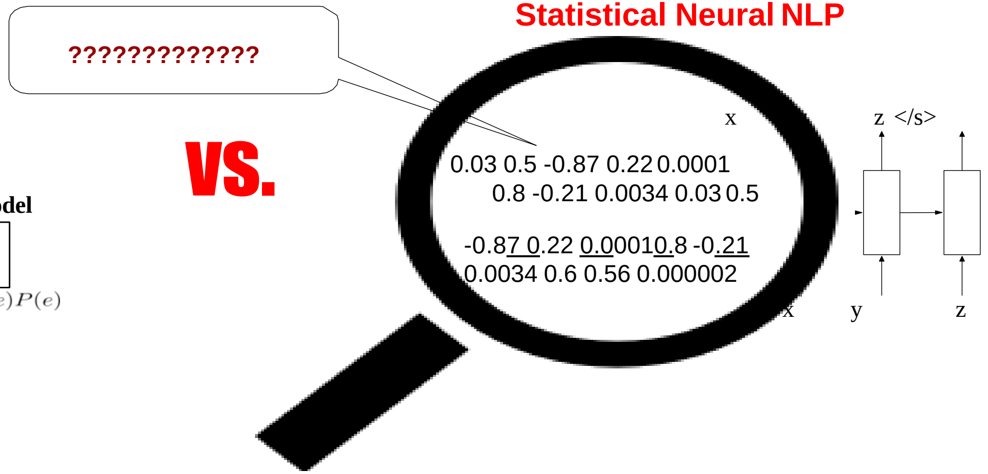
Building Blocks in Neural NLP Models
Vectors, matrices ⇒ not clear yet how to interpret

Statistical NLP



VS.

Statistical Neural NLP



How to interpret continuous representations?

How to integrate auxiliary knowledge into neural network architectures?

Problems in Low-Resource NLP

- 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
  - extreme case: 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

Unsupervised Learning

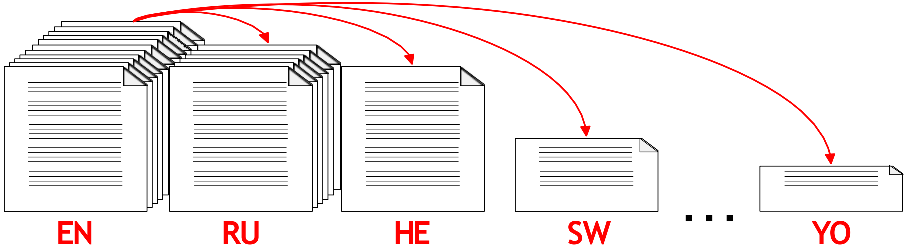
Unsupervised feature induction: Brown clustering, Word vectors

Unsupervised POS tagging

Unsupervised dependency parsing

. . .

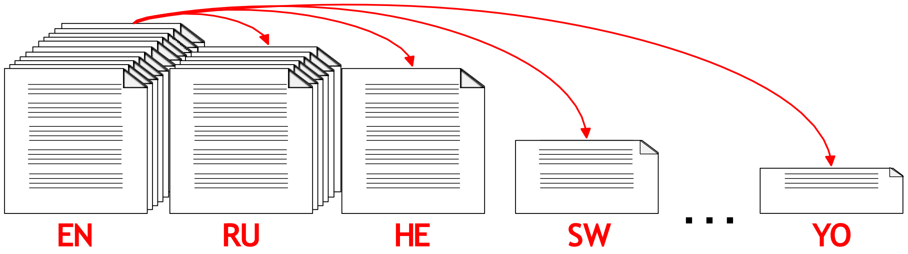
Transfer Learning



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

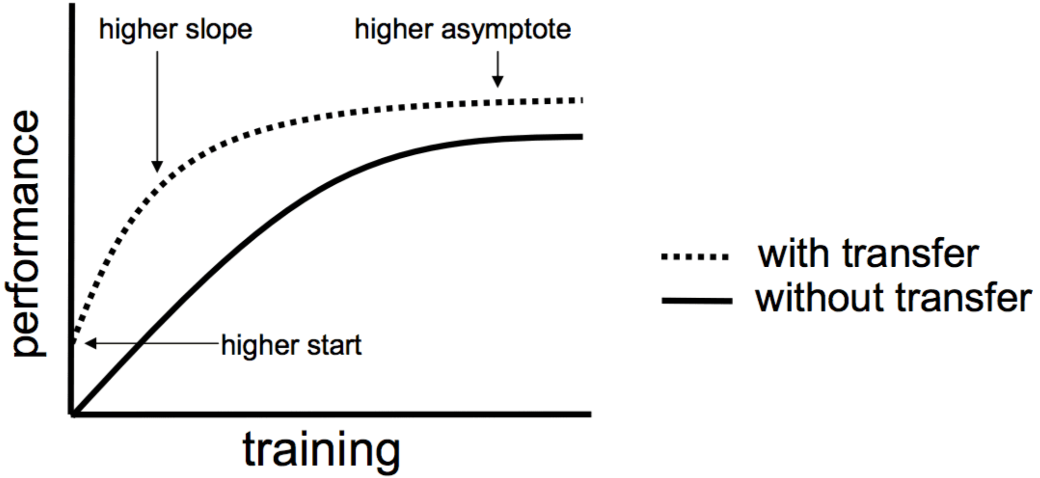
Transfer Learning



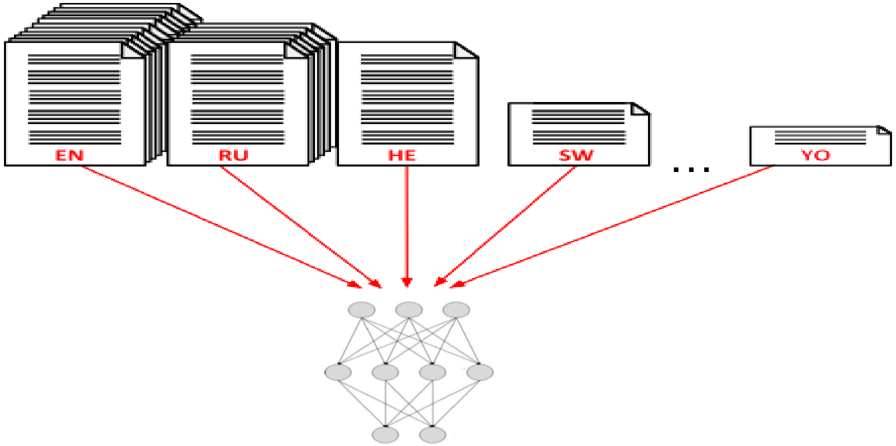
Zero-shot learning – train a model in one domains and assume it generalizes more or less out-of-the-box in a low-resource domain

One-shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it

Transfer Learning: 3 Advantages



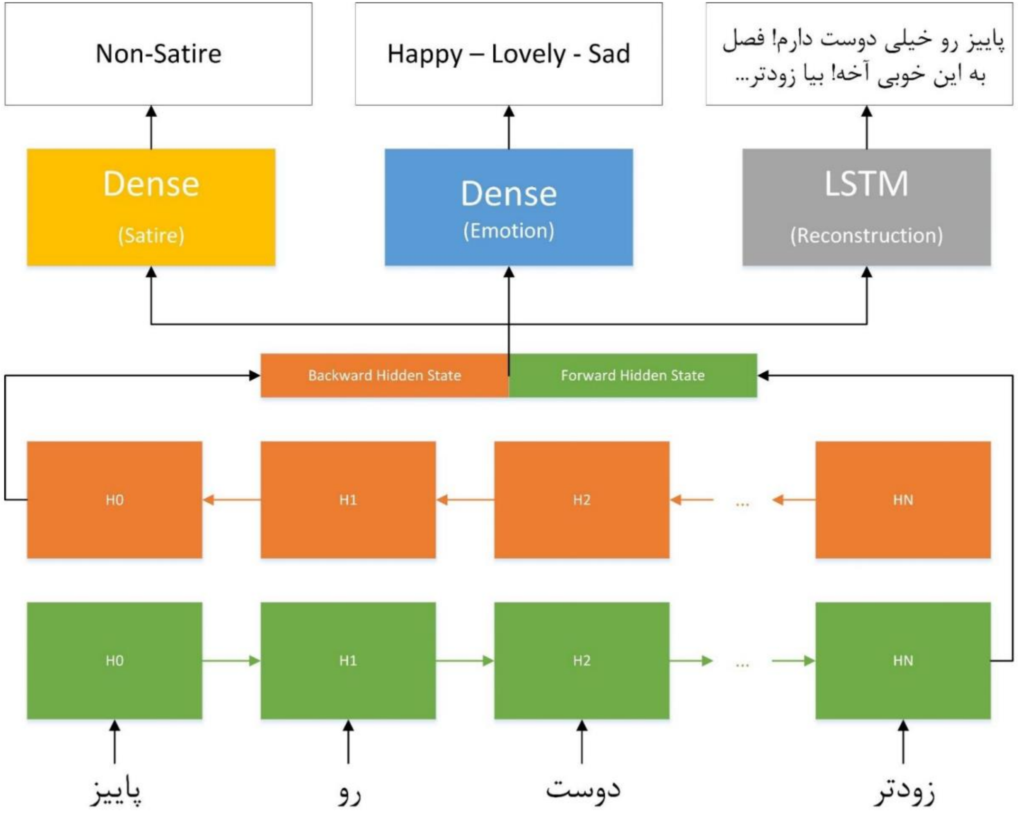
Joint Multilingual or “Polyglot” Learning



Joint resource-rich and resource-poor learning using a language-universal representation.

- Convert data in all languages to a shared representation
- Train a single model on a mix of datasets in all languages, to enable parameter sharing where possible

Case Study: Satire Detection



Satire Dataset: 2K tweets
Emotion Dataset: 300K
Reconstruction Dataset: >200M

Satire Performance (F1)

Single Task: 0.55
Multi-Task: 0.68

Another Approach: Task Modeling

Basic Idea: Model your problem in an easy-to-acquire-label setting

Example: Sentiment / Emotion analysis

Problem: People do not annotate emotions in text very often…

But: People use emoji all the time!

Solution: Approximate emotion prediction by emoji prediction!

Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm: https://arxiv.org/pdf/1708.00524.pdf

Emoji: Proxy for Emotions

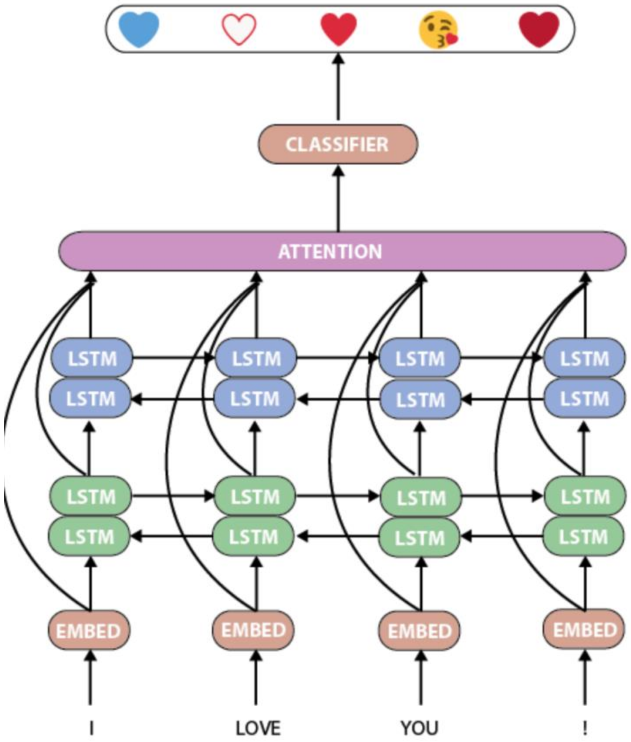
I love mom's cooking	😋 49.1%	😍 8.8%	❤️ 3.1%	☺️ 3.0%	♡ 2.9%
I love how you never reply back..	😒 14.0%	😑 8.3%	😠 6.3%	😐 5.4%	💔 5.1%
I love cruising with my homies	😎 34.0%	👌 6.6%	✌️ 5.7%	😌 4.1%	💯 3.8%
I love messing with yo mind!!	😜 17.2%	😈 11.8%	😏 8.0%	😉 6.4%	🙊 5.3%
I love you and now you're just gone..	💔 39.1%	😔 11.0%	😞 7.3%	😪 5.3%	😢 4.5%
This is shit	😠 7.0%	😡 6.4%	😞 6.0%	😒 6.0%	😤 5.8%
This is the shit	🎧 10.9%	🎶 9.7%	👌 6.5%	😎 5.7%	😏 4.8%

Source: https://www.media.mit.edu/posts/what-can-we-learn-from-emojis/

Emoji Classifier

DeepMoji Model

- Predict Emoji
- Map Emoji to Emotions



Source: http://sharif.edu/~kharrazi/data_talks/slides/data-talks-sabeti-98-7-24-slides.pdf

Emoji Prediction

پاییز رو خیلی دوست دارم! فصل
به این خوبی آخه! بیا زودتر...

Summary

- Low-Resource NLP is hard!
- Ambiguity (word senses, part-of-speech, syntactic structure…)
- Linguistic diversity at all levels of language structure
  - Tokenization, morphology, part-of-speech, syntax, semantics, discourse…

- Paradigm shifts in NLP
  - Rule-based NLP: high precision, low recall
  - Statistical NLP: Needs more data
  - Neural NLP: Needs MORE more data!

- Most promising approach: Transfer Learning

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Low-Resource NLP: Introduction

Approaches to Low-Resource NLP

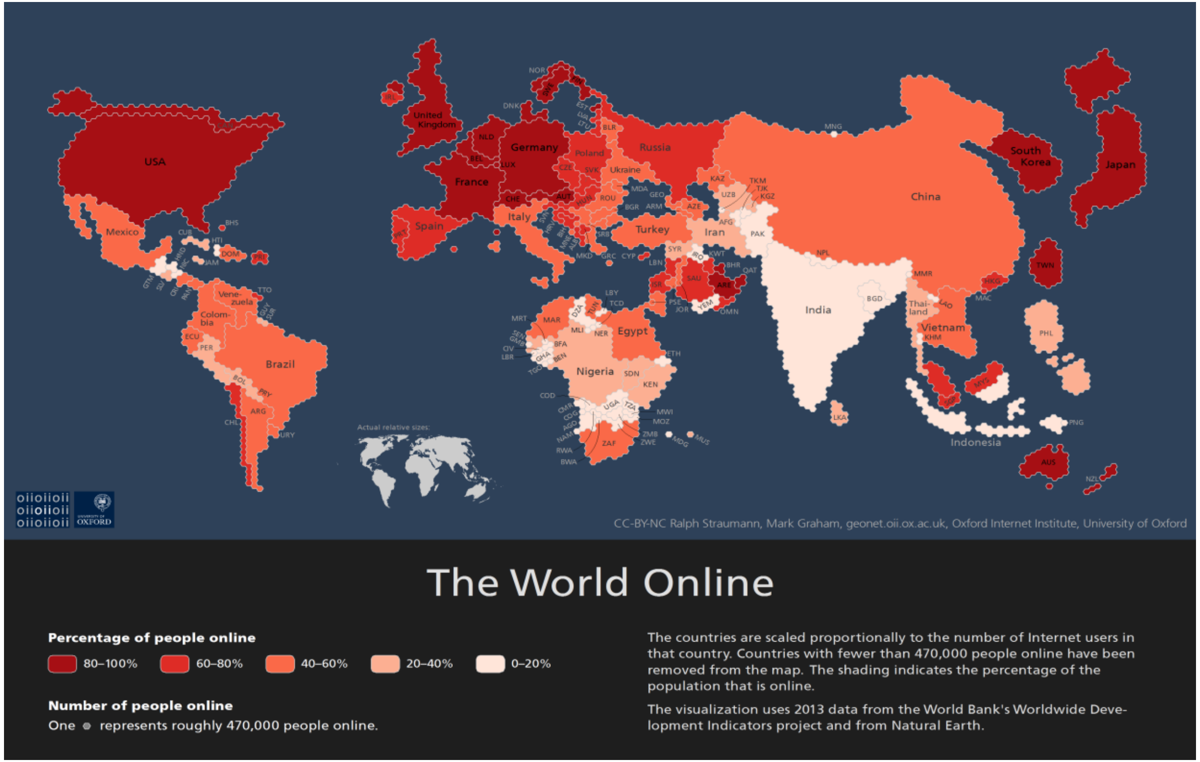
Low-Resource NLP: Lorelei Program

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?
Social good reasons

Translation systems

Speech interfaces

Dialogue systems

Educational applications

Emergency response applications

Monitoring democratic processes

Social good reasons: Emergency response

About 3 million people were affected by the quake

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

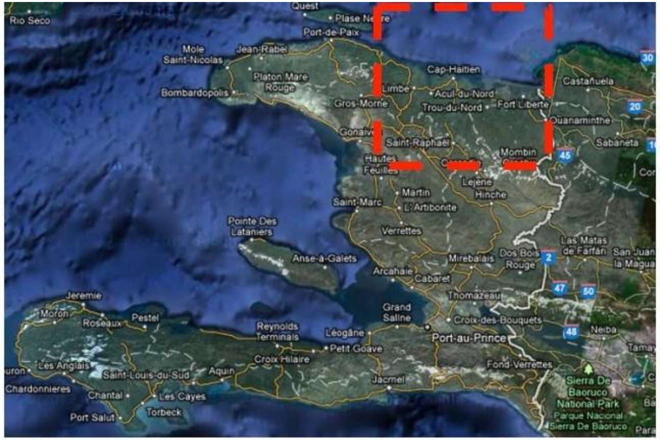
* Slide by Rob Munro http://web.stanford.edu/class/cs124

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?
Identifying outbreaks of diseases

Language Detection

Keyword Filter
“flu”, “sick”

Analytics
Which symptoms?
Are they hungover?

Why Care About Low-Resource NLP?
Identifying outbreaks of diseases

Language Detection

Government Investment in Languages

- Language Technologies mostly developed for High Resource Languages
  - English, Spanish, German, Arabic, Mandarin
- What about the other 6995 languages?
  - Maybe 30 have good resources (ASR, Treebanks, Parsers)
- What about those around 300-1000?
  - > 1 Millions speakers, writing systems…
- If no immediate commercial value no support happens

US Government LT Investment

- DARPA (Defense Advance Research Projects Agency)
  - Invested in MT from 1940s
  - Invested in ASR from 1970s
  - Invested in Dialog systems from 1990s
  - Invested in Speech Translation from 1990s
- Case study Lorelei (2016-2021)

The Scenario

- Disaster happens! (e.g. earthquake)
- Area effected doesn’t use major language
- Communication is in local language
  - News, TV/Radio, Social Media
- What is going on?
  - Where should you provide support
  - Who is affected
  - How many people need help
  - What is the urgency

Lorelei Incident

- Disaster happens! (e.g. earthquake)
- Communication is in local language
  - News, TV/Radio, Social Media
- Provide
  - Machine Translation
  - Named Entity Recognition
  - Situation Frames (11 types) plus location, status, urgency, “gravity”
    - evac, food, infra, med, search…

Lorelei Incident

- Disaster happens! (e.g. earthquake)
- Communication is in local language
    - News, TV/Radio, Social Media
- Provide
    - Machine Translation
    - Named Entity Recognition
    - Situation Frames (11 types) plus location, status, urgency, “gravity”
        - evac, food, infra, med, search…
- Do this in
    - 24 hours
    - 7 days
- You are told the language at hour 0

Lorelei Evaluation Exercises

- May 2016: Dry Run (Mandarin)
- July 2016: Uighur (Turkic Language spoken in Western China)
- July 2017: Tigrinya and Oromo (spoken in Eritrea and Ethiopia)
- July 2018: Kinyarwanda and Sinhala (spoken in Uganda and Sri Lanka)
- July 2019: ??? and ???

Techniques

- Perform in pronunciation space
  - Not words, morphemes or character space

- Cross Lingual Transfer
  - If word3(L1) co-occurs with word1(L1), word2(L1)
  - And word3(L2) co-occurs with word1(L2), word2(L2)
  - And trans(word1(L1)) = word1(L2) and trans(word2(L1)) = word2(L2)
  - Maybe trans(word3(L1)) = word3(L2)?

Techniques

- Use available resources
  - Religious Texts (Bible, Quran, Unix Manuals…)
  - Wikipedia
  - Native Informant

- Global Linguistic Knowledge
  - High morphology language more likely to have free word order
  - Close language borrowing

Lorelei Advances

- Techniques for low resource languages
  - Translation, interpretation, sentiment
  - Both particular languages, and general techniques

- Machine Learning
  - Better use of limited data
  - Use transfer learning
  - Not naive just end-to-end
  - Using large mono-lingual dataset to improve models
  - Using structure to make learning easier

In Two Weeks:

Privacy & Security I