Challenges of language technologies for the indigenous languages of the Americas

Manuel Mager[1], Ximena Gutierrez-Vasques[2], Gerardo Sierra[2] and Ivan Meza[1]

[1]IIMAS-UNAM
[2]GIL-UNAM

August 21, 2018

Objectives

This is a survey paper on NLP research for the indigenous languages of the Americas

Objectives

This is a survey paper on NLP research for the indigenous languages of the Americas

Our aims:

Objectives

This is a survey paper on NLP research for the indigenous languages of the Americas

Our aims:

- To explore the current research in NLP for the indigenous languages spoken in the American continent and to encourage research for these languages.

Objectives

This is a survey paper on NLP research for the indigenous languages of the Americas

Our aims:

- To explore the current research in NLP for the indigenous languages spoken in the American continent and to encourage research for these languages.
- To point out some of the challenges that arise when dealing with these languages.

Question 1/4:

Why should we study indigenous languages?

Indigenous languages of the Americas

- Wide range of linguistic families in the Americas

Indigenous languages of the Americas

- Wide range of linguistic families in the Americas
- They exhibit linguistic phenomena that are different from the most common languages usually studied in Natural Language Processing (NLP)

Indigenous languages of the Americas

Tones

- Otomi language

High tone /dá-tsot'e/ *(1.CPL-arrive)* 'I arrived'
Low tone /da-tsot'e/ *(3.IRR-arrive)* 'He would arrive'

- Mixtec language

nu$^{\mathbf{3}}$mi$^{\mathbf{3}}$ *(3.IRR-hug)* 'He would hug'
nu$^{\mathbf{14}}$mi$^{\mathbf{3}}$ *(3.NEG.IRR-hug)* 'He would not hug'
nu$^{\mathbf{13}}$mi$^{\mathbf{3}}$ *(3.CPL-hug)* 'He hugged'

Morphology

Indigenous languages have a high morpheme per word rate and many of them are polysynthetic.

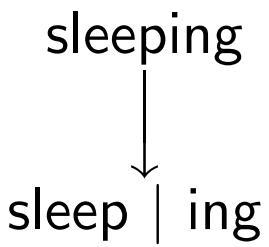
English example

sleeping

Morphology

Indigenous languages have a high morpheme per word rate and many of them are polysynthetic.

English example



Morphology

Indigenous languages have a high morpheme per word rate and many of them are polysynthetic.

English example

sleeping

↓

sleep | ing

German example

Telekommunikationsdienstleistungsunternehmen

Morphology

Indigenous languages have a high morpheme per word rate and many of them are polysynthetic.

English example

sleeping

↓

sleep | ing

German example

Telekommunikationsdienstleistungsunternehmen

↓

Tele | kommunikation | s | dienst | leistung | s | unter | nehmen

Morphology

Example in Wixarika

Tsimekam+kakatenixetsihanuyutits++kiriyeku
kuyatsit+iriex+aximekaitsiek+t+kaku

↓

Tsi | me | ka | m+ | ka | ka | te | ni | xe | tsi | hanu | yu | ti |
ts++ki | ri | ye | ku | ku| ya | tsi | t+i | rie | x+a | xime | kai | tsie
| k+ | t+ | kaku

Morphology

Example in Wixarika

Tsimekam+kakatenixetsihanuyutits++kiriyeku
kuyatsit+iriex+aximekaitsiek+t+kaku

↓

Tsi | me | ka | m+ | ka | ka | te | ni | xe | tsi | hanu | yu | ti |
ts++ki | ri | ye | ku | ku| ya | tsi | t+i | rie | x+a | xime | kai | tsie
| k+ | t+ | kaku

Morphology

Example in Wixarika

Tsimekam+kakatenixetsihanuyutits++kiriyeku
kuyatsit+iriex+aximekaitsiek+t+kaku

Tsi | me | ka | m+ | ka | ka | te | ni | xe | tsi | hanu | yu | ti |
ts++ki | ri | ye | ku | ku| ya | tsi | t+i | rie | x+a | xime | kai | tsie
| k+ | t+ | kaku

(even when they were about to make their dogs fight on the hill)

Indigenous languages of the Americas

Other characteristics:

- Lack of orthographic normalization

Indigenous languages of the Americas

Other characteristics:

- Lack of orthographic normalization
- Wide dialectal variation

Indigenous languages of the Americas

Other characteristics:

- Lack of orthographic normalization
- Wide dialectal variation
- Limited digital text production

Indigenous languages of the Americas

Other characteristics:

- Lack of orthographic normalization
- Wide dialectal variation
- Limited digital text production

They are challenging!

The numbers

- Around **140** linguistic families in the world

The numbers

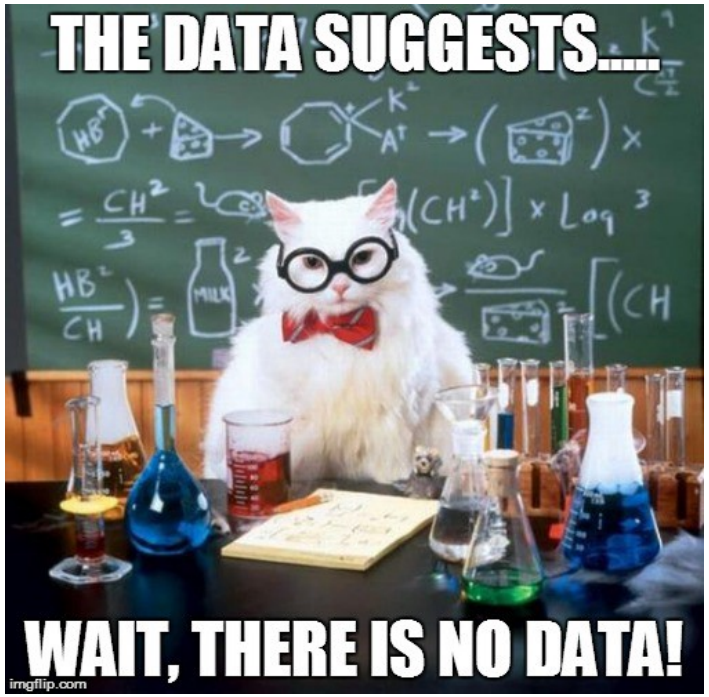
- Around **140** linguistic families in the world
- Almost **40%** are native to the Americas

The numbers

- Around **140** linguistic families in the world
- Almost **40%** are native to the Americas
- **900** different indigenous languages are spoken in this region (approximately)

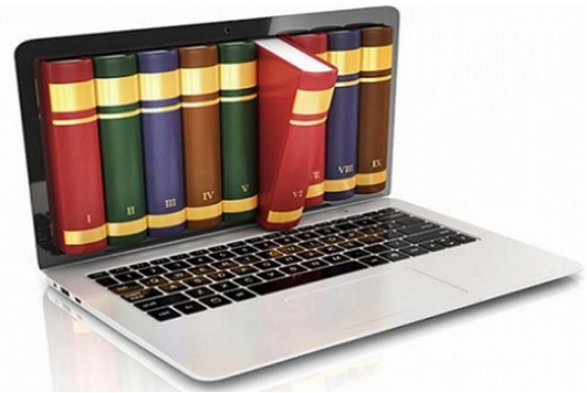
Question 2/4:

With what resources do we count and how much is there?



Corpus and digital resources

- Most of the current state of the art methods require **vast amounts** of corpora in order to achieve **good performance**
- **Indigenous languages** of the Americas do not have an important **web presence** or **text production** comparable to richer resourced languages. It is difficult to find websites that offer their content the native languages.



What kind of data is available for those languages?

- Parallel corpora
- Dictionaries
- Speech
- Morphology
- Not so common: Treebanks and POS tagging

We need more standardized datasets

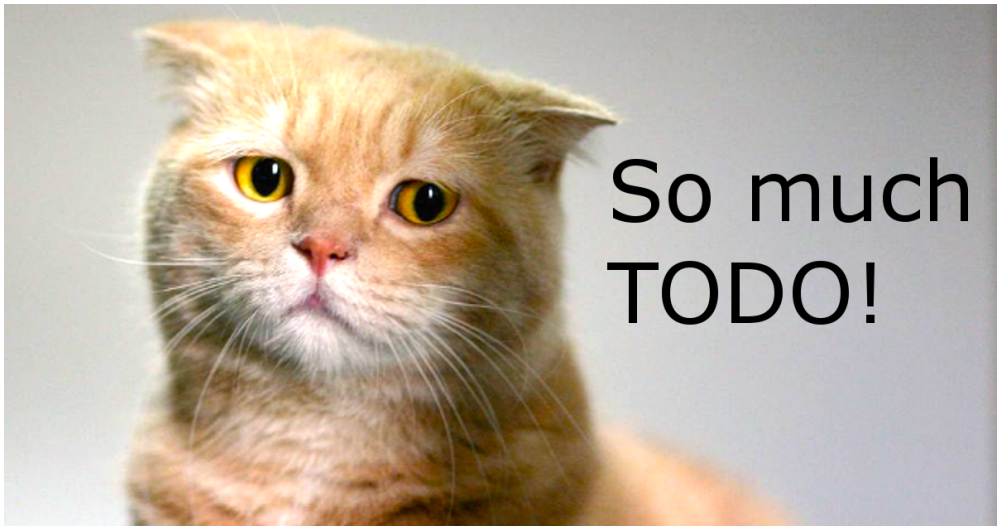
Maximum amount of data.

- Parallel corpora $\sim$ **18K phrases**
- Dictionaries $\sim$ **3.5K words**
- Speech $\sim$ **10 hours** (annotated)
- Morphology $\sim$ **2K roots**
- Not so common: Treebanks and POS Tagging **2K sentences**

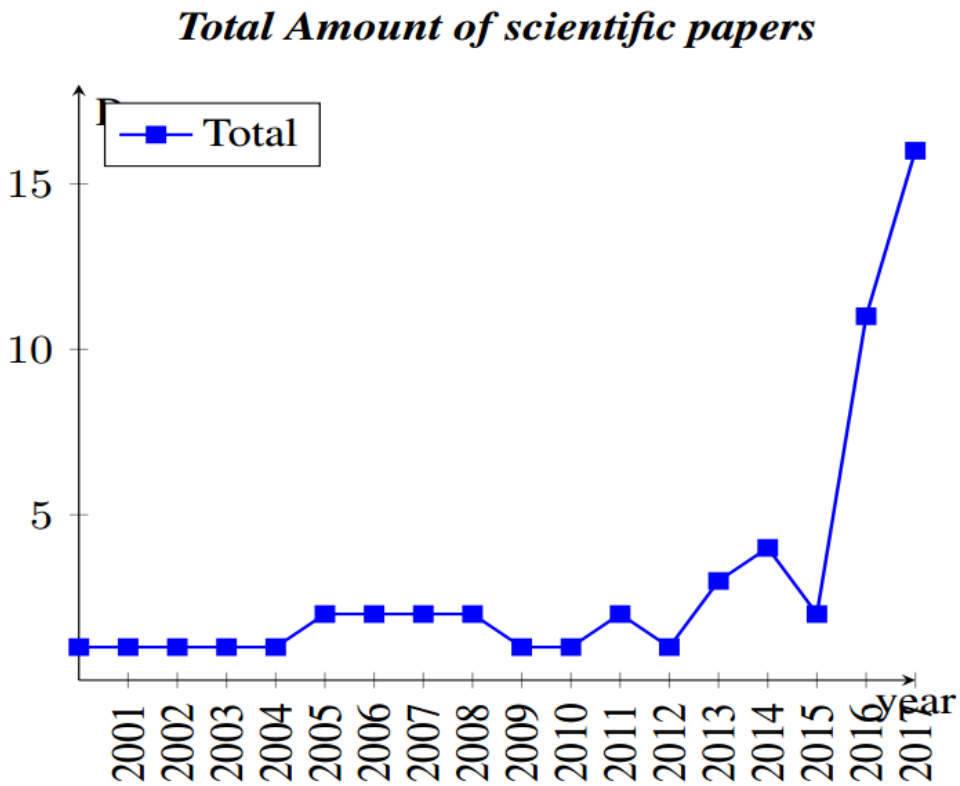
We need more standardized datasets

Question 2/4:

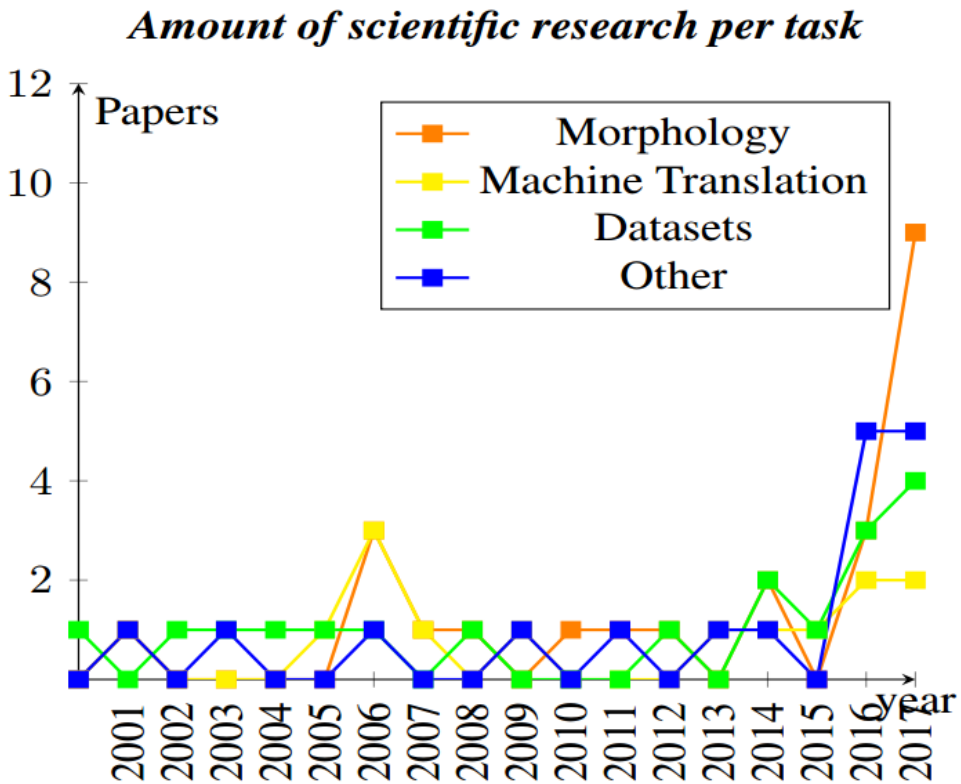
How much have we done and how much do we lack?



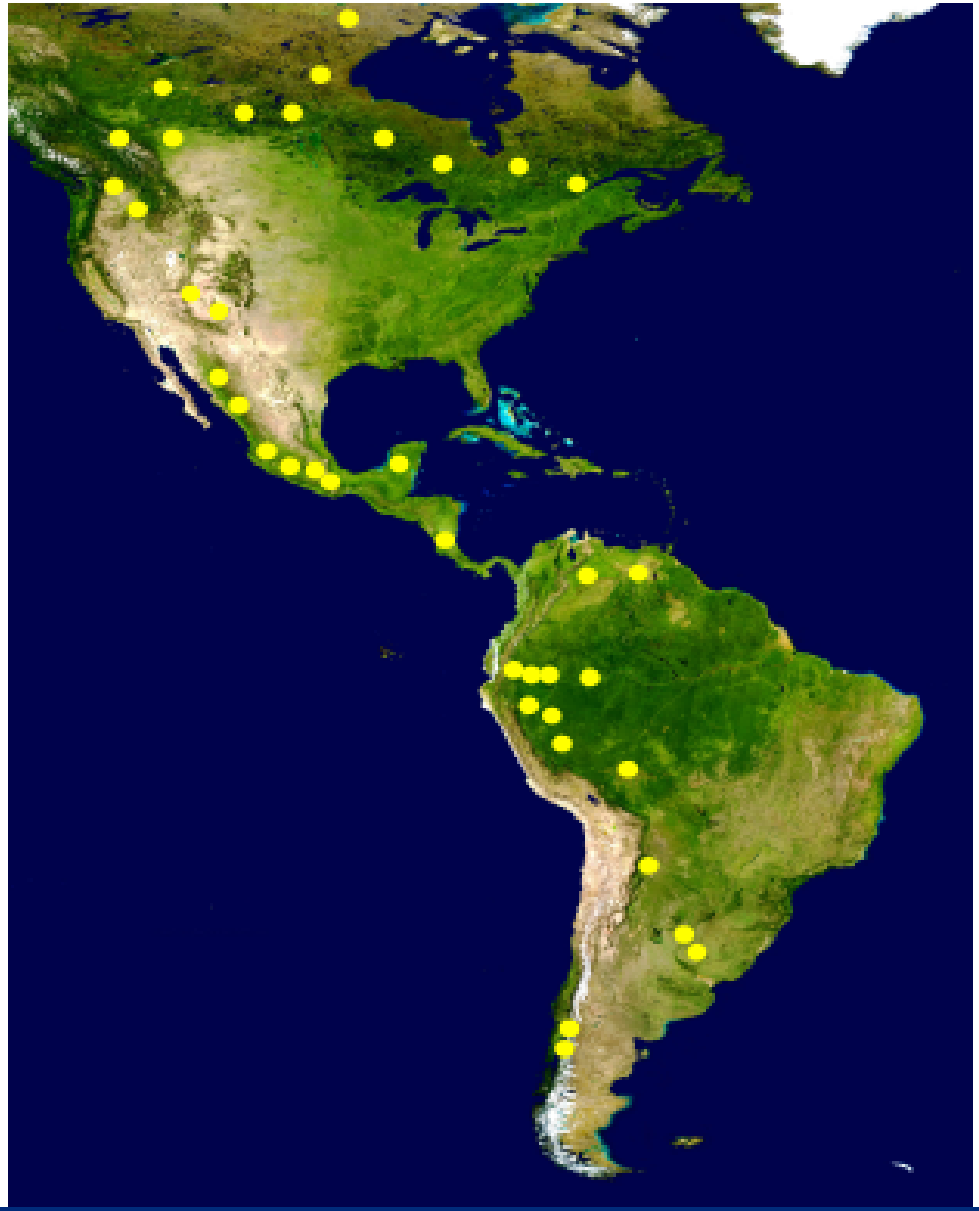
NLP research on indigenous languages in the Americas is getting more popular.



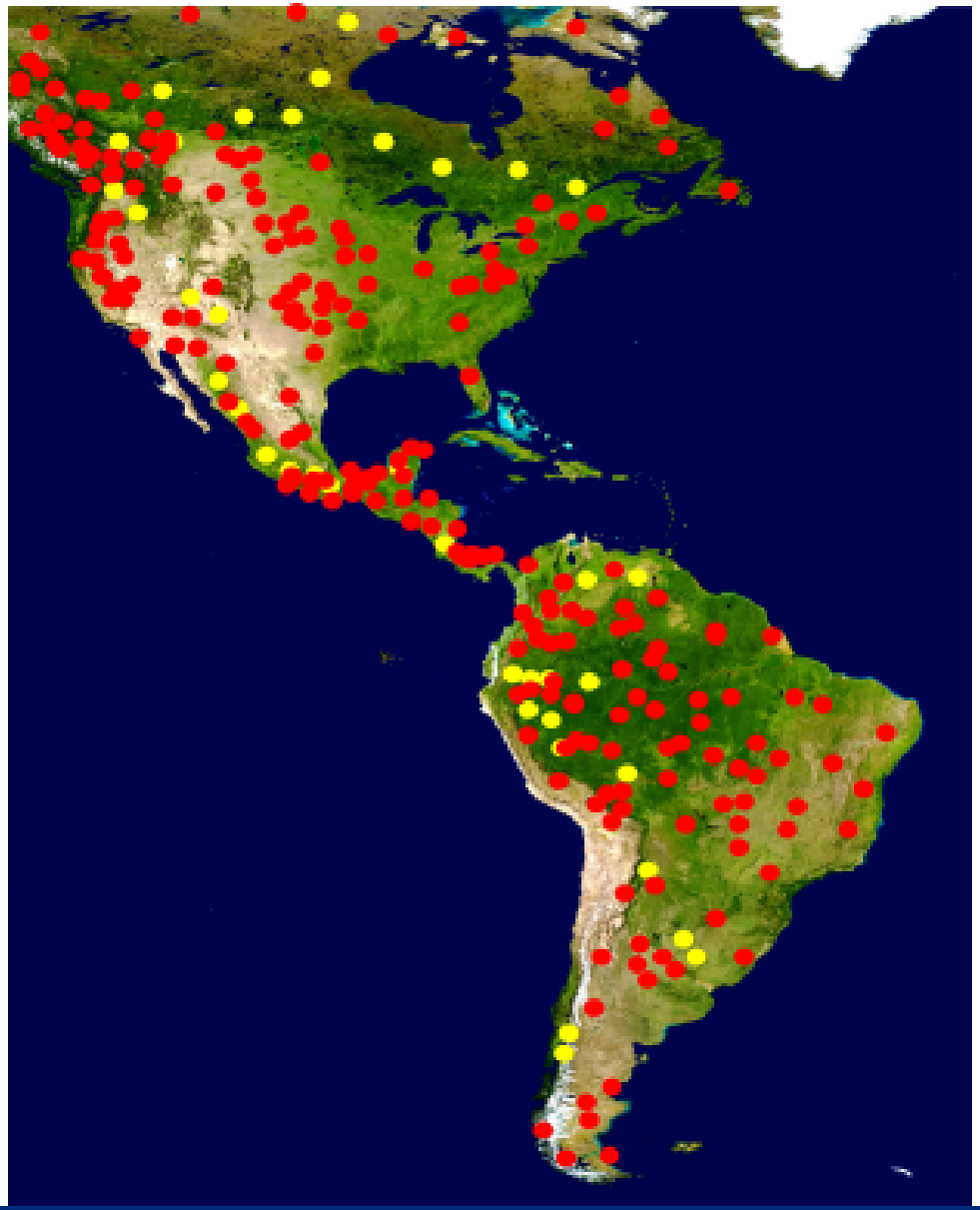
Most popular tasks



Studied languages

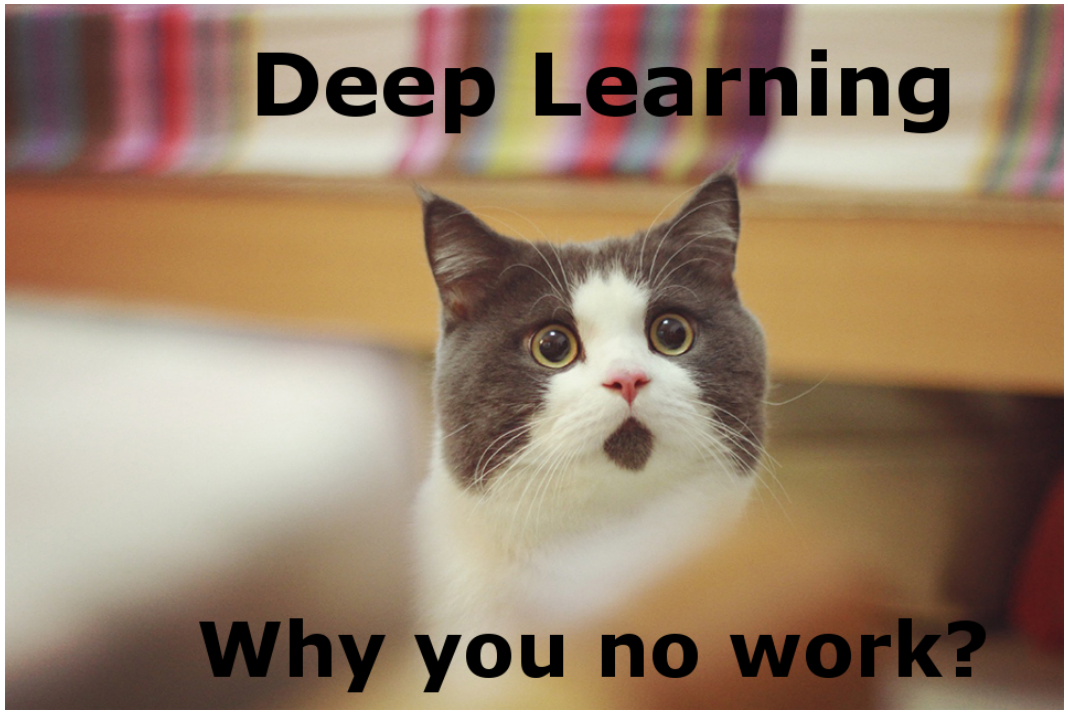


But..



Question 3/4:

What progress and challenges are there for each NLP task?



Morphology

- **Lemmatization and stemming:** typical methods for reducing morphological variation in NLP. Plenty of tools available for English and for a reduced set of languages

Morphology

- **Lemmatization and stemming:** typical methods for reducing morphological variation in NLP. Plenty of tools available for English and for a reduced set of languages
- However, **not all languages are suffixal** (it is not enough to remove inflectional endings in order to obtain a stem)

Morphology

- **Lemmatization and stemming:** typical methods for reducing morphological variation in NLP. Plenty of tools available for English and for a reduced set of languages
- However, **not all languages are suffixal** (it is not enough to remove inflectional endings in order to obtain a stem)
- We need to develop **morphological tools** for languages of the **Americas**, i.e., morphological analysis, morphological segmentation, inflection/reinflection, etc.

Morphology

What has been done so far?

- Rule-based tools for morphological analysis (most of the work)

Morphology

What has been done so far?

- Rule-based tools for morphological analysis (most of the work)
- Unsupervised/semisupervised methods for morphological segmentation

Morphology

What has been done so far?

- Rule-based tools for morphological analysis (most of the work)
- Unsupervised/semisupervised methods for morphological segmentation
- **Some neural systems for low resources!**

Morphology

What has been done so far?

- Rule-based tools for morphological analysis (most of the work)
- Unsupervised/semisupervised methods for morphological segmentation
- **Some neural systems for low resources!**
- Shared tasks that include American languages (the CoNLL-SIGMORPHON Shared Task)

Machine translation

Constructing viable and production quality in MT is a great challenge.

- Rule based MT (popular for low-resource languages, requires linguistic knowledge)

Machine translation

Constructing viable and production quality in MT is a great challenge.

- Rule based MT (popular for low-resource languages, requires linguistic knowledge)
- Building SMT systems using small parallel corpora

Machine translation

Constructing viable and production quality in MT is a great challenge.

- Rule based MT (popular for low-resource languages, requires linguistic knowledge)
- Building SMT systems using small parallel corpora
- Currently, NMT systems have no good results.

Machine translation

Constructing viable and production quality in MT is a great challenge.

- Rule based MT (popular for low-resource languages, requires linguistic knowledge)
- Building SMT systems using small parallel corpora
- Currently, NMT systems have no good results.

Sub-word level models and ML research in low-resource settings could enhance MT for these languages

Multilinguality and code-switching

- Speech synthesis and recognition

Multilinguality and code-switching

- Speech synthesis and recognition
- Part-of-Speech Tagging

Multilinguality and code-switching

- Speech synthesis and recognition
- Part-of-Speech Tagging
- Spell Checking

Multilinguality and code-switching

- Speech synthesis and recognition
- Part-of-Speech Tagging
- Spell Checking
- Optic Character Recognizing

Multilinguality and code-switching

- Speech synthesis and recognition
- Part-of-Speech Tagging
- Spell Checking
- Optic Character Recognizing
- Language Identification

Multilinguality and code-switching

- Speech synthesis and recognition
- Part-of-Speech Tagging
- Spell Checking
- Optic Character Recognizing
- Language Identification
- Parsing

Multilinguality and code-switching

- Speech synthesis and recognition
- Part-of-Speech Tagging
- Spell Checking
- Optic Character Recognizing
- Language Identification
- Parsing
- Code-switching language identification

Final remarks

- We noticed that **North American** languages are the most studied

Final remarks

- We noticed that **North American** languages are the most studied
- NLP research for Americas indigenous languages can **broad the understanding** of human languages (more general computational models)

Final remarks

- We noticed that **North American** languages are the most studied
- NLP research for Americas indigenous languages can **broad the understanding** of human languages (more general computational models)
- Positive **social impact** for the speakers (maintaining the living cultural heritage that each language represents)

Final remarks

- We noticed that **North American** languages are the most studied
- NLP research for Americas indigenous languages can **broad the understanding** of human languages (more general computational models)
- Positive **social impact** for the speakers (maintaining the living cultural heritage that each language represents)
- Shared **tasks** and special **workshops** help to encourage more research

Question 4/4:

What is next? (and conclusions)

Conclusions

- We presented some features and characteristics of the indigenous languages of the Americas

Conclusions

- We presented some features and characteristics of the indigenous languages of the Americas
- The study of such languages can lead us for a more complete understanding of human languages and advance towards universal NLP models.

Conclusions

- We presented some features and characteristics of the indigenous languages of the Americas
- The study of such languages can lead us for a more complete understanding of human languages and advance towards universal NLP models.
- The interest for those languages are growing in NLP community

Conclusions

- We presented some features and characteristics of the indigenous languages of the Americas
- The study of such languages can lead us for a more complete understanding of human languages and advance towards universal NLP models.
- The interest for those languages are growing in NLP community
- Most research has been done for MT and Morphology

Conclusions

- We presented some features and characteristics of the indigenous languages of the Americas
- The study of such languages can lead us for a more complete understanding of human languages and advance towards universal NLP models.
- The interest for those languages are growing in NLP community
- Most research has been done for MT and Morphology
- However, more diversity in tasks is growing

Conclusions

- We presented some features and characteristics of the indigenous languages of the Americas
- The study of such languages can lead us for a more complete understanding of human languages and advance towards universal NLP models.
- The interest for those languages are growing in NLP community
- Most research has been done for MT and Morphology
- However, more diversity in tasks is growing
- A complete list of papers and resources can be found in the paper

Check out out web page with an updated list of resources and papers!

`https://github.com/pywirrarika/naki`

Future work

Upcoming work:

- Comparing morphological complexity of Spanish, Otomi and Nahuatl. (Gutierrez-Vasquez and Minjagos). Complexity Workshop, COLING!
- Lost in Translation: Analysis of Information Loss During Machine Translation Between Polysynthetic and Fusional Languages (Mager et al.,) Polysynthetic workshop, COLING!
- Codeswitching (*work in progress*)
- Please contact us for more information!

Future work

Upcoming work:

- Comparing morphological complexity of Spanish, Otomi and Nahuatl. (Gutierrez-Vasquez and Minjagos). Complexity Workshop, COLING!
- Lost in Translation: Analysis of Information Loss During Machine Translation Between Polysynthetic and Fusional Languages (Mager et al.,) Polysynthetic workshop, COLING!
- Codeswitching (*work in progress*)
- Please contact us for more information!

Thank you!