www.idiap.ch

- **Areas of activity**: multilingual and multimodal interaction and multimedia information management, including human behavior modeling.
- **Staff**: 120+ (+50 across 16 start-ups)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

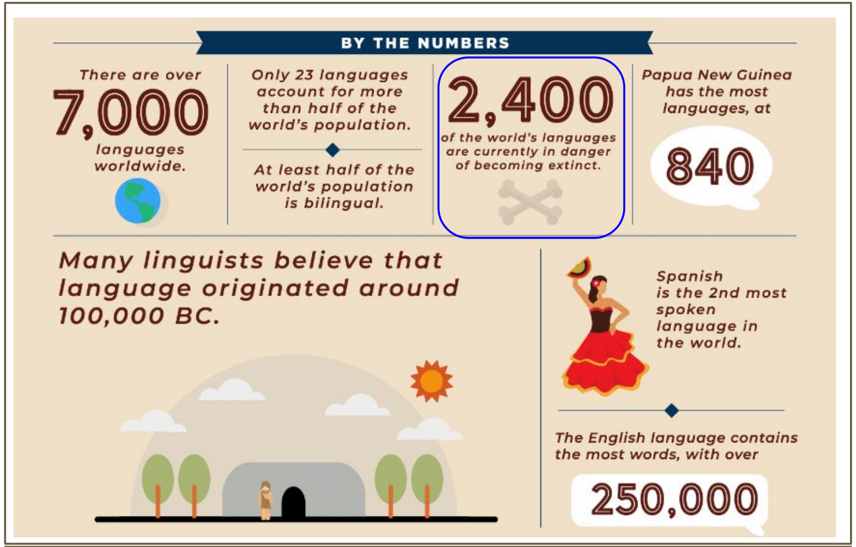
Multimodal Multilingual Corpus Development for Machine Translation

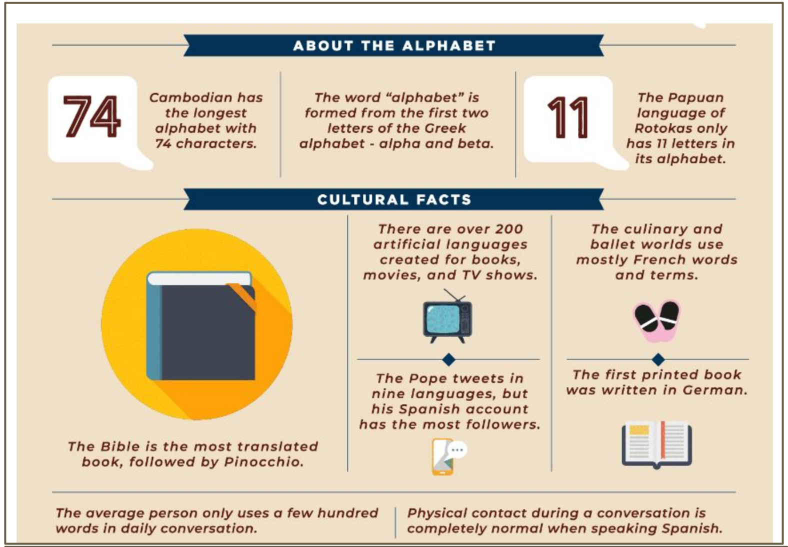
Dr. Shantipriya Parida
Idiap Research Institute
Martigny, Switzerland

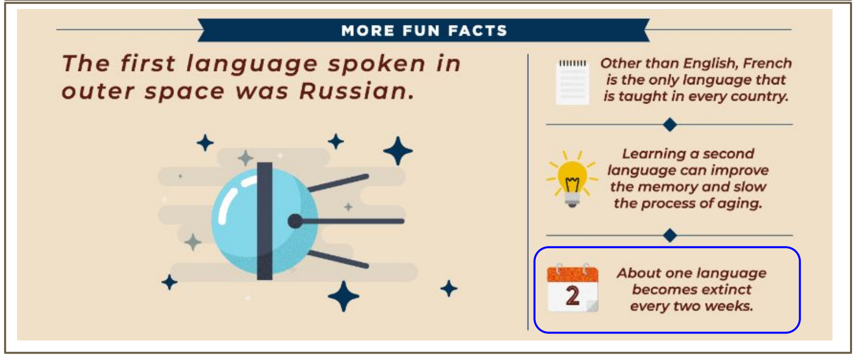
Agenda

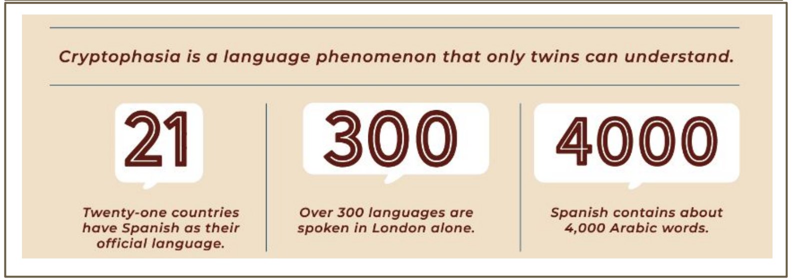
- **Overview**
- Corpus Development
- Case Study1 : Hindi Visual Genome
- Case Study2 : Malayalam Visual Genome
- Conclusion

Some Facts



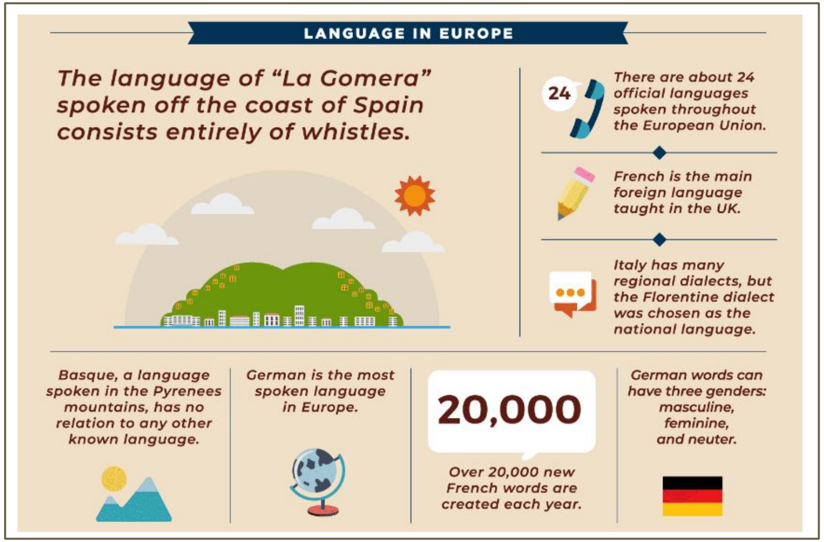


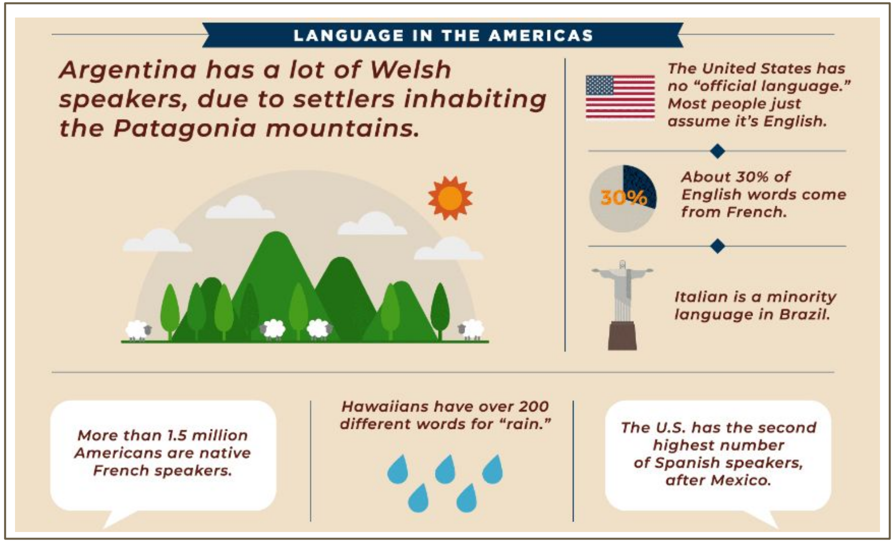




Source: https://takelessons.com/blog/language-facts-z14

Some Facts





Source: https://takelessons.com/blog/language-facts-z14

Some Facts

- How many facts (from above slides) already you knew ?.
- Do you have any interesting facts about languages (e.g. Indian languages) to share ?.

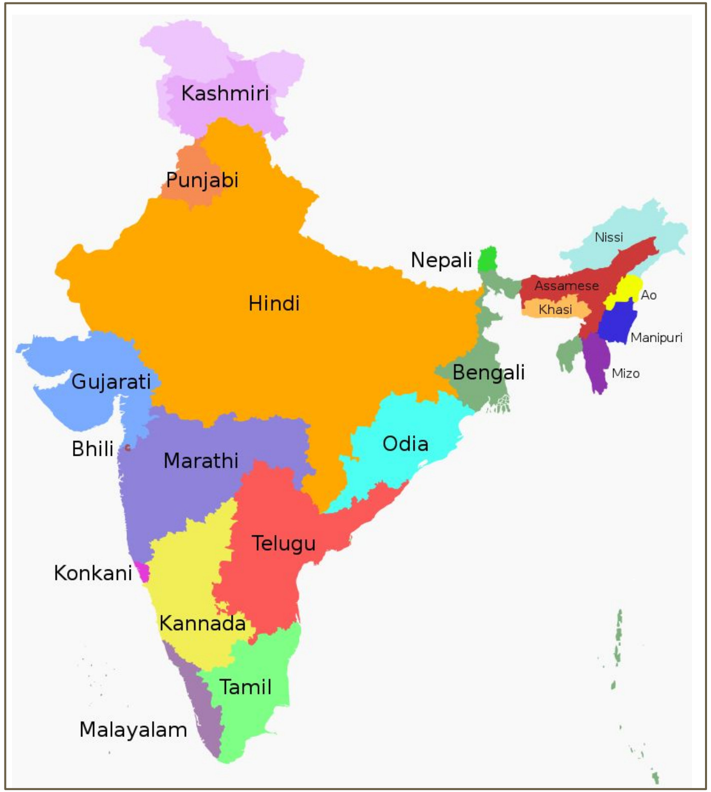
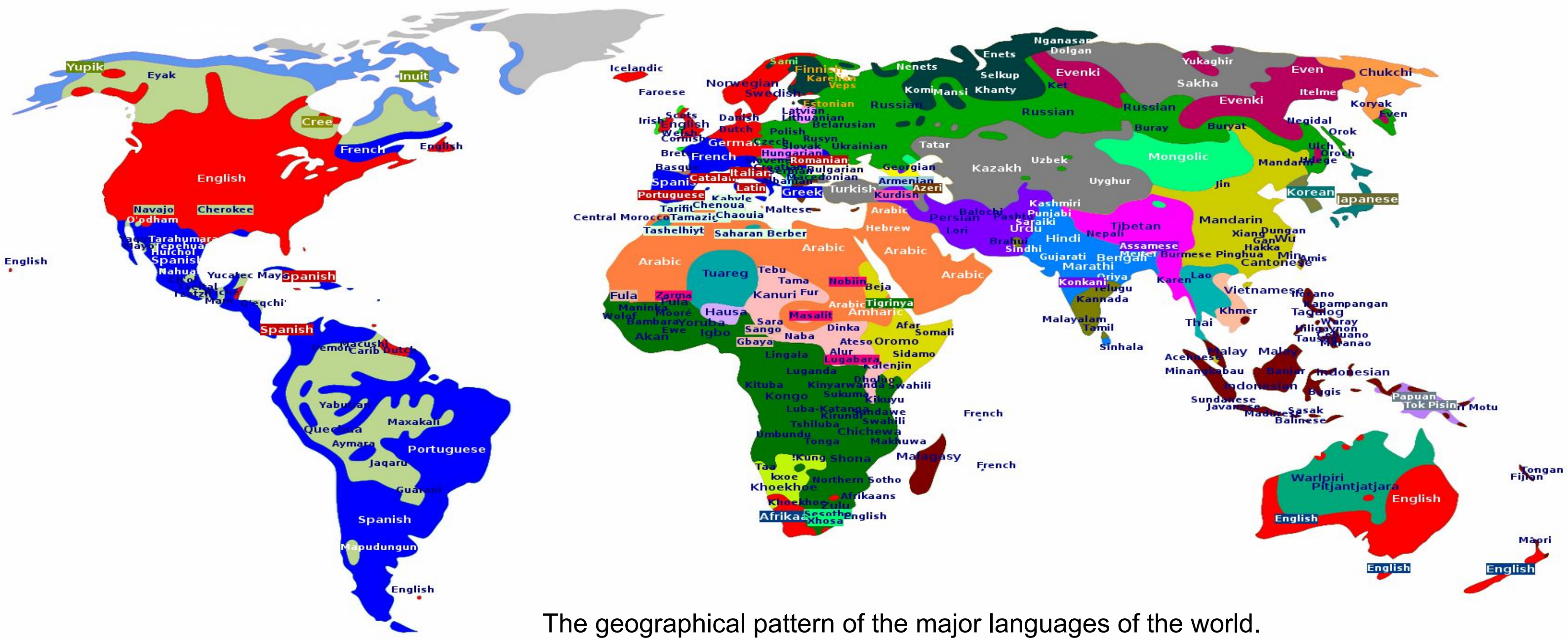


Image source:
https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India#/media/File:Language_region_maps_of_India.svg

Multilinguality

- The ethnologue.com website lists over **7000** languages in the world.



The geographical pattern of the major languages of the world.
Source: https://en.wikipedia.org/wiki/Template:Distribution_of_languages_in_the_world

Need for Language Resource

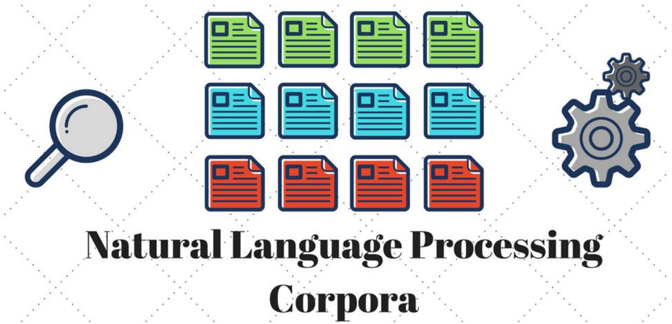
- Wikipedia has texts in 313 languages.
- Natural language technology development depends on large numbers of language resources (text / speech).
- Lack of language resources affects the development of natural language technologies.

Agenda

- Overview
- **Corpus Development**
- Case Study1 : Hindi Visual Genome
- Case Study2 : Malayalam Visual Genome
- Conclusion

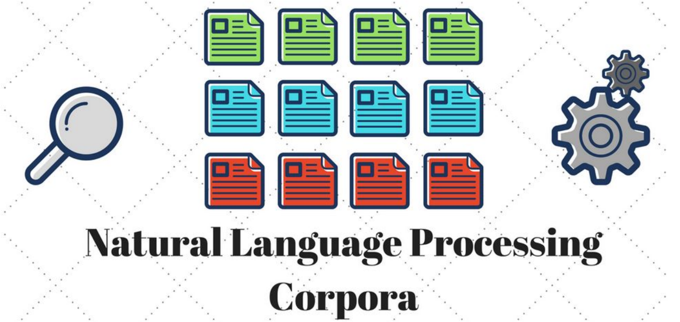
Corpus

- Corpus (plural corpora) : A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech.
- A corpus can be made up of everything from newspapers, novels, recipes and radio broadcasts to television shows, movies and tweets.
- In NLP, a corpus contains text and speech data that can be used to train AI and machine learning systems.
- Generally, the larger the size of a corpus, the better (prioritize quantity over quality).



Corpus

- High quality data is crucial
  - Accuracy
    - Ensure values and metadata contained within the corpus are accurate so the machine learning algorithm can learn to perform a task efficiently and effectively.
  - Completeness
    - Ensuring that the data in the corpus doesn’t have any gaps or missing information.
  - Timeliness
    - Making sure the corpus is up-to-date and the data remains relevant.
- Clean Data (eliminate any errors or duplicate data)
- Balance



Source:
https://www.definedcrowd.com/the-challenge-of-building-corpus-for-nlp-libraries/

Corpus - How to Build ?

- Data Collection
  - Data type
    - Text/Image/Speech/Video
  - Identify source
    - Web, Social Media, Books, Recordings
  - Web scraping
    - Identify URLs (e.g. language, text, tags)
  - Bots
  - Optical Character Recognition (OCR)
  - Extract data
    - tools: Python, BeautifulSoup
- Data Processing
  - segmentation, alignment
    - Purnaviram, Hunalign
- Finalization and Release
  - Split train/dev/test set
  - Baseline
  - License
  - Release platform
  - Share/organize shared task
    - WMT, WAT, ICON, etc...

Image source:
https://medium.com/analytics-vidhya/web-scraping-and-coursera-8db6af45d83f

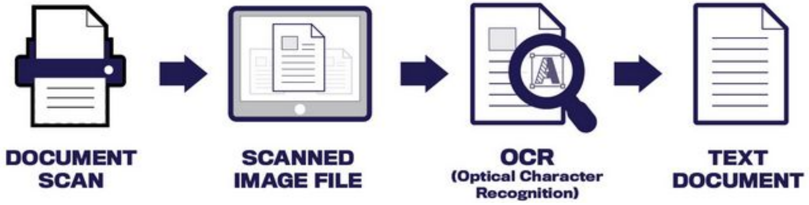


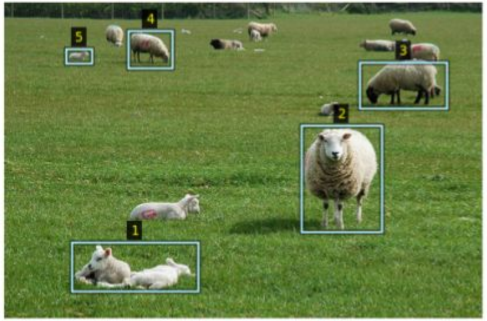
Image source: Image source:
https://medium.com/states-title/using-nlp-bert-to-improve-ocr-accuracy-385c98a
e174c

Agenda

- Overview
- Corpus Development
- **Case Study1 : Hindi Visual Genome**
- Case Study2 : Malayalam Visual Genome
- Conclusion

Motivation

Do Visual Context Disambiguates ?



Caption 1: Two lambs lying in the sun.
Hindi MT: दो भेड़ के बच्चे सूरज में **झूठ बोल** रहे है
Gloss: Two baby sheep are **telling lies** …

Selected surrounding captions:

2. Sheep standing in the grass
3. Sheep with black face and legs
4. Sheep eating grass
5. Lamb sitting in grass.

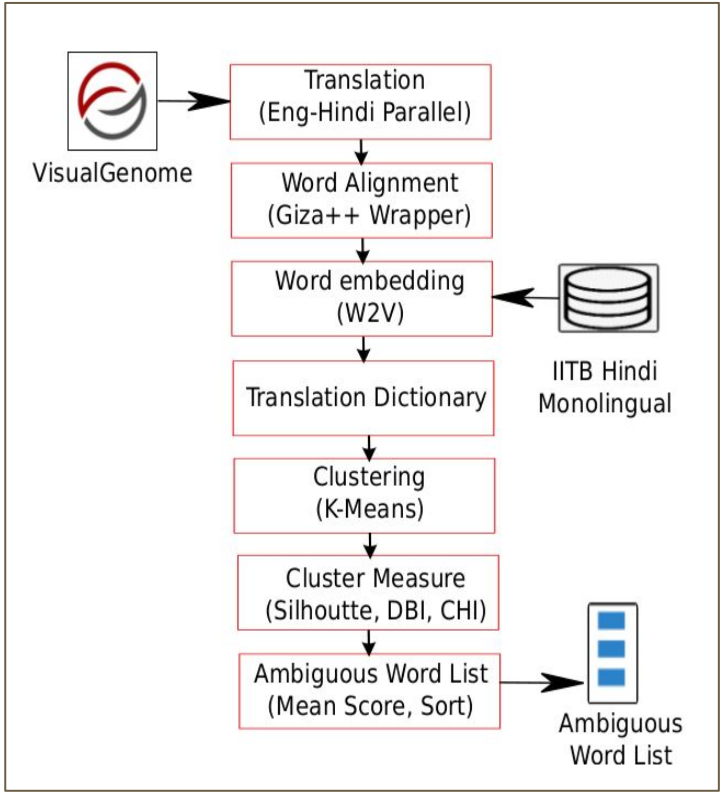
Multimodal Corpus

- Multi-modal content is gaining popularity in machine translation (MT) community due to its appealing chances to improve translation quality.
- It has application in commercial application
    - Translation of of image captions in online news article
    - Machine translation of e-commerce product listings.
- Although neural machine translation (NMT) models very good for large parallel texts, some inputs can remain genuinely ambiguous, especially if the input context is limited.
    - Exa: “mouse” in English (source) which can be translated into different words in Hindi based on the context (e.g. either a computer mouse or a small rodent)

Steps (Training and Test)

- The starting point were 31,525 randomly selected images from Visual Genome
- We translated all 31,525 captions into Hindi using the NMT model (Tensor-to-Tensor)
- We uploaded the image, the source English caption and its Hindi machine translation into a "Translation Validation Website"
- Volunteers post-edited all the Hindi translations.
- We manually verified and finalized the post-edited files to obtain the training and test data.

Steps (Challenge Test set)



Overall pipeline for ambiguous word finding from input corpus

1. Translate all English captions of visual Genome (3.15 million unique strings) using Google translate.
2. Apply word alignment on the synthetic parallel corpus using GIZA++ Wrapper.
3. Extract all pairs of aligned words in the form of a "translation dictionary". Dictionary contains key/value pairs of the English word (E) and all its Hindi translations (H_1 , H_2, . . . H_n), E → {H_1 , H_2, . . . H_n }.
4. Train Hindi word2vec (W2V) word embeddings. We used the gensim implementation and trained it on IITB Hindi Monolingual Corpus which contains about 45 million Hindi sentences.
5. For each English word from the translation dictionary, get all Hindi translation words and their embeddings.
6. Apply K-means clustering algorithm to the embedded Hindi words to organize them according to their word similarity.
7. Evaluate the obtained clusters with the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harbaz Index (CHI).
8. Sort the list in descending order to get the most ambiguous words.
9. Manually check the list and extract the most ambiguous English words.

Challenge Test set

	Word	Segment Count
1	Stand	180
2	Court	179
3	Players	137
4	Cross	137
5	Second	117
6	Block	116
7	Fast	73
8	Date	56
9	Characters	70
10	Stamp	60
11	English	42
12	Fair	41
13	Fine	45
14	Press	35
15	Forms	44
16	Springs	30
17	Models	25
18	Forces	9
19	Penalty	4
	Total	1400

Challenge test set: ambiguous words

Image		
	English Input:	gold religious **cross** on top of golden ball
	Translated Output:	सोने की गेंद के शीर्ष पर स्वर्ण धार्मिक क्रॉसें .
	Gloss:	Gold religious cross on top of golden ball
	English Input:	a blue wall beside tennis **court**
	Translated Output:	टेनिस कोर्ट के पास एक नीली दीवार हैं ।
	Gloss:	Blue wall near the tennis court
	English Input:	the tennis **court** is made up of sand and dirt
	Translated Output:	टेनिस कोर्ट रेत और गंदगी से बनी है।
	Gloss:	Tennis court is made of sand and dirt
	English Input:	A crack on the **court**
	Translated Output:	अदालत पर एक crack
	Gloss:	A crack on the judicial court

Sample Challenge Test set machine translation output (ENHI Multimodal Task, WAT 2019)
System description paper: Idiap NMT System for WAT 2019Multi-Modal Translation Task

Availability

Hindi Visual Genome

Hindi-English Multimodal Dataset

https://ufal.mff.cuni.cz/hindi-visual-genome

Hindi Visual Genome 1.0	Hindi Visual Genome 1.1
Used in WAT 2019	Used in WAT 2020, Using in WAT2021

WAT 2019 ENHI Multimodal Task

- English→Hindi multimodal translation task is based on the first English-Hindi multi-modal corpus (Hindi Visual Genome, HVG in short).
- Multi-modal task is introduced first time in WAT 2019.

Street sign advising of penalty.

The penalty box is white lined.

An illustration of two meanings of the word “penalty” exemplified with two images (Hindi Visual Genome)

Dataset

Dataset	Items	Tokens English	Tokens Hindi
Training Set	28,932	143,178	136,722
D-Test	998	4,922	4,695
E-Test (EV)	1,595	7,852	7,535
C-Test (CH)	1,400	8,185	8,665

Data for the English→Hindi multi-modal translation task. One item consists of source English sentence, target Hindi sentence, and a rectangular region within an image. The total number of English and Hindi tokens in the dataset also listed. The abbreviations EV and CH are used in the official task names in WAT scoring tables.

Source Text : Man stand of skateboard
Reference : आदमी स्केटबोर्ड पर खड़ा है

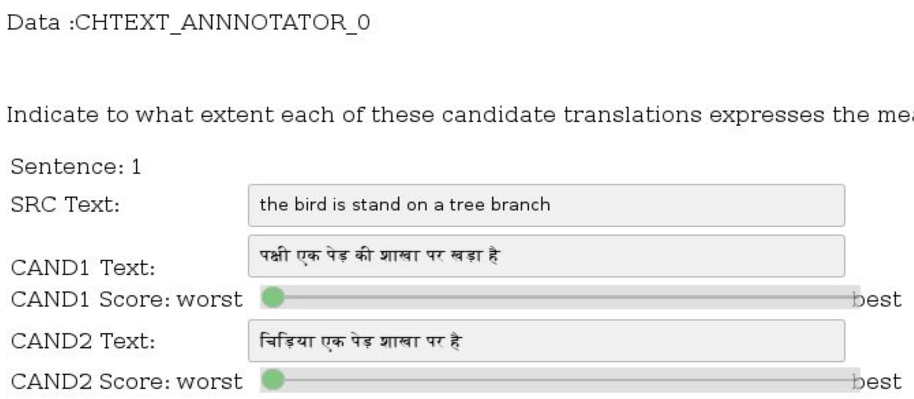
Illustration of an item

Tracks

- **Text-Only Translation (labeled “TEXT” in WAT official tables) :** The task is to translate short English captions (text) into Hindi. No visual information can be used. (need to be specified other resources if used in the corresponding system description paper).
- **Hindi Captioning (labeled “HI”):** The task is generate captions in Hindi for the given rectangular region in an input image.
- **Multi-Modal Translation: (labeled “MM”):** Given an image, a rectangular region in it and an English caption for the rectangular region, the task is to translate the English text into Hindi. Both textual and visual information can be used.

Results (Manual Evaluation)

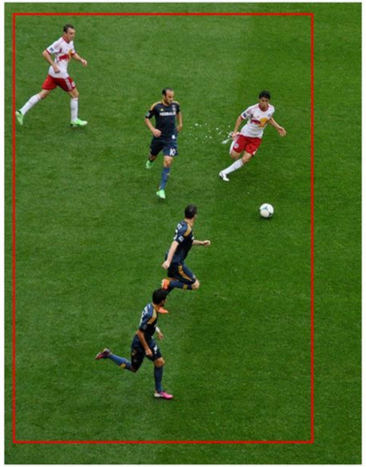
- Manual Evaluation follow Direct Assessment (DA) technique by asking annotator to assign 0-100 for each candidate.
- Collected DA scores averaged for each system and track (denoted “Ave”).
- Standardized per annotator and then averaged (denoted “Ave Z”).
  - Scores are scaled, so average score of each annotator is 0 and standard deviation is 1.



Manual evaluation of text-only translation.

Results (Manual Evaluation)

Data :CHHI_ANNNOTATOR_1



Sentence: 1

Indicate how plausible these captions are for the highlighted area of the image.
Judge each of the captions independently of the other. Each of the captions may be focusing on a different aspect of the area in the image.

CAND1 Text: टेनिस खेल

CAND1 Score: worst best

CAND2 Text: फुटबाल खिलाड़ी एक्शन में

CAND2 Score: worst best

Manual evaluation of Hindi captioning.

Data :CHMM_ANNNOTATOR_3

Sentence: 1

Is the English text (SRC) a good caption for the highlighted area of the image? : Yes No

SRC Text: Four baseball players on field.

Indicate to what extent each of these candidate translations expresses
the meaning of the English source text (independently of the other candidate).

CAND1 Text: क्षेत्र में बेसबॉल खिलाड़ी

CAND1 Score: worst best

CAND2 Text: क्षेत्र में चार बेसबॉल खिलाड़ी।

CAND2 Score: worst best

Manual evaluation of multi-modal translation.

Results (Manual Evaluation)

	Team ID	Data ID	Ave	Ave Z
EV TEXT	IDIAP	2956	72.85	0.70
	Reference		71.34	0.66
	683	3285	68.89	0.57
	683	3286	61.64	0.36
	NITSNLP	3299	52.53	0.00
CH TEXT	**Reference**		79.23	0.94
	IDIAP	3277	60.81	0.25
	IDIAP	3267	60.17	0.25
	683	3284	45.69	-0.28
	683	3287	45.52	-0.24
	NITSNLP	3300	28.48	-0.81
EV MM	**Reference**		70.04	0.60
	683	3271	69.17	0.61
	PUP-IND	3296	62.42	0.35
	PUP-IND	3295	60.22	0.28
	NITSNLP	3288	58.98	0.25
CH MM	**Reference**		75.96	0.76
	683	3270	54.51	0.08
	NITSNLP	3298	48.45	-0.20
	PUP-IND	3281	48.06	-0.13
	PUP-IND	3280	47.06	-0.17
EV HI	**Reference**		68.80	0.52
	NITSNLP	3289	51.78	-0.05
CH HI	**Reference**		72.60	0.61
	NITSNLP	3297	44.46	-0.35
	683	3304	26.54	-0.94

Manual evaluation result for WAT Multi-Modal Tasks.

HVG Validation

- One of the participant team spotted few error in the HVG dataset.
- We made use of the manual annotations to validate English sources in HVG.

Source Good?	C-Test	E-Test
Yes	1586 (78.7 %)	1348 (66.9 %)
No	20 (1.0 %)	46 (2.3 %)
No Answer	410 (20.3 %)	622 (30.9 %)
Total	2016 (100.0 %)	2016 (100.0 %)

Appropriateness of source English captions in the 4032 assessments collected for the multi-modal track.

Results (WAT2020)

	Team ID	Data ID	Ignoring Unscored Ave	Ignoring Unscored Ave Z	Unscored = Worst Ave	Unscored = Worst Ave Z
EV TEXT	ODIANLP	3711	83.38	0.34	78.25	0.53
	Reference	-	82.19	0.29	75.14	0.47
	CNLP-NITS	3897	80.01	0.23	70.46	0.37
	2019:IDIAP	2019:2956	76.94	0.15	67.64	0.30
	iiitsc	4030	74.27	0.07	58.60	0.07
CH TEXT	Reference	-	88.07	0.47	85.44	0.71
	ODIANLP	3713	75.21	0.08	63.60	0.18
	2019:IDIAP	2019:3277	67.79	-0.10	56.29	0.03
	CNLP-NITS	3898	59.61	-0.38	40.40	-0.36
	iiitsc	4031	54.85	-0.53	36.78	-0.47
EV MM	Reference	-	86.82	0.45	83.82	0.68
	CNLP-NITS	3896	81.75	0.28	73.78	0.43
	2019:638	2019:3271	74.82	0.07	63.47	0.18
	2019:NITSNLP	2019:3288	59.31	-0.39	42.88	-0.31
CH MM	Reference	-	90.66	0.53	88.53	0.78
	CNLP-NITS	3894	68.72	-0.11	55.80	0.01
	2019:638	2019:3270	57.03	-0.45	41.79	-0.33
EV HI	Reference	-	90.26	0.53	80.45	0.58
	ODIANLP	3779	47.16	-0.73	10.69	-1.10
CH HI	Reference	-	88.94	0.51	78.53	0.53
	2019:NITSNLP	2019:3297	58.56	-0.37	21.29	-0.84
	ODIANLP	3759	52.10	-0.57	10.47	-1.11

Manual evaluation result for WAT2020 Multi-Modal Tasks.

Results (WAT2020)

	English Input: Translated Output: Gloss:	a man trying to **cross** एक आदमी क्रॉस करने की कोशिश कर रहा है A man trying to cross
	English Input: Translated Output: Gloss:	the woman is waiting to **cross** the street महिला सड़क पार करने की प्रतीक्षा कर रही है। The woman is waiting to cross the street
	English Input: Translated Output: Gloss:	the lady appears to be going **cross** country skiing लगता है कि महिला क्रॉस कंट्री स्कीइंग जा रही है It seems that the lady is going for cross country skiing
	English Input: Translated Output: Gloss:	a **cross** sign on top of the tower टॉवर के शीर्ष पर एक पार संकेत A par sign on top of tower

Sample Challenge Test set machine translation output (ENHI Multimodal Task, WAT 2020)
System description paper: ODIANLP’s Participation in WAT2020

Agenda

- Overview
- Corpus Development
- Case Study1 : Hindi Visual Genome
- **Case Study2 : Malayalam Visual Genome**
- Conclusion

Malayalam Visual Genome (MVG)

- Malayalam Visual Genome (MVG) has the similar goal as HVG for Malayalam language.
- MVG is a multimodal dataset consisting of text and images.
- First Multi-modal dataset in Malayalam for multimodal translation and image captioning tasks.
- MVG contains 29K segments for training, 1K and 1.6K segments are provided in development and test sets, and additional challenge test set consists of 1.4K segments.
- Prepared by the native speakers [postediting](#).

Malayalam Visual Genome

English-Malayalam Multimodal Dataset

https://ufal.mff.cuni.cz/malayalam-visual-genome

[Malayalam Visual Genome 1.0](#)

Using in [WAT2021](#)

Malayalam Visual Genome (MVG)

Sample items from the randomly selected segments (train/dtest/etest)

Image	Image_ID	X	Y	Width	Height	English Text	Malayalam Text
	2323457	20	150	325	121	Many giraffes at a zoo	ഒരു മൃഗശാലയിലെ നിരവധി ജിറാഫുകൾ
	2335684	61	191	437	182	Fruit stand outside market	ഫ്രൂട്ട് സ്റ്റാൻഡ് മാർക്കറ്റിന് പുറത്താണ്

Sample item from the challenge test set (chtest)

Image	Image_ID	X	Y	Width	Height	English Text	Malayalam Text
	2372733	26	107	152	218	The tennis court is made up of sand and dirt	ടെന്നീസ് കോർട്ട് മണലും അഴുക്കും ചേർന്നതാണ്

Conclusion and Possible Research Direction

- Findings
  - Text-only system with larger data outperformed multi-modal systems.
  - It has been observed that in many instances image helps to resolve ambiguities.
- Research direction
  - Do you think image can helps disambiguation ?. Verify where the machine translation system fails for the Indian languages. Try to to analyze how to resolve this issue. Can you able to generate a challenging test set for this ?.
  - Can we generate better captions using (HVG/MVG) utilizing the regions ?.
  - How to utilize different modalities (text, image, speech) for corpus development ?.
- Going forward...
  - Building Multilingual Multimodal Corpus.

Agenda

- Overview
- Corpus Development
- Case Study1 : Hindi Visual Genome
- Case Study2 : Malayalam Visual Genome
- **Conclusion**

References

[1] Parida, S., & Bojar, O. (2018). **Translating Short Segments with NMT: A Case Study in English-to-Hindi**. In 21st Annual Conference of the European Association for Machine Translation (p. 229).

[2] Parida, S., Bojar, O., & Dash, S. R. (2019). **Hindi Visual Genome: A Dataset for Multi-Modal English to Hindi Machine Translation**. *Computación y Sistemas*, *23*(4), 1499-1505.

[3] Parida, S., Bojar, O., & Motlicek, P. (2019, November). **Idiap NMT System for WAT 2019 Multimodal Translation Task.** In Proceedings of the 6th Workshop on Asian Translation (pp. 175-180).

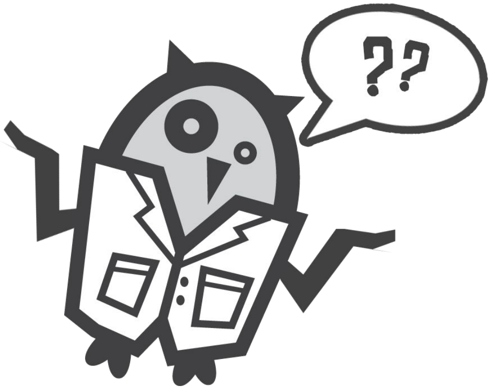
[4] Parida, S., Motlicek, P., Dash, A. R., Dash, S. R., Mallick, D. K., Biswal, S. P., ... & Bojar, O. (2020, December). **ODIANLP's Participation in WAT2020**. In *Proceedings of the 7th Workshop on Asian Translation* (pp. 103-108).

[5] Nakazawa, T., Higashiyama, S., Ding, C., Mino, H., Goto, I., Kazawa, H., ... & Kurohashi, S. (2017, November). **Overview of the 4th Workshop on Asian Translation**. In Proceedings of the 4th Workshop on Asian Translation (WAT2017) (pp. 1-54).

[6] Nakazawa, T., Nakayama, H., Ding, C., Dabre, R., Higashiyama, S., Mino, H., ... & Kurohashi, S. (2020, December). **Overview of the 7th workshop on Asian translation**. In *Proceedings of the 7th Workshop on Asian Translation* (pp. 1-44).

[7] Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., ... & Kurohashi, S. **Overview of the 6th Workshop on Asian Translation**.

Q&A



Contact information:

- Email: shantipriya.parida@idiap.ch
- Twitter: @Shantipriyapar3
- Web : https://www.idiap.ch/~sparida/

Resource

- **OdiEnCorp 2.0** (Odia-English parallel corpus)
- **OdiEnCorp 1.0** (Odia-English parallel and Odia monolingual corpus)
- **Hindi Visual Genome 1.0** (English to Hindi Multimodal dataset)
- **Hindi Visual Genome 1.1** (English to Hindi Multimodal dataset)
- **Malayalam Visual Genome 1.0** (English to Malayalam Multimodal dataset)
- **English->Hindi Machine Translation System**
- **Odia-NLP-Resource-Catalog** (A catalog for Odia language NLP resources)

Note: The released corpora are available freely for non-commercial research purpose

Thanks to all of our collaborators

Dr. Shantipriya Parida
Idiap Research Institute, Switzerland

Assoc. Prof. Ondřej Bojar
Charles University, Czech Republic

Assoc. Prof. Satya Ranjan Dash
KIIT University, India

We are indebted to the researchers and volunteers associated with the corpora development and special thanks to the WAT shared task participants and researchers using our corpora and providing feedback for improvements.

Thank You