How can Research Help in Reducing the Language Gap on Wikipedia

WIKIMANIA
STOCKHOLM

Hady Elsahar
wd:Q50290890
[[User:Hadyelsahar]]
@hadyelsahar

Lucie-Aimée Kaffee
wd:Q37860261
[[User:Frimelle]]
@frimelle

- **Information Poverty** & **Language Gap** on Wikipedia

- Research Topics

- Scribe Project 

WIKIMANIA STOCKHOLM

Key Points

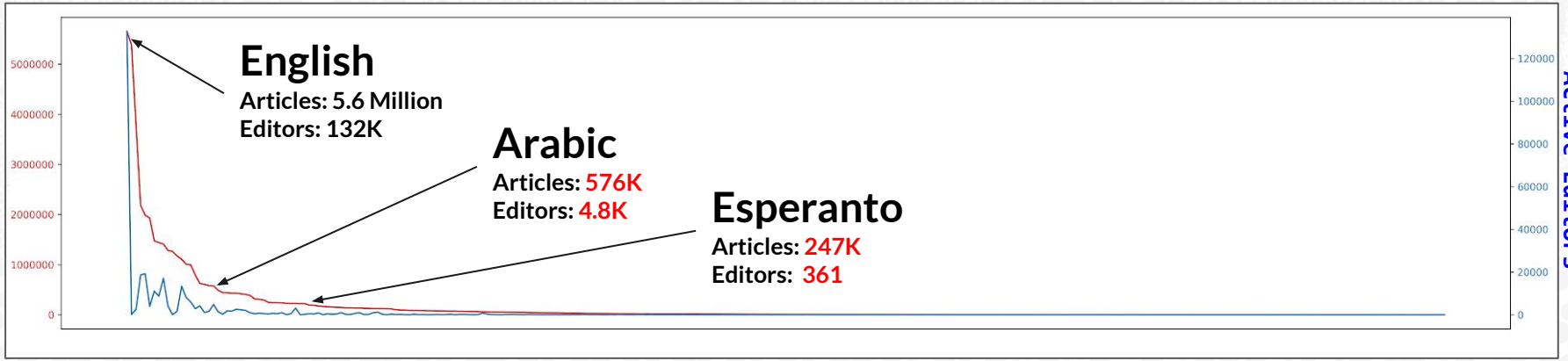
WIKIPEDIA
The Free Encyclopedia

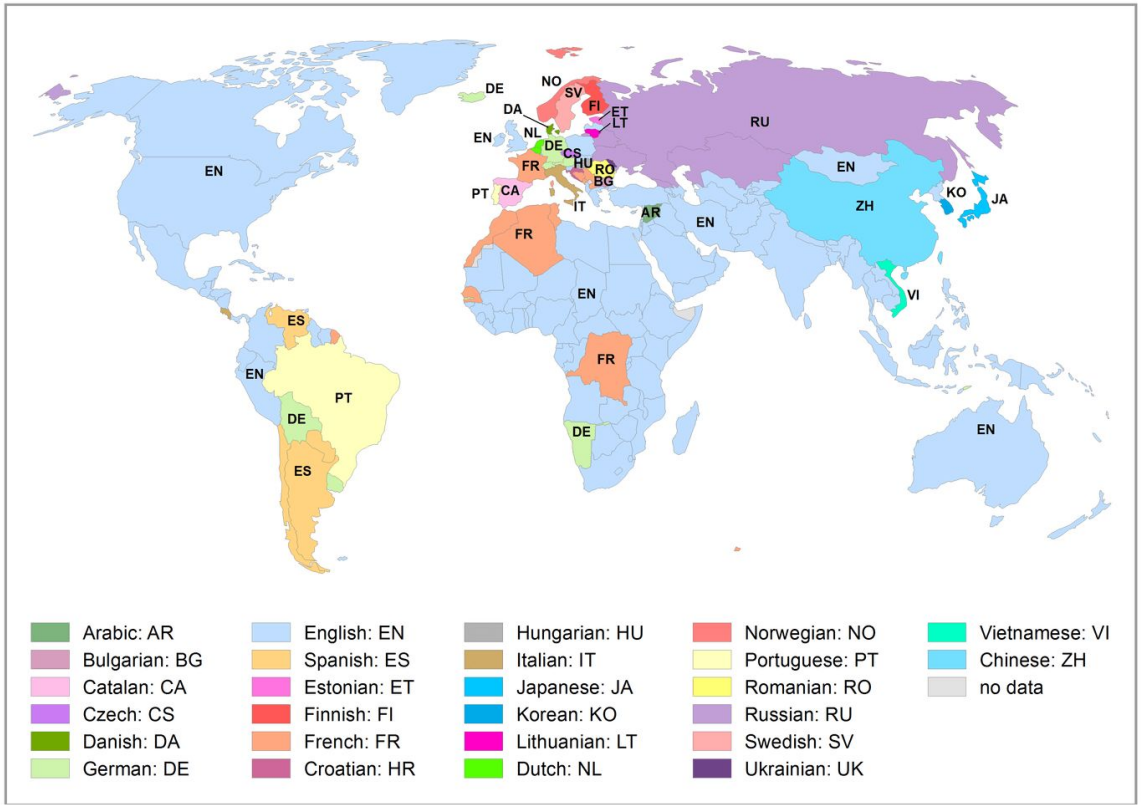
=

∑ All Human Knowledge

?

Language Distribution of Articles and Active Users on Wikipedia

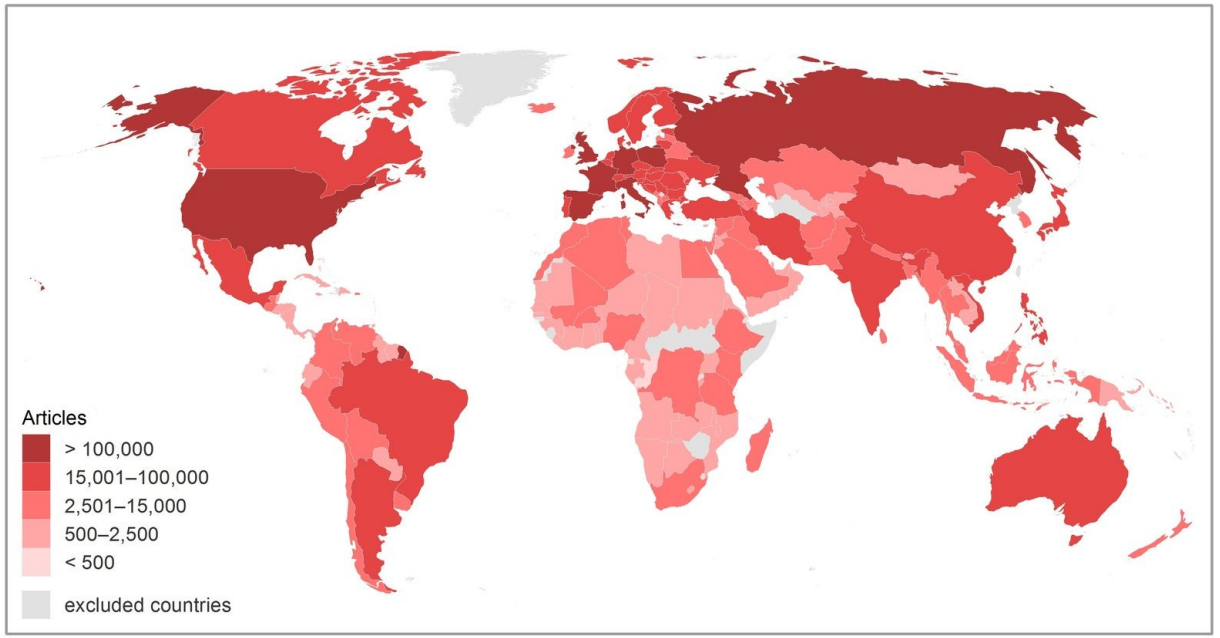




Dominant language of Wikipedia articles (geo tagged by country)
(Graham et. al. 2014)

Language Gap or Content Gap

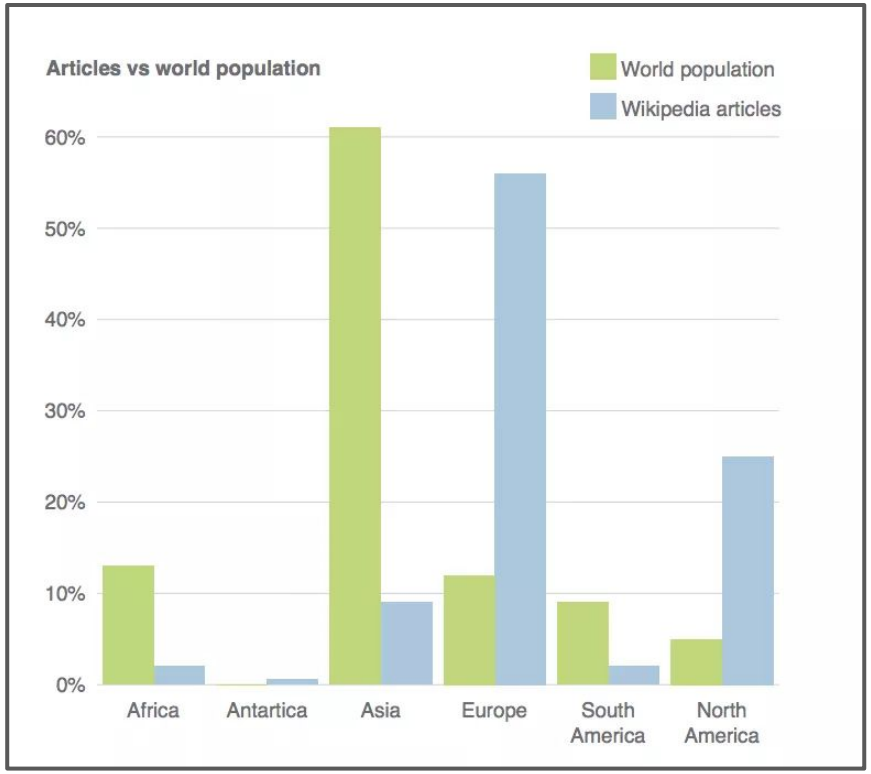
Few articles are being written by locals



Number of geotagged Wikipedia articles per country (Graham et. al. 2014)

Information Poverty

"There are more articles about Japan than about the whole middle east combined."



Articles vs world population (Graham et. al. 2011)

Information Poverty

The World relevant topics are not well represented on Wikipedia

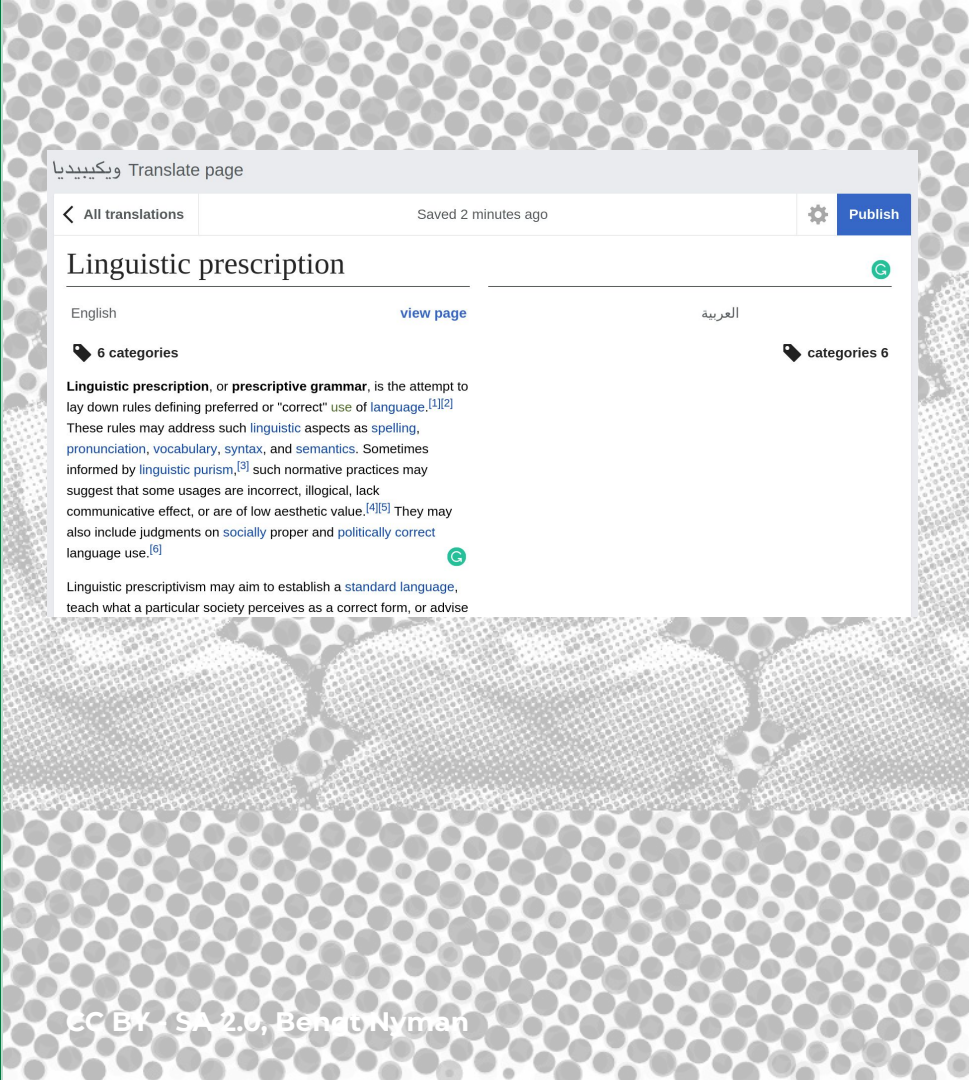
Solutions to Information poverty?

Current

- Automatic Machine Translation

Alternatives

- Natural Language Generation
- Summarization
- Cross Lingual Word Embeddings

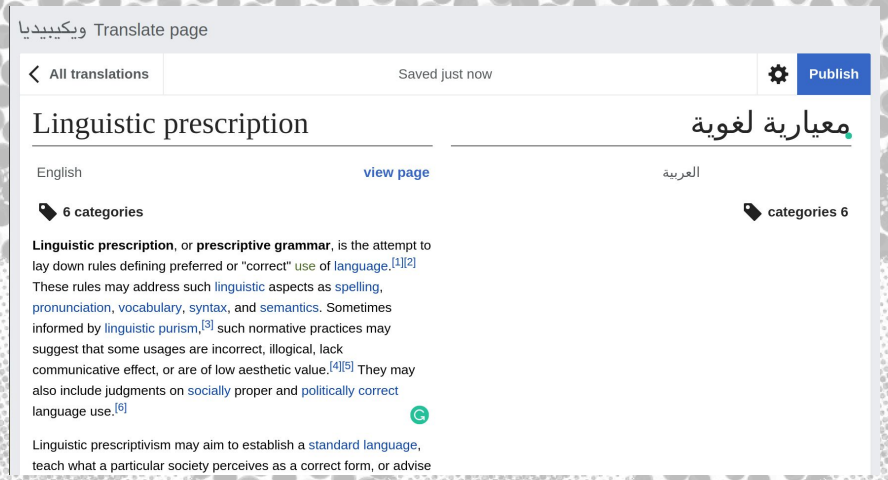


Content Translation Tool

Relies on automatic Machine Translation

- Google Translate
- Yandex Translate

WIKIMANIA STOCKHOLM

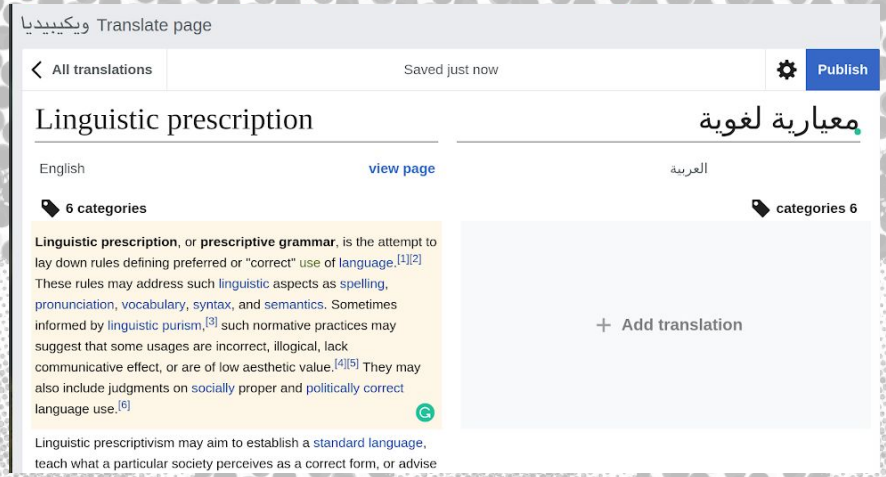


Content Translation Tool

Relies on automatic Machine Translation

- Google Translate
- Yandex Translate

WIKIMANIA STOCKHOLM



Content Translation Tool

Relies on automatic Machine Translation

- Google Translate
- Yandex Translate

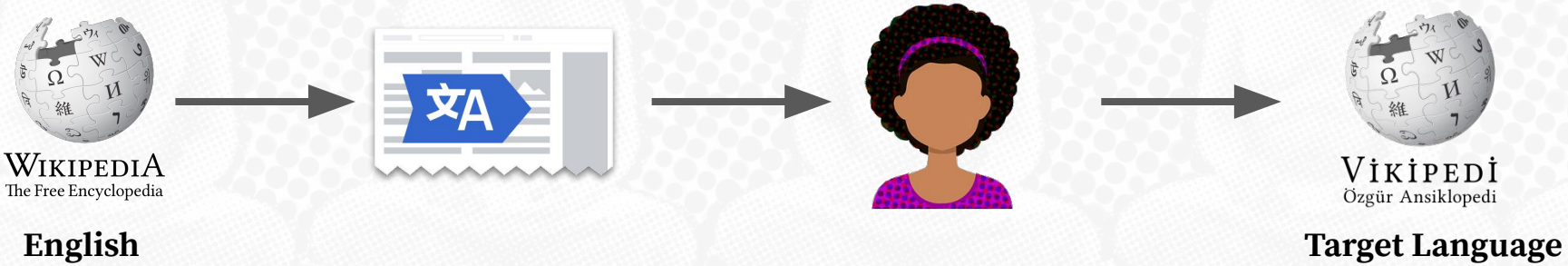
WIKIMANIA STOCKHOLM

Content Translation Tool

Relies on automatic Machine Translation

- Google Translate
- Yandex Translate

Machine Translation with human in the loop.

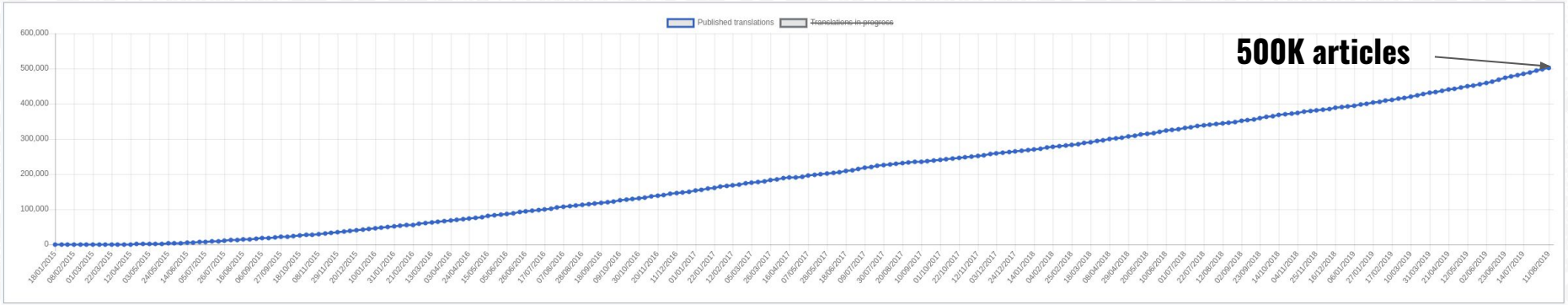


English

Target Language

Machine Translation with human in the loop

Successful in filling part of the language Gap



Translations to all languages

https://en.wikipedia.org/wiki/Special:ContentTranslationStats

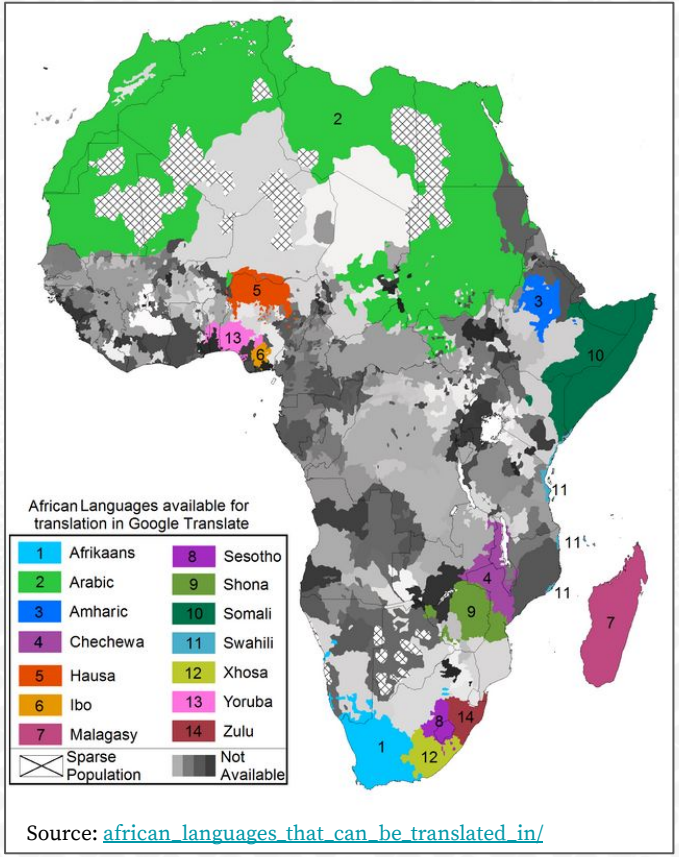
Machine Translation with human in the loop

Fails in representing all languages

301 on Wikipedia

103 on Google Translate

2000-6000 in the world



Source: [african_languages_that_can_be_translated_in/]

Machine Translation with human in the loop

Quality? For small languages

- No enough parallel data
- **Poor Quality** makes Translation is not useful to editors

Source: https://www.theverge.com/2019/5/8/18526739/wikipedia-translation-tool-machine-learning-ai-english

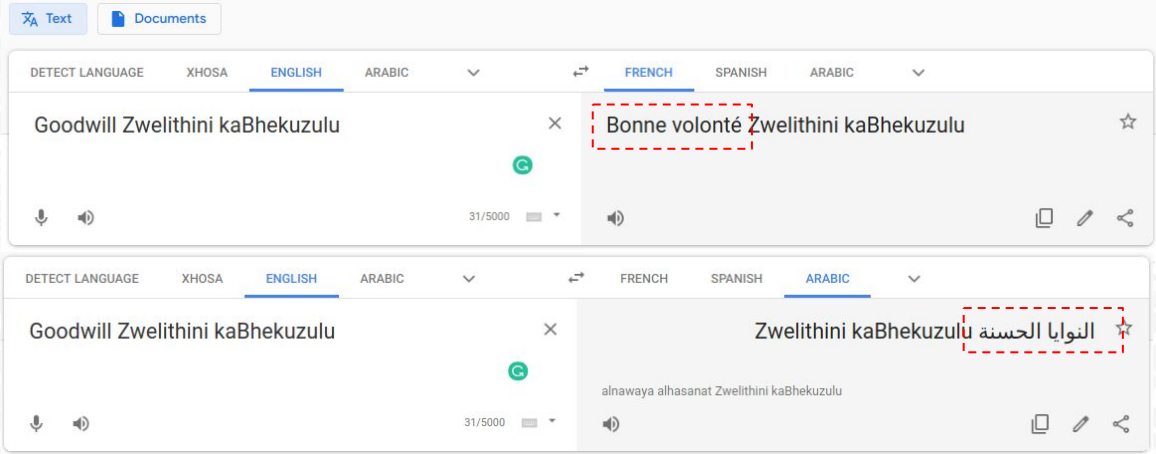
Machine Translation with human in the loop

Goodwill Zwelithini

King of the Zulus

Reign	17 September 1968 – present
Coronation	3 December 1971
Predecessor	Cyprian Bhekuzulu kaSolomon
Born	July 14, 1948 (age 71) Nongoma, Union of South Africa
Wives	Mantfombi Dlamini (great wife) Sibongile Winifred Dlamini Buthle MaMathe Thandekile Jane Ndlovu Nompumelelo Mchiza Zola Zelusiwe Mafu
Issue	27 including: [show]
Full name	Goodwill Zwelithini kaBhekuzulu
House	House of Zulu
Father	Cyprian Bhekuzulu kaSolomon

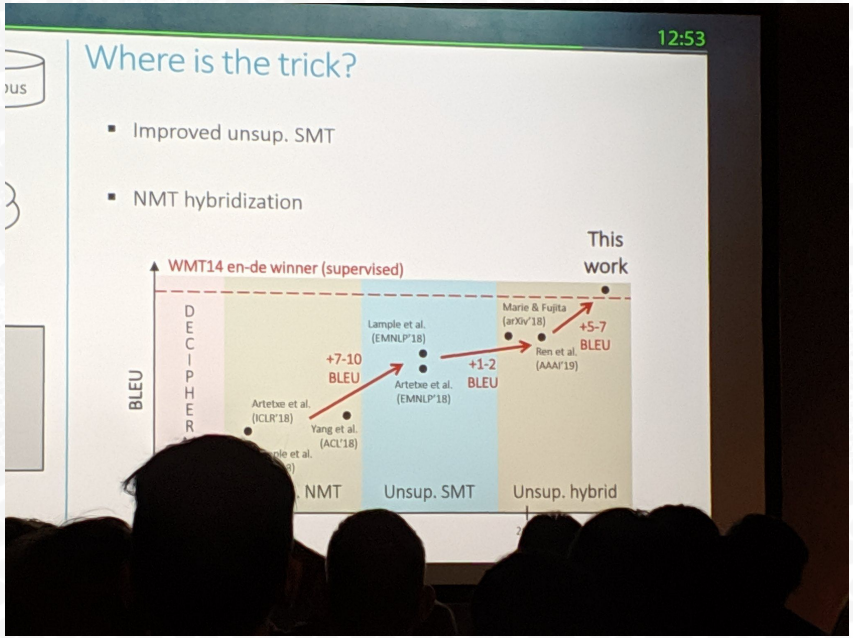
Translation of Labels?



Unsupervised Machine Translation

A solution to **Quality** and **Coverage**

- No need for parallel corpora
- Very suitable for small languages
- Getting better rapidly



(Artetxe et al. 2019) An Effective Approach to Unsupervised Machine Translation. ACL2019

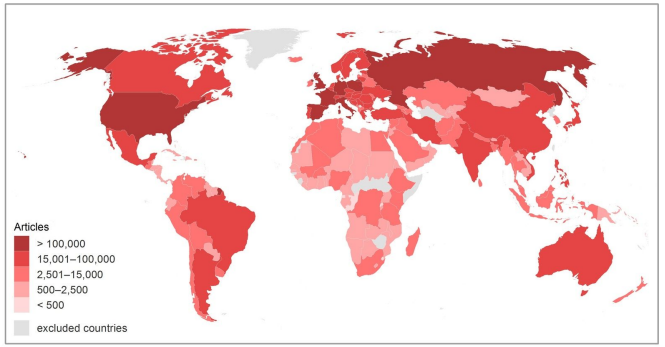
Solutions

Label Translation

Translation from English will not solve Information Poverty

Large language wikipedias
get to decide what is notable.

Small Language Wikipedias
will enforce that bias through translations.



Number of geotagged Wikipedia articles per country (Graham et. al. 2014)

Translation of English References

Harms

The Neutral Point of View

English: Alaa Salah is a Sudanese student and anti-government protester. The image of Salah has been dubbed as "Woman in White" or "Lady Liberty" of Sudan.

آلاء صلاح [edit|edit source]

Wikidata: آلاء صلاح – Q63126345: "ناشطة سودانية" · Aliases: None

آلاء صلاح في الصورة التي أصبحت مشهورة كرمز للاحتجاجات السودانية 2018-19.

آلاء صلاح هي طالبة سودانية ومتظاهرة مناهضة للحكومة. اكتسبت شهرتها من صورة التقطتها (لانا هارون)، وانتشرت في شهر أبريل 2019. وأصبحت آلاء ايقونة ورمز للمرأة في الثورة السودانية والتي تعرف باسم "كنداكة" أو "الملكة النوبية" أو "تمثال الحرية السودانية".[1][2][3]

Contents [hide]

1 الحياة المبكرة والتعليم
2 احتجاجات السودان
3 الصورة
4 انظر أيضا
5 المراجع

الحياة المبكرة والتعليم [edit|edit source]

ولدت آلاء صلاح في عام 1996 أو 1997. وكانت والدتها تعمل مصممة أزياء ووالدها يعمل في البناء والتشييد.[4] تدرس الهندسة والعمارة في جامعة السودان العالمية في الخرطوم.[5]

- CNN (en)
- 2x Guardian (en)
- Le Monde (fr)
- BuzzFeed (en)
- ABC (en)
- smh (en)
- 2x NYT (en)
- Nzherald (en)
- France24(ar)

المراجع [edit|edit source]

1. ^ CNN، Gianluca Mezzofiore. "This woman has come to symbolize Sudan's protests". CNN. مؤرشف من الأصل في 7 مايو 2019. اطلع عليه بتاريخ 11 أبريل 2019.
2. ^ أ ب "Woman in white goes viral as symbol of Sudan's uprising". The Sydney Morning Herald (باللغة الإنجليزية). 11-04-2019. مؤرشف من الأصل في 29 أبريل 2019. اطلع عليه بتاريخ 11 أبريل 2019.
3. ^ "حيونى كنداكه".. امرأة سودانية تتحول لايقونة "الثورة"'. فرانس 24 / France 24. 2019-04-11. مؤرشف من الأصل في 12 أبريل 2019. اطلع عليه بتاريخ 19 أبريل 2019.
4. ^ أ ب "I was raised to love our home': Sudan's singing protester speaks out". The Guardian (باللغة الإنجليزية). Salih، Zeinab Mohammed (2019-04-10). ISSN 0261-3077. مؤرشف من الأصل في 9 مايو 2019. اطلع عليه بتاريخ 11 أبريل 2019.
5. ^ أ ب "This Woman Stood On Top Of A Car And Became An Icon Of Sudan's Historic Protests". BuzzFeed News (باللغة الإنجليزية). مؤرشف من الأصل في 29 مايو 2019. اطلع عليه بتاريخ 10 أبريل 2019.
6. ^ "Sudanese police fire on protesters demanding president step down". The Guardian (باللغة الإنجليزية). Reuters (2019-01-17). ISSN 0261-3077. مؤرشف من الأصل في 27 أبريل 2019. اطلع عليه بتاريخ 11 أبريل 2019.
7. ^ "EXCLUSIVE: Sudanese spy chief 'met head of Mossad to discuss Bashir succession plan'". Middle East Eye (باللغة الإنجليزية). مؤرشف من الأصل في 19 مايو 2019. اطلع عليه بتاريخ 11 أبريل 2019.
8. ^ "Le mouvement de protestation s'embrase au Soudan" (باللغة الفرنسية). 08-04-2019. مؤرشف من الأصل في 29 مايو 2019. اطلع عليه بتاريخ 11 أبريل 2019.
9. ^ "Sudan's Military to Make Announcement Amid Protests Against Omar Hassan al-Bashir". The New York Times (باللغة الإنجليزية). Mullany، Gerry (2019-04-11). ISSN 0362-4331. مؤرشف من الأصل في 15 مايو 2019. اطلع عليه بتاريخ 11 أبريل 2019.
10. ^ "Mood in Sudan shifts to anger as the army prepares to seize power". The Guardian (باللغة الإنجليزية). Maclean، Ruth (2019-04-11). ISSN 0261-3077. مؤرشف من الأصل في 29 مايو 2019. اطلع عليه بتاريخ 12 أبريل 2019.
11. ^ "Poetic photo of Sudan's 'Lady Liberty' sheds light on anti-government protests". ABC News (باللغة الإنجليزية). 10-04-2019. مؤرشف من الأصل في 11 أبريل 2019. اطلع عليه بتاريخ 11 أبريل 2019.
12. ^ أ ب "'It's Going to Be the Image of the Revolution'". The New York Times (باللغة الإنجليزية). Friedman، Vanessa (2019-04-10). ISSN 0362-4331. مؤرشف من الأصل في 28 مايو 2019. اطلع عليه بتاريخ 11 أبريل 2019.
13. ^ "The woman in white: Why a photo from the Sudan protests has gone viral". The Guardian (باللغة الإنجليزية). "O'Grady، Siobhán (2019-04-09). ISSN 1170-0777. مؤرشف من الأصل في 13 مايو 2019. اطلع عليه بتاريخ 10 أبريل 2019.

Source: https://en.wikipedia.org/wiki/Alaa_Salah (en, ar)

Current

- Automatic Machine Translation

Alternatives

- **Natural Language Generation**
- Summarization
- Cross Lingual Transfer Learning

Natural Language Generation

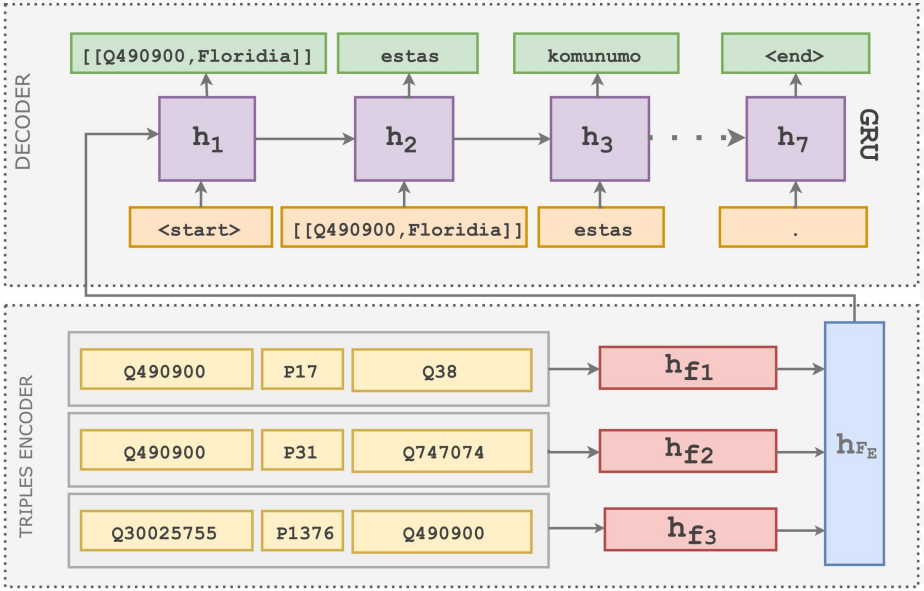
Generation of Wikipedia Articles from **Wikidata**

WIKIDATA

WIKIPEDIA
The Free Encyclopedia

WIKIMANIA STOCKHOLM

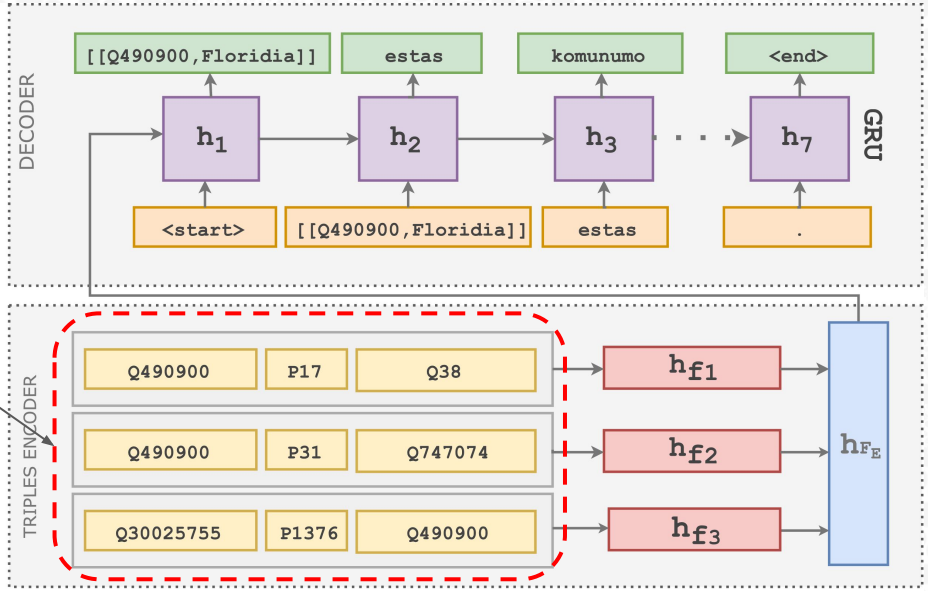
Using same type of Encoder-Decoder Models of MT



Learning to generate Wikipedia summaries for underserved languages from Wikidata (Kaffee, Elsahar & Vougiouklis et al. NAACL 2018)

Using same type of Encoder-Decoder Models of MT

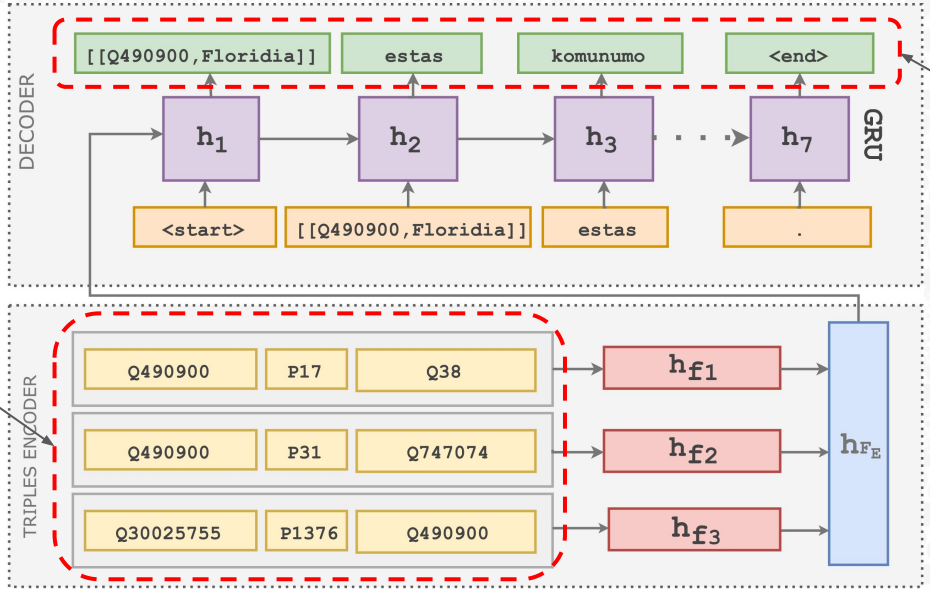
Encoded Wikidata Triples



Learning to generate Wikipedia summaries for underserved languages from Wikidata (Kaffee, Elsahar & Vougiouklis et al. NAACL 2018)

Using same type of Encoder-Decoder Models of MT

Encoded Wikidata Triples

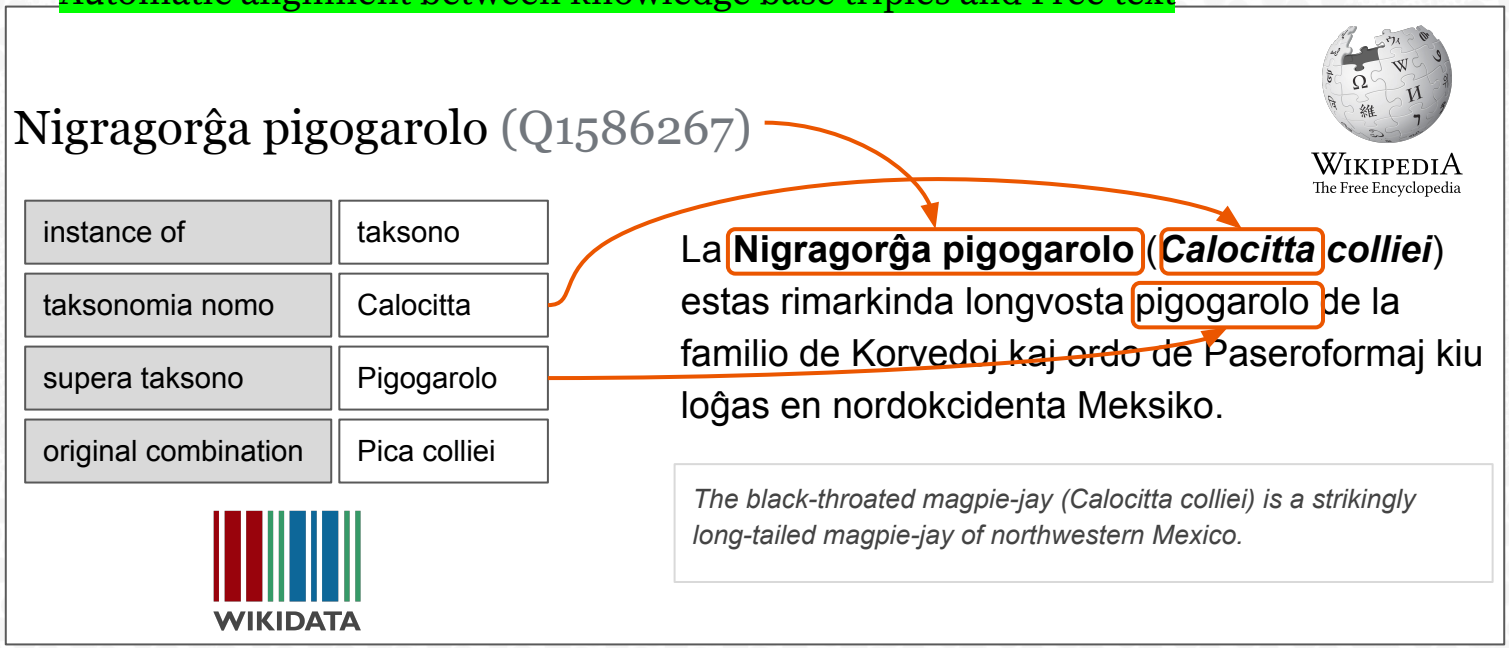


Generation of 1-2 sentences descriptions of the input triples.

Learning to generate Wikipedia summaries for underserved languages from Wikidata (Kaffee, Elsahar & Vougiouklis et al. NAACL 2018)

Annotated Data For Free

Automatic alignment between knowledge base triples and Free text



T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples (Elsahar et al. LREC 2017)

Good

- Integrating Wikidata knowledge
- Better than MT for small languages
- Translation of Names (using Wikidata)

Bad

- Generation of More than 1 Paragraph
- factual mistakes

Natural Language Generation

Generation of Wikipedia Articles from **Wikidata**

Current

- Automatic Machine Translation

Alternatives

- **Natural Language Generation**
- Summarization
- Cross-Lingual Transfer Learning

Building sentence repsentations using:

- Frequency
- Length
- Similarity
- Location
- Centrality

Optimizing:

- Min Redundancy
- Max Relevance
- Max Informativeness

A Simple Theoretical Model of Importance for Summarization
(Peyrard ACL2019)

WIKIMANIA STOCKHOLM

Multi-document Unsupervised Extractive Summarization

Current

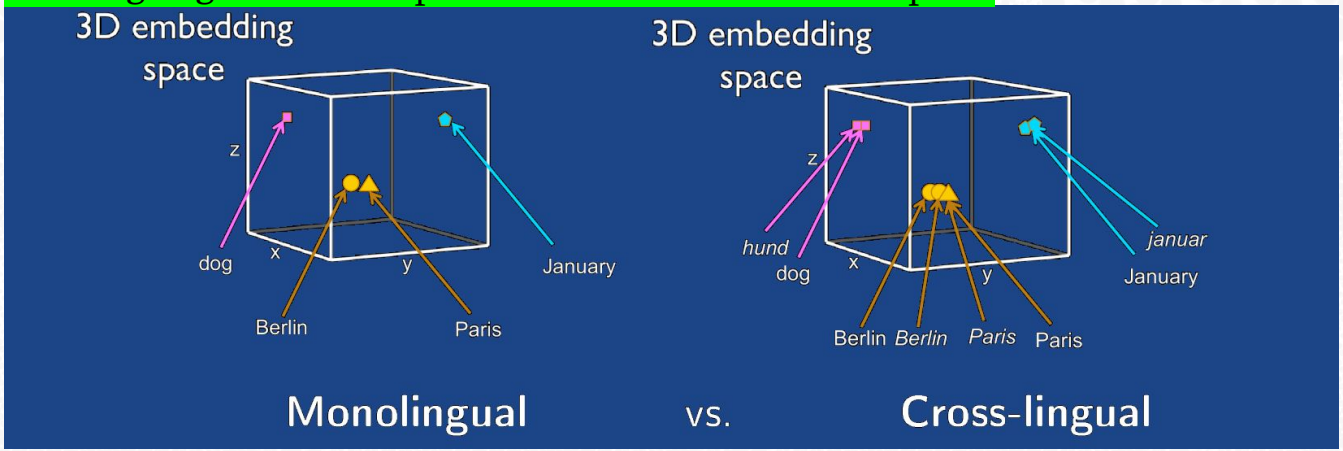
- Automatic Machine Translation

Alternatives

- **Natural Language Generation**
- Summarization
- Cross-Lingual Transfer Learning

Unsupervised Cross-Lingual Word Embeddings

Building aligned word representations in the feature space.



“Unsupervised” = Requires very minimal datasets

- Small bilingual dictionaries
- Weak supervision (shared vocabulary)

Source: https://tinyurl.com/xlingual

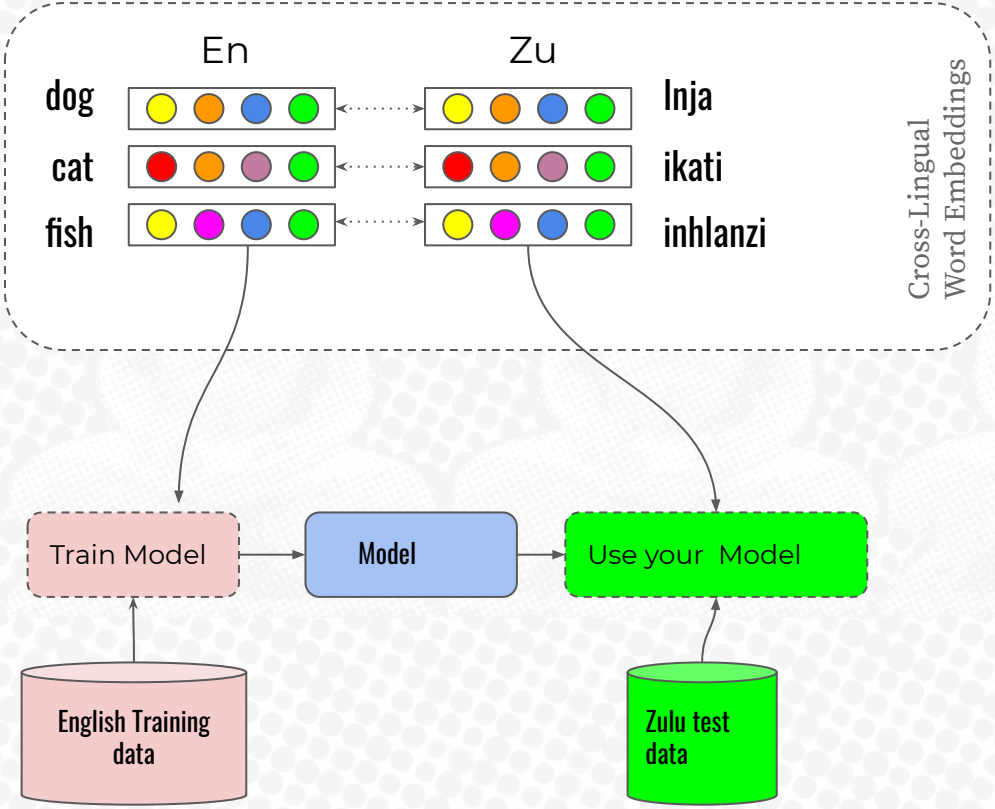
Unsupervised Cross-Lingual Word Embeddings

Useful for

“Cross lingual Transfer Learning”

Acquiring good results for tasks on languages with
limited/no annotated datasets (e.g. Zulu)

Using large annotated datasets available in other languages (e.g. English)

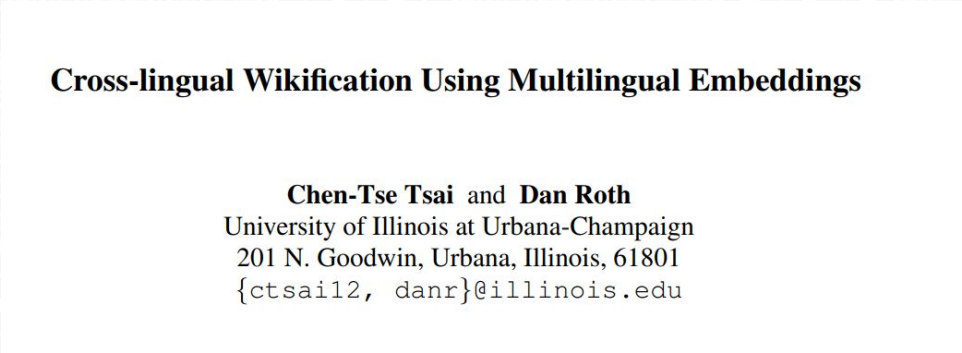


Adapted from: https://tinyurl.com/xlingual

Unsupervised Cross-Lingual Word Embeddings

Applications: Reference Retrieval

Retreival using **Entities** (Entity Linking)



Cross-lingual Wikification Using Multilingual Embeddings

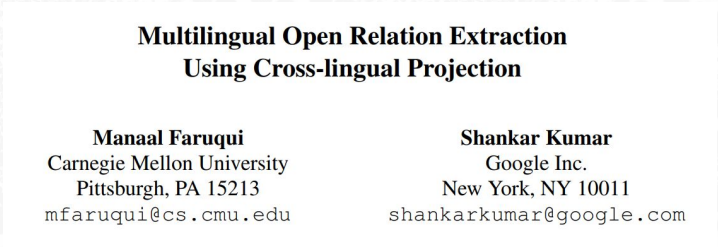
Chen-Tse Tsai and **Dan Roth**
University of Illinois at Urbana-Champaign
201 N. Goodwin, Urbana, Illinois, 61801
{ctsai12, danr}@illinois.edu

Cross-lingual Wikification Using Multilingual Embeddings (Tsai et al. NAACL 2016)

Unsupervised Cross-Lingual Word Embeddings

Applications: Reference Retrieval

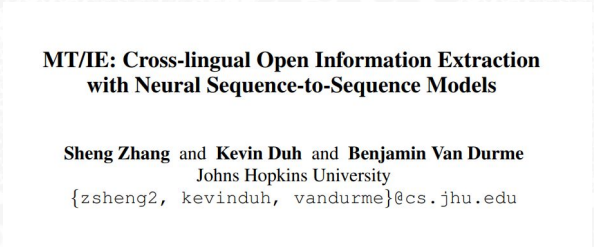
Retreival using **Facts** (OpenIE)



Multilingual Open Relation Extraction Using Cross-lingual Projection

Manaal Faruqui
Carnegie Mellon University
Pittsburgh, PA 15213
mfaruqui@cs.cmu.edu

Shankar Kumar
Google Inc.
New York, NY 10011
shankarkumar@google.com



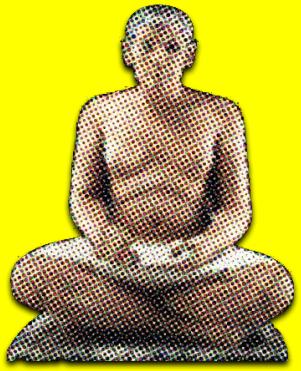
MT/IE: Cross-lingual Open Information Extraction with Neural Sequence-to-Sequence Models

Sheng Zhang and **Kevin Duh** and **Benjamin Van Durme**
Johns Hopkins University
{zsheng2, kevinduh, vandurme}@cs.jhu.edu

- Multilingual Open Relation Extraction Using Cross-lingual Projection (Faruqui et al. NAACL 2015)
- MT/IE: Cross-lingual Open Information Extraction with Neural Sequence-to-Sequence Models (Zhang et al. EACL2017)

Research solutions vs problems

	Language Gap	Information Poverty	Reference Bias	Quality	Label Translation	Long articles Generation
Machine Translation	✓	□	□	□	(□)	✓
NLG (data2text)	✓	✓	□	✓	(✓)	□
Summarization	✓	✓	✓	✓	(✓)	✓
Crosslingual Transfer Learning	—	—	✓	—	—	—



Scribe

Wikimedia Funded Project (2019/2020)

Helping editors of under-resourced languages to create new high-quality Wikipedia articles

https://meta.wikimedia.org/wiki/Scribe

Unsupervised Summarization human in the loop.

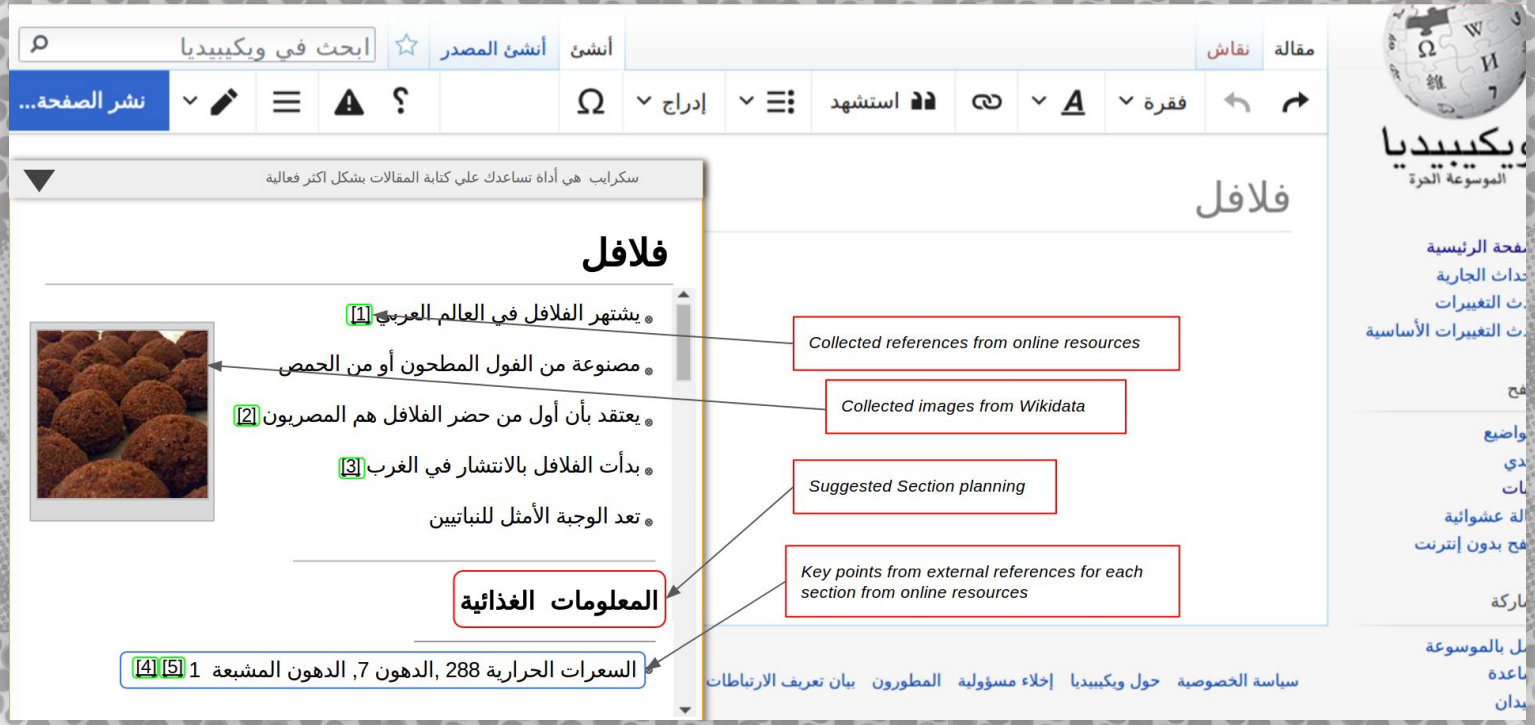
The world online Knowledge

Scribe

Target Wikipedia

Reference retrieval & filtering

Summarization


مقالة
نقاش
أنشئ
أنشئ المصدر
ابحث في ويكيبيديا
فقرة
استشهد
إدراج
نشر الصفحة...
فلافل
سكرايب هي أداة تساعدك علي كتابة المقالات بشكل اكثر فعالية
فلافل
يشتهر الفلافل في العالم العربي [1]
مصنوعة من الفول المطحون أو من الحمص
يعتقد بأن أول من حضر الفلافل هم المصريون [2]
بدأت الفلافل بالانتشار في الغرب [3]
تعد الوجبة الأمثل للنباتيين
المعلومات الغذائية
السعرات الحرارية 288 ,الدهون 7, الدهون المشبعة 1 [4] [5]
Collected references from online resources
Collected images from Wikidata
Suggested Section planning
Key points from external references for each section from online resources
سياسة الخصوصية
حول ويكيبيديا
إخلاء مسؤولية
المطورون
بيان تعريف الارتباطات

THANK YOU!

Interested in Scribe?

- **Know More:**
  https://meta.wikimedia.org/wiki/Scribe

- **You also care about filling the language Gap?**
  Collaborate with us:

Hady Elsahar
wd:Q50290890
[[User:Hadyelsahar]]
@hadyelsahar
hadyelsahar@gmail.com

Lucie-Aimée Kaffee
Q37860261
[[User:Frimelle]]
@frimelle
lucie.kaffee@gmail.com