

Mind the (Language) Gap: Neural Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders

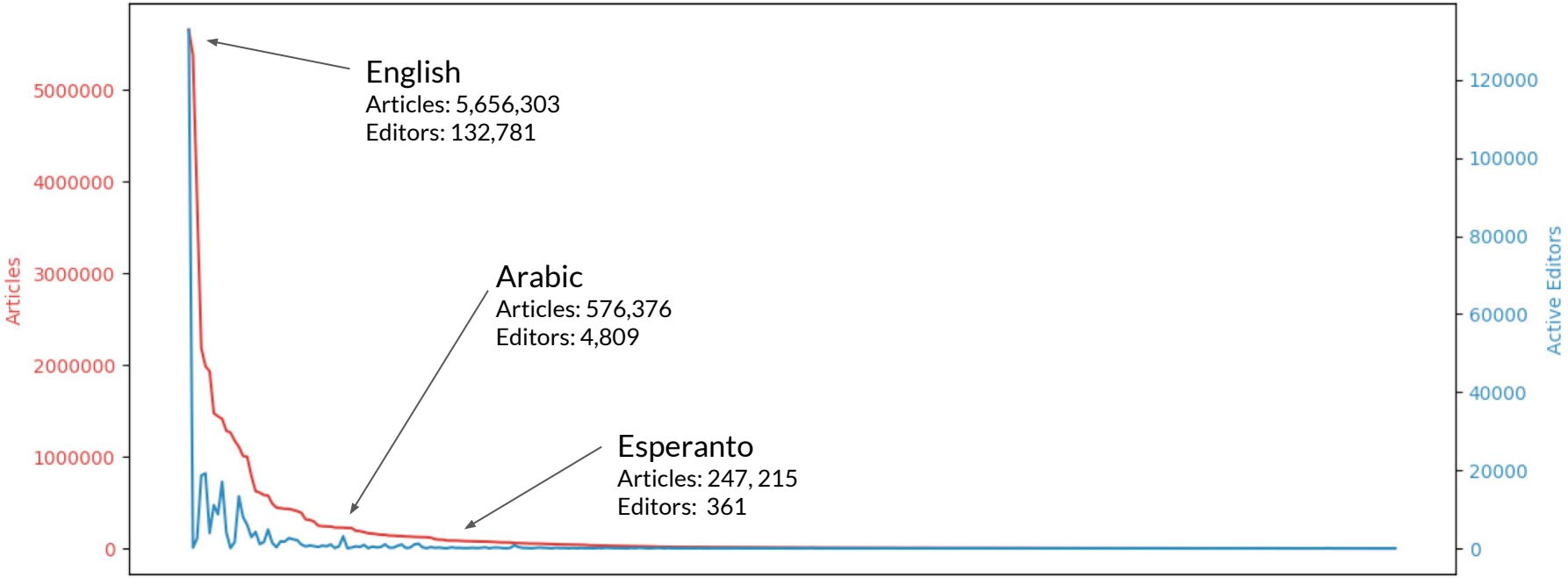
Lucie-Aimée Kaffee*, Hady Elsahar*, Pavlos Vougiouklis*, Christophe Gravier, Frédérique Laforest, Jonathon Hare, and Elena Simperl

* the authors contributed equally to this work

WDAqua

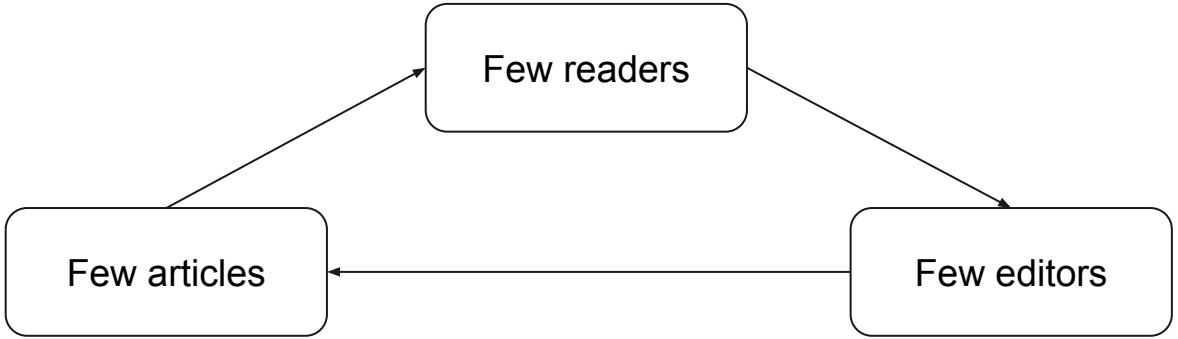
UNIVERSITY OF Southampton

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE



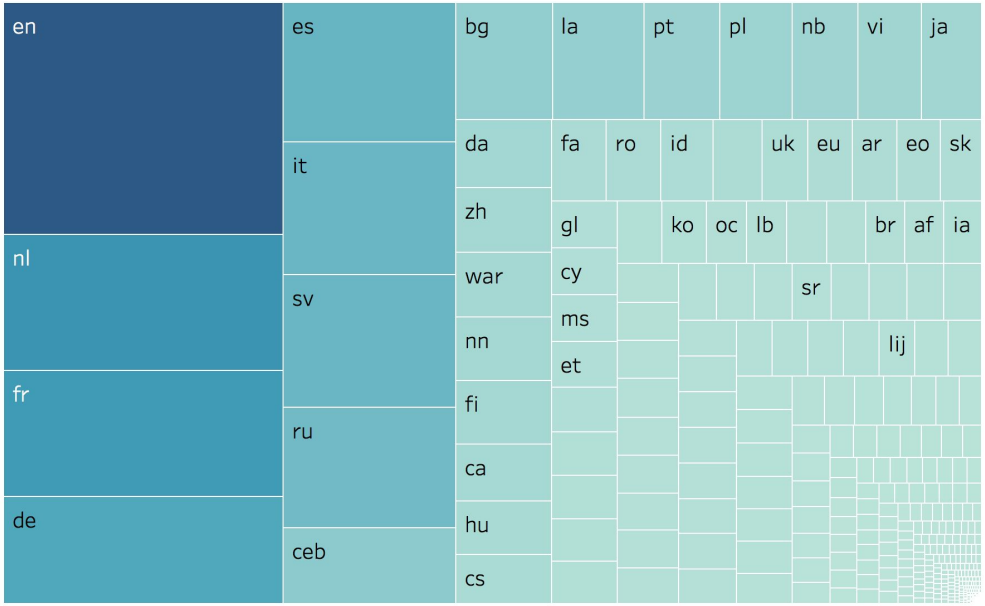
Wikipedia is available in 285 languages, but the content is unevenly distributed

Vicious cycle of lack of information

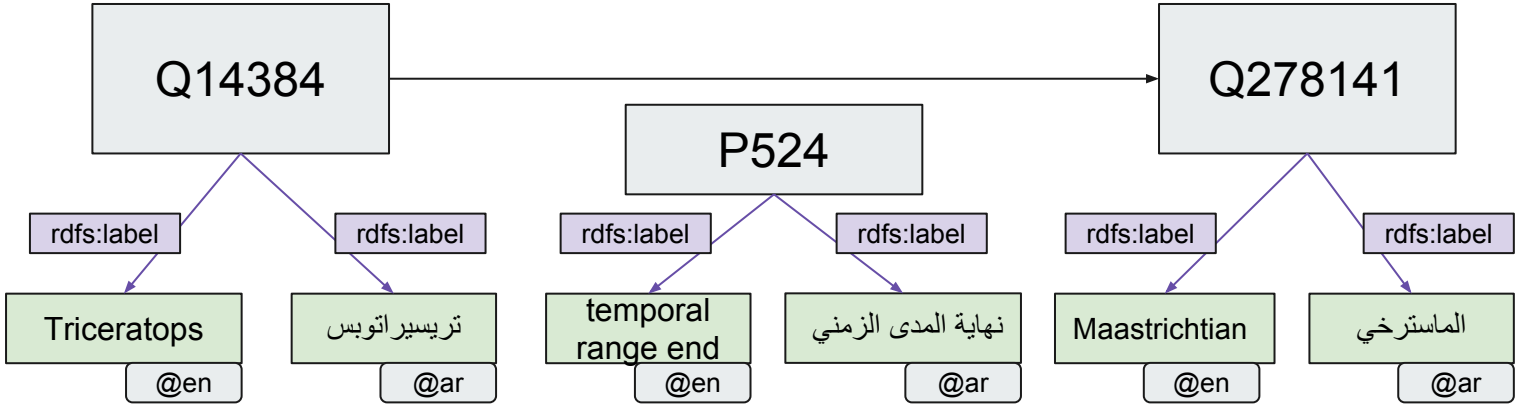


Wikidata (a source of multilingual structured information)

- ➔ Knowledge base maintained and edited by a community of users
- ➔ 48,775,926 items
- ➔ Each entity can have labels in >400 languages
- ➔ Variety of languages well covered (Kaffee et al. 2017)



Multilinguality in Wikidata



WIKIDATA

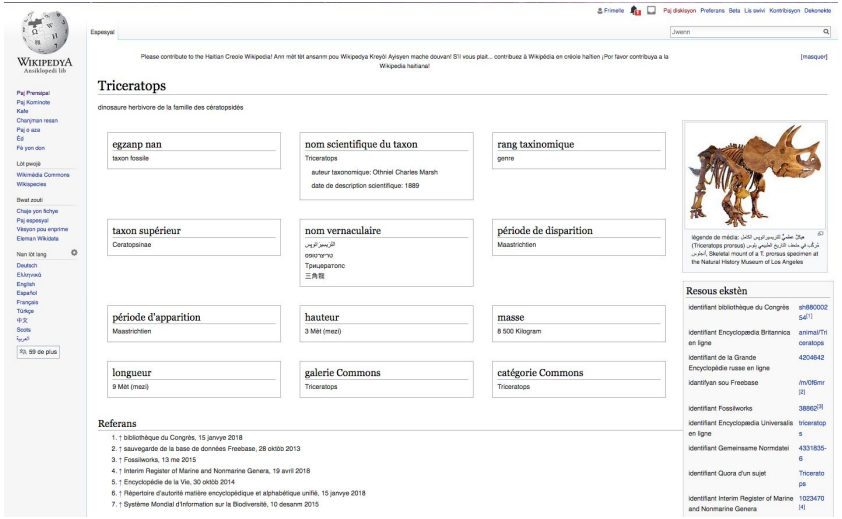
WIKIPEDIA
The Free Encyclopedia

ArticlePlaceholder

displays Wikidata triples on Wikipedia in tabular way

Dynamically generated when data changes, not stub articles

Currently deployed on 14 under-resourced Wikipedias (e.g. Gujarati, Haitian Creole, Urdu)



celtic knot

decorative knot

بیرونی وسائل

فری بیس آئی ڈی /m/ofk_3[1]

کومنز نگارخانہ

Knots in traditional art

ذیلی درجہ

Islamic interlace patterns
decorative knot

کومنز زمرہ

Celtic knots

3D model

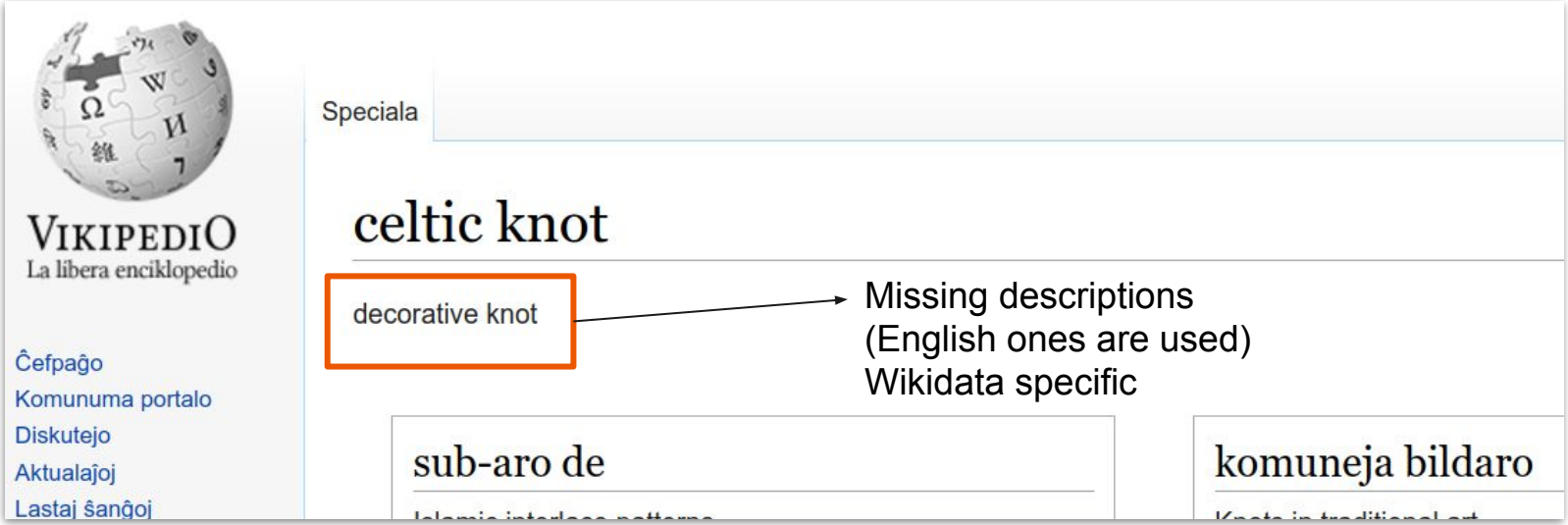
حوالہ

1. ↑ فری بیس ڈیٹا ڈمپس، 28 اکتوبر 2013

مضمون تحریر کریں

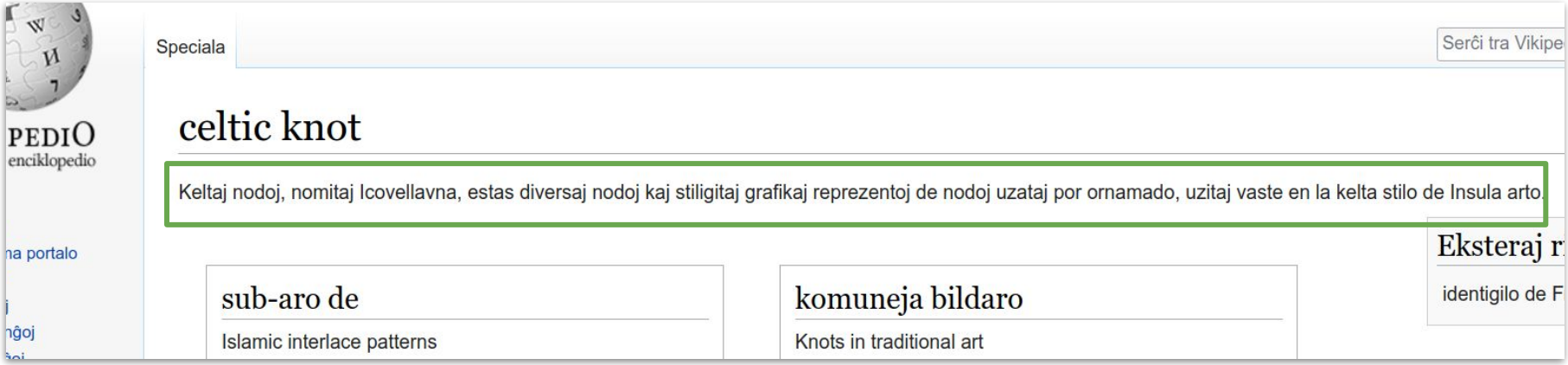
https://ur.wikipedia.org/wiki/Special:AboutTopic/Q242009

ArticlePlaceholders **“Textual Extension”**



ArticlePlaceholders **“Textual Extension”**

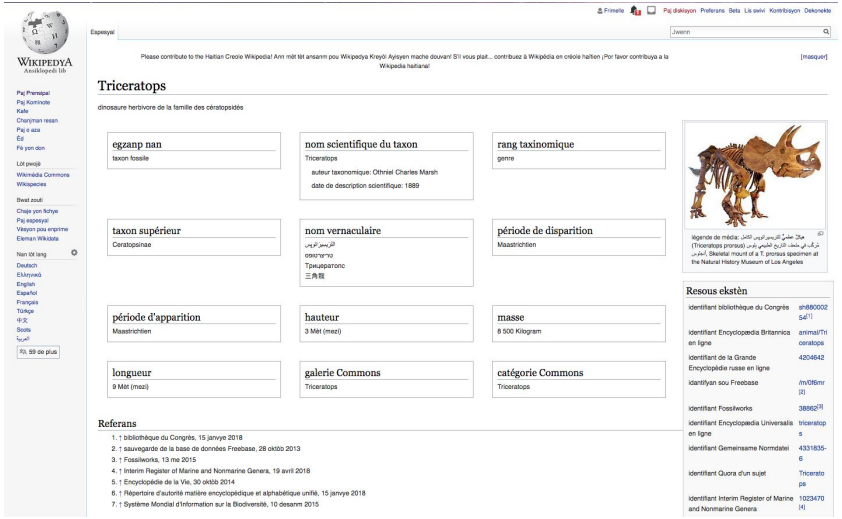
Enriching ArticlePlaceholder with textual summaries generated from Wikidata triples using **Neural Language Generation from Structured Knowledge bases**



Enriching ArticlePlaceholder with textual summaries generated from Wikidata triples using Neural Language Generation from Structured Knowledge bases

→ More pleasant to readers than tables
→ Can serve as a starting point to a wikipedia article

Working on under-resourced languages (testing on Arabic and Esperanto)

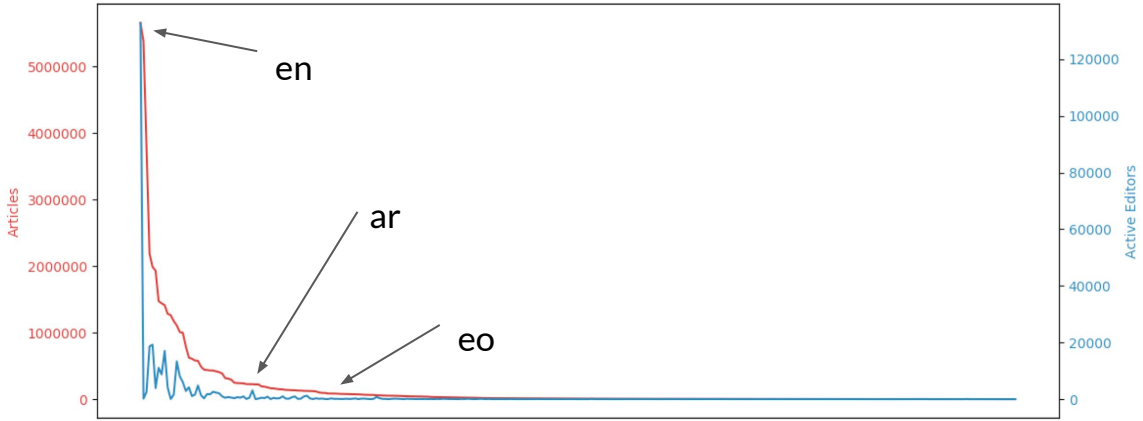


Esperanto

- **Esperanto** is an artificial language
- Easy to learn
- Engaged Wikipedia community
- A good starting point

Arabic

- **Arabic** is the 5th most spoken language in the world
- Content online in Arabic is sparse however



Reasonator

- displays Wikidata information
- manual created templates
- community effort
- can only display as much information as available in Wikidata

celtic knot (Q242009)

Knotenmuster

decorative knot

Other properties

sub-aro de	Islamic interlace patterns ornements évoquant des cordes sans extrémités et enchevêtrées
	decorative knot

Classification

Name	Description
Ento	io, kio ekzistas
Objekto	technical term in modern philosophy often used in contrast to the term subject
Concrete object	object with a physical referent
Fizika substanco	substance composed of quantum particle(s)/field(s), such as matter and/or radiation; that of which objects/systems are composed; physical stuff that can be considered concrete (not strictly abstract)
Fizika korpo	singular aggregation of substance(s) such as matter or radiation, with overall properties such as mass, position or momentum
Artefakto	physical object made or shaped by human hand
Maŝinelemento	elementary component of a machine
Rapidumskatolo	machine in a power transmission system for controlled application of the power;gearbox,uses gears/gear trains to provide speed,torque conversions from a rotating power source to another device;reduces the higher engine speed to the slower wheel speed
Pivoto	assembly of bodies connected to manage forces and movement
Nodo	method of fastening or securing linear material, such as rope, by tying or interweaving
Decorative knot	
Interlace	decorative element of bands or portions of other motifs looped, braided, and knotted in complex geometric patterns
Islamic interlace patterns	ornements évoquant des cordes sans extrémités et enchevêtrées
Celtic knot	decorative knot

Free images Google search

External sources

3D model	Celtic knot.stl
Freebase	/m/0fk_3

Wikimedia projects

Big Wikipedias	
cs	Pletenec (ornament)
de	Knotenmuster
en	Celtic knot
es	Nudo celta
nl	Keltische knoop
pl	Węzeł celtycki
pt	Nó celta
ru	Кельтский узел
Wikimedia Commons	
commons	Category:Celtic knots
Other Wikipedias	
eu	Zelta korapilo
gl	Nó celta
no	Keltisk knute

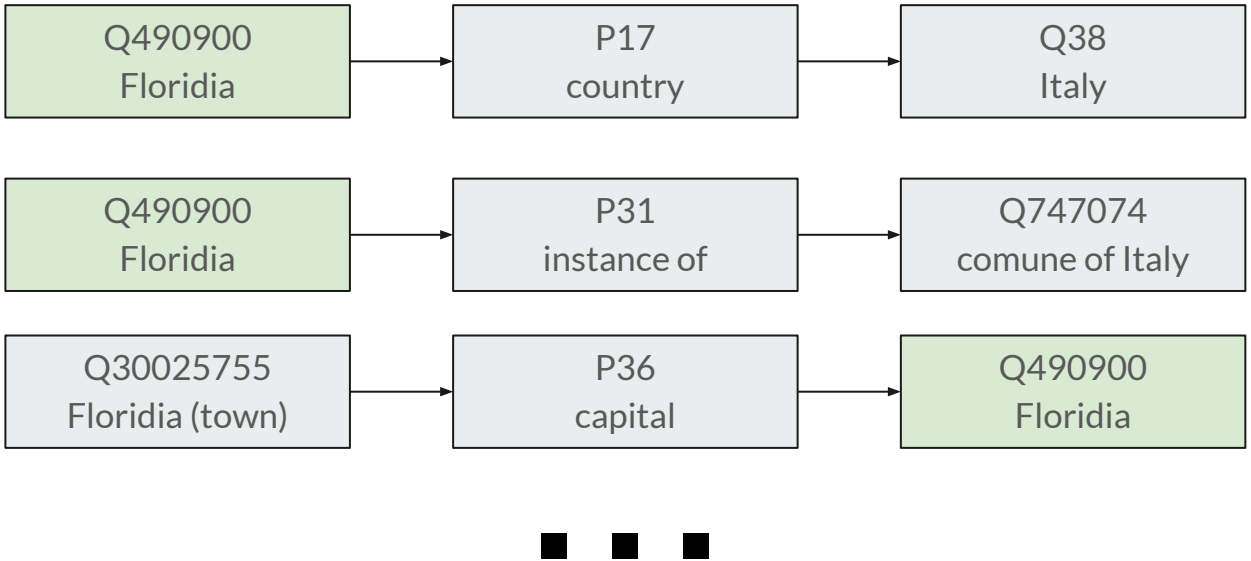
Concept cloud

Related media

Komuneja kategorio : Celtic knots
komuneja bildaro : Knots in traditional art

https://tools.wmflabs.org/reasonator/?q=Q242009&lang=eo

Sample Input (from ArticlePlaceholder)

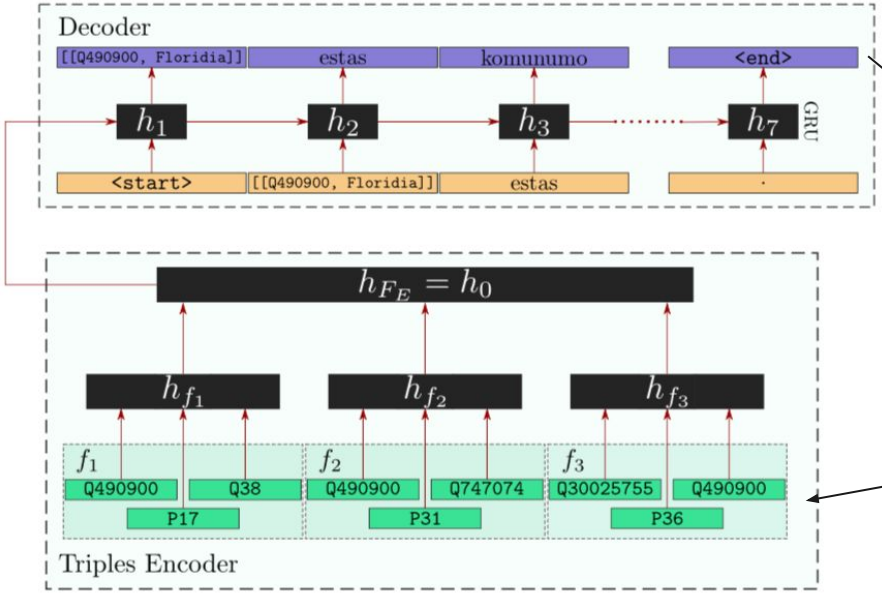


Neural Text Generation

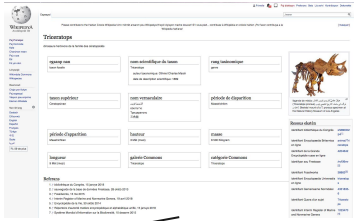
Feed-forward architecture encodes triples from the ArticlePlaceholder into a vector of fixed dimensionality

RNN-based decoder generates text summaries, one token at a time

Based on [Vougiouklis et al. (2017)]



ArticlePlaceholder



Dataset Preparation

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

WIKIPEDIA
The Free Encyclopedia

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Dataset Preparation

Nigragorĝa pigogarolo (Q1586267)

instance of	taksono
taksonomia nomo	Calocitta
supera taksono	Pigogarolo
original combination	Pica colliei

La **Nigragorĝa pigogarolo** (***Calocitta colliei***) estas rimarkinda longvosta pigogarolo de la familio de Korvedoj kaj ordo de Paseroformaj kiu loĝas en nordokcidenta Meksiko.

The black-throated magpie-jay (Calocitta colliei) is a strikingly long-tailed magpie-jay of northwestern Mexico.

Wikimedia's global language fallback chain

Property Placeholder

- ➔ for underserved languages we experience more out of vocabulary words due to limited training data
- ➔ to overcome this problem, we introduce property placeholder

Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (P225)	Calocitta
supera taksono (P171)	Pigogarolo

Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (P225)	Calocitta
supera taksono (P171)	Pigogarolo

La Nigragorĝa pigogarolo (**Calocitta** colliei) estas rimarkinda longvosta **pigogarolo** de la familio…

Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (P225)	Calocitta

supera taksono (P171)	Pigogarolo

La Nigragorĝa pigogarolo (**Calocitta** colliei) estas rimarkinda longvosta **pigogarolo** de la familio…

Property Placeholder

Nigragorĝa pigogarolo (Q1586267)

taksonomia nomo (**P225**)	Calocitta
supera taksono (**P171**)	Pigogarolo

La Nigragorĝa pigogarolo (**Calocitta** colliei) estas rimarkinda longvosta **pigogarolo** de la familio…

La Nigragorĝa pigogarolo (**[[P225]]** colliei) estas rimarkinda longvosta **[[P171]]** de la familio…

Evaluation

- **Automatic Evaluation**

- **Community Study**
  - Readers Evaluation
  - Editors Evaluation

Generated Examples

Q106693 Group 14 (chemical series):

مجموعة الكربون هي العناصر الموجودة الموجودة في **الجدول الدوري للعناصر**

Karbongrupo estas **elemento en grupo** 0 **de la perioda tabelo** laŭ la IUPAC-sistemo .

The carbon group is a periodic table group consisting of carbon, silico-n, germanium, tin, lead, and flerovium.

Q16885 *Thelxinoe* (natural satellite):

ثيليكسيون هو قمر طبيعي غير نظامي يتحرك بحركة تراجعية تابع لكوكب المشتري .

Telksino estas **neregula satelito** de **Jupitero** , kiu havas retrogradan orbiton .

Thelxinoe (/θɛlkˈsɪnoʊˌiː/ thelk-SIN-o-ee; Greek: Θελξινόη), also known as Jupiter XLII, is a natural satellite of Jupiter.

Automatic Evaluation

- Baselines
  - Machine Translation → already used in Wikipedia
  - Kneser-Ney → 5-gram language model
  - Template Retrieval → widely used for text generation

- Automatic Evaluation Metrics: BLEU 1 - 4, METEOR, ROUGE

Model	BLEU 4	ROUGE-L	METEOR
MT	9.11	30.51	30.10
KN	0.61	17.09	29.02
KN_ext	13.42	28.52	30.43
TR	25.98	43.58	33.33
TR_ext	32.51	50.57	34.25
Ours	39.20	64.64	45.99
+Placeholders	**39.51**	**64.69**	**46.17**

Arabic

Model	BLEU 4	ROUGE-L	METEOR
MT	9.11	30.51	30.1
KN	2.79	36.90	30.74
KN_ext	8.79	44.77	33.71
TR	24.30	45.92	20.46
TR_ext	32.41	57.62	31.04
Ours	34.85	**67.02**	**41.13**
+Placeholders	**34.95**	66.61	40.74

Esperanto

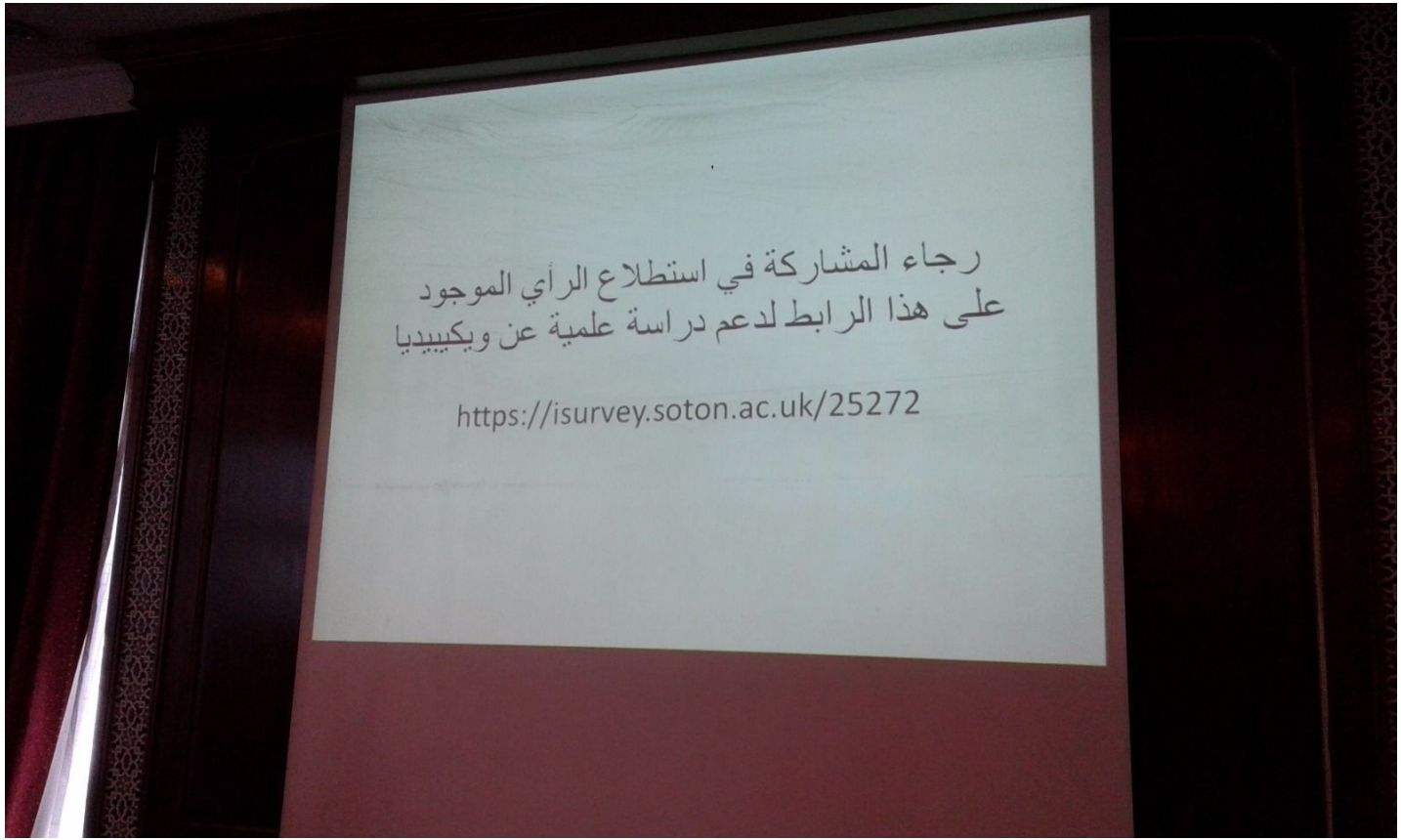
Results of the automatic evaluation: Our network outperforms all baselines in Automatic Evaluation

Community Study

- ➔ Two 15 days online surveys, aimed at readers and editors in Esperanto and Arabic
- ➔ Aiming to test our work with the actual Wikipedia community, outreach on Wikipedia platforms
- ➔ Two groups: **readers** and **editors**

Recruitment

- ➔ **Readers:** Social media (Reddit Esperanto, Twitter, Facebook)
- ➔ **Editors:** Social media, mailing lists and Wikipedia community pages



Link to our survey featured in the WikiArabia opening remarks

Reader evaluation

- Fluency: Is the text understandable and grammatically correct?
  - Scores from 0 to 6
- Appropriateness: Does the summary ‘feel’ like a Wikipedia article?
- Three different sources:
  - Generated sentences
  - Wikipedia
  - News

Participation

		Participants	Sentences	Participants Sentences > 50 %	Avg Sentence/ Participant	Total Annotations
Arabic	Fluency Approp.	27 27	60 60	5 5	15.03 14.78	406 399
Esperanto	Fluency Approp.	27 27	60 60	3 3	8.7 8.63	235 233

Instructions

من فضلك قم بتقييم جودة النص علي مقياس من صفر 0 إلى ستة 6

Generated Summary

خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

قم بتقييم جودة هذا النص:

- 0
- 1
- 2
- 3
- 4
- 5
- 6

Scores

Instructions

قيم إذا كنت تعتقد أن هذا النص من الممكن ان يكون مقتبس من ا لويكيبيديا العربية ام لا.

لا تعتمد على أي مصادر خارجية لمعرفة الإجابة (مثل محرك بحث جوجل أو ويكيبيديا)

Generated Summary

خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

هل تعتقد أن الجملة السابقة بإمكانها أن تستخدم كأول جملة في مقال من مقالات ويكيبيديا العربية ؟

		Fluency		Appropriateness
		Mean	SD	Part of Wikipedia
Arabic	Ours	4.7	1.2	77%
	Wikipedia	4.6	0.9	74%
	News	5.3	0.4	35%
Esper.	Ours	4.5	1.5	69%
	Wikipedia	4.9	1.2	84%
	News	4.2	1.2	52%

Results of the reader study: We generate sentences of comparable fluency, that “feel” like Wikipedia sentences

Editor evaluation

- ➔ Editors were asked to edit the article starting from our summary and the corresponding triples (2-3 sentences)
- ➔ How much of the text was reused?

How much of the text was reused by editiors?

Greedy String-Tiling (GST) (0 → 1):

- Mainly used for Plagiarism detection
- Detects whole block moves unlike Levenshtein distance.
- minimum match length (**mml**) factor to ignore copying of small subsequences

$$gstscore(S, D) = \frac{\sum_{t_i \in T} |t_i|}{|S|}$$

Length of the Longest common tiles (> mml) between source and edited text

Length of the source text

- **Dividing results into:**
  - WD (wholly derived)
  - PD (partially derived)
  - ND (non-derived)

Participation

		Participants	Sentences	Participants Sentences > 50 %	Avg Sentence/ Participant	Total Annotations
Arabic	Editors	7	30	2	4	33
Esperanto	Editors	8	30	2	4.75	38

Instructions

من فضلك قم بكتابة فقرة من الممكن أن تستخدم كأول فقرة في صفحة الويكيبيديا الخاصة بهذا الموضوع باستخدام المعلومات المعطاة لك فقط.

Generated Summary

خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة (كلمة ناقصة) ، ويكون على شكل بلورات بيضاء.

كلوريد الزرنيخ الخماسي -- حالة خاصة من -- مركب كيميائي
كلوريد الزرنيخ الخماسي -- الصيغة الكيميائية -- AtClu2085
كلوريد الزرنيخ الخماسي -- يتكون من -- زرنيخ
كلوريد الزرنيخ الخماسي -- يتكون من -- كلور
كلوريد الزرنيخ الخماسي -- مواصفات الإدخال النصي المبسط للجزيئات -- [At+5].[Cl-].[Cl-].[Cl-].[Cl-].[Cl-]
كلوريد الزرنيخ الخماسي -- تصنيف كومنز -- Artenic

Wikidata triples

Editing field

	Category	Examples	%
Arabic	WD	.خماسي كلوريد الزرنيخ مركب كيميائي له الصيغة(كلمة ناقصة)، ويكون على شكل بلورات بيضاء خماسي كلوريد الزرنيخ هو مُركب كميائي له الصيغة(AtClu2085)، ويكون على شكل بلورات بيضاء.	45.45%
	PD	بيتش باتوم أوهايو (بالإنجليزية (كلمة ناقصة) Ohio)هي منطقة سكنية تقع في الولايات المتحدة في(كلمة ناقصة.) بيتش باتوم (بالإنجليزية:Beach Batom)هي قرية تقع في الولايات المتحدة الامريكية في بروك كاونتي.	33.33%
	ND	دير علا هي بلدة تقع في جنوب غرب إيران. دير علا، أو بيثر، هي قرية أردنية	21.21%
Esperanto	WD	Zederik estas komunumo en la nederlanda provinco Zuid-Holland. Zederik estas komunumo en la nederlanda provinco Zuid-Hooland kaj estas ĉirkaŭata de la municipoj Lopik kaj Zederik.	78.98%
	PD	Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando, kiu havis (manka nombro) loĝantojn en (jaro). Nova Pádua estas municipo en la brazila subŝtato Suda Rio-Grando.	15.79%
	ND	Ibiúna estas municipo de la brazila subŝtato San-Paŭlio, kiu taksis (manka nombro) enloĝantojn en (jaro). Ibiúna estas brazila [[municipo]] kiu troviĝas en la administra unuo [[San-Paŭlo]].	5.26%

Results of the editor study: We generate sentences that are highly reused by editors

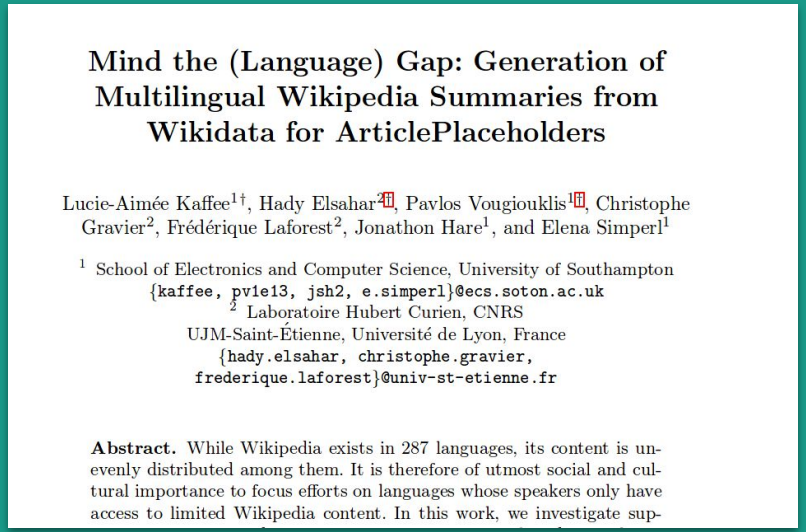
Conclusions

- ➔ Neural NLG approaches can work for underresourced languages with different properties (Arabic and Esperanto)
- ➔ Generated summaries are useful for article creation
- ➔ **Wikipedia's ArticlePlaceholder** is a good use case for NLG tasks.
- ➔ Engaging the community of readers / editors is the way to go when doing NLG for Wikipedia rather than automatic evaluation.

Questions

- https://tinyurl.com/y7opjsnl
  Paper published at **ESWC 2018**

- http://aclweb.org/anthology/N18-2101
  Neural Network published at **NAACL 2018**



Mind the (Language) Gap: Generation of Multilingual Wikipedia Summaries from Wikidata for ArticlePlaceholders

Lucie-Aimée Kaffee[1†], Hady Elsahar[2], Pavlos Vougiouklis[1], Christophe Gravier[2], Frédérique Laforest[2], Jonathon Hare[1], and Elena Simperl[1]

[1] School of Electronics and Computer Science, University of Southampton
{kaffee, pv1e13, jsh2, e.simperl}@ecs.soton.ac.uk
[2] Laboratoire Hubert Curien, CNRS
UJM-Saint-Étienne, Université de Lyon, France
{hady.elsahar, christophe.gravier, frederique.laforest}@univ-st-etienne.fr

Abstract. While Wikipedia exists in 287 languages, its content is unevenly distributed among them. It is therefore of utmost social and cultural importance to focus efforts on languages whose speakers only have access to limited Wikipedia content. In this work, we investigate sup-

Lucie-Aimée Kaffee
kaffee@soton.ac.uk

Hady Elsahar
hadyelsahar@gmail.com

Pavlos Vougiouklis
pv1e13@ecs.soton.ac.uk