Centre for
Data Analytics

Insight

IRIS:
English-Irish Translation Sytem

Mihael Arcan, UNLP, Insight@NUI Galway
Supervised by Dr. Paul Buitelaar

DCU NUI Galway Oé Gaillimh UCC UCD DUBLIN sfi

About UNLP @ Insight Galway and myself

IRIS: English-Irish Translation System

http://server1.nlp.insight-centre.org/iris/

Configuration of the Translation System

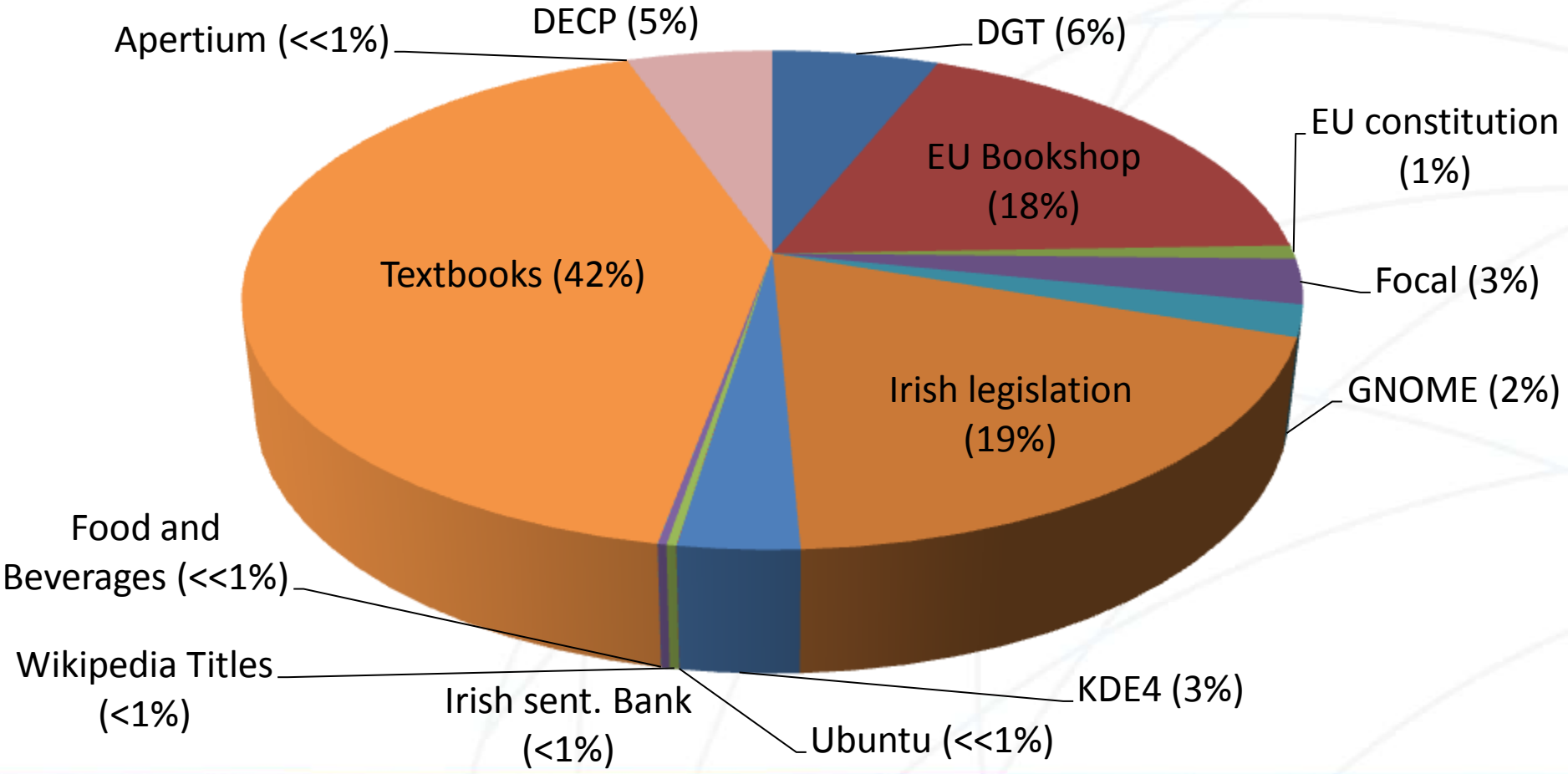
- IRIS System Interface
  - English-Irish bilingual interface
  - Allows users to upload new monolingual or parallel data used for training

- Translation System
  - Moses (default settings), giza++ word alignments, KenLM
  - source/target = tokenized, source = lowercased, target = detokenized
  - phrase table filtered on p(e|f) > 0.0001, max 5-grams, OnDisk binarization

- Web Service API
  - Input = sentence
  - Output = json object, with n-best translations

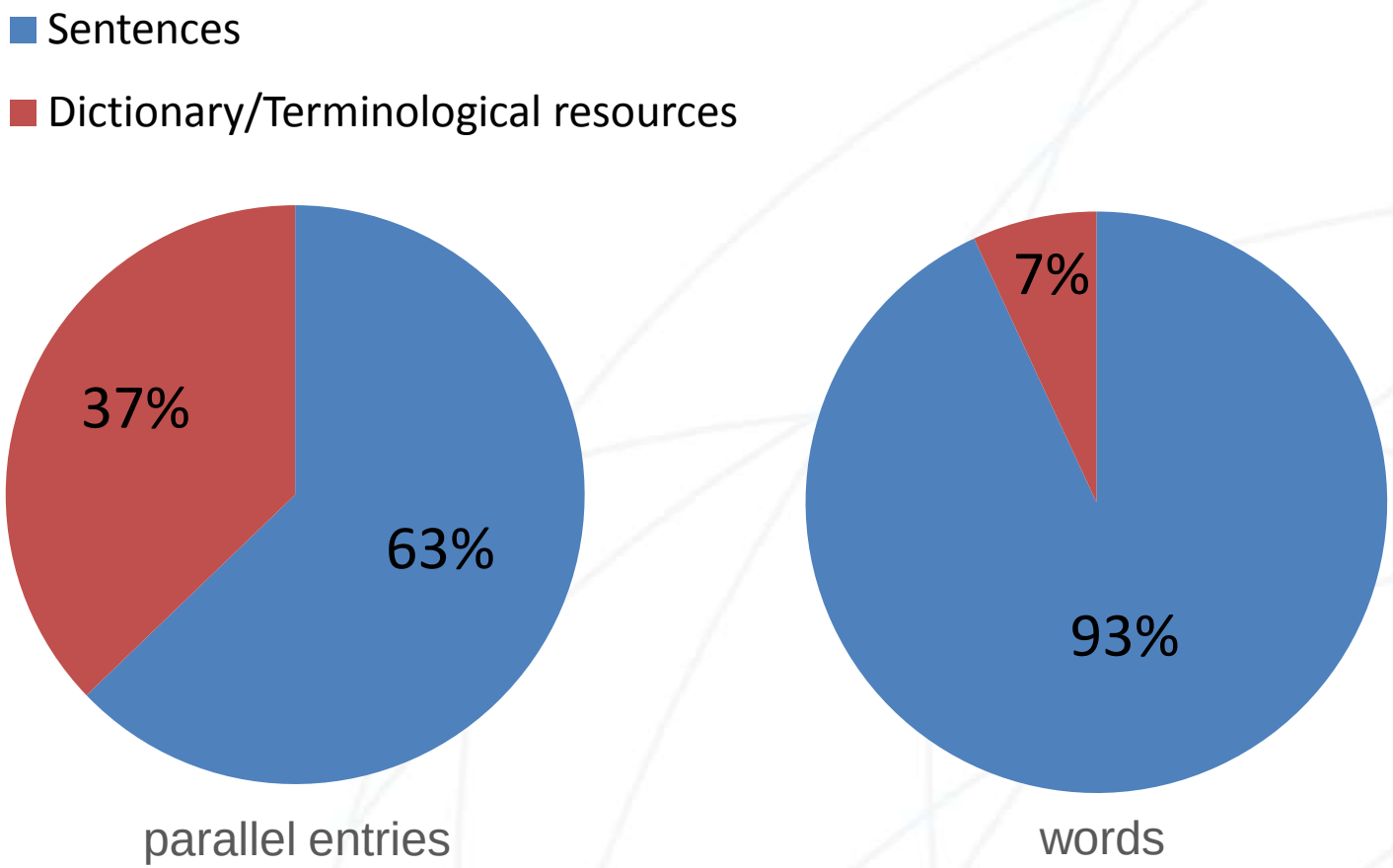
Parallel Corpora

~14/15M source/target words
~ 1M parallel entries



Data Structure

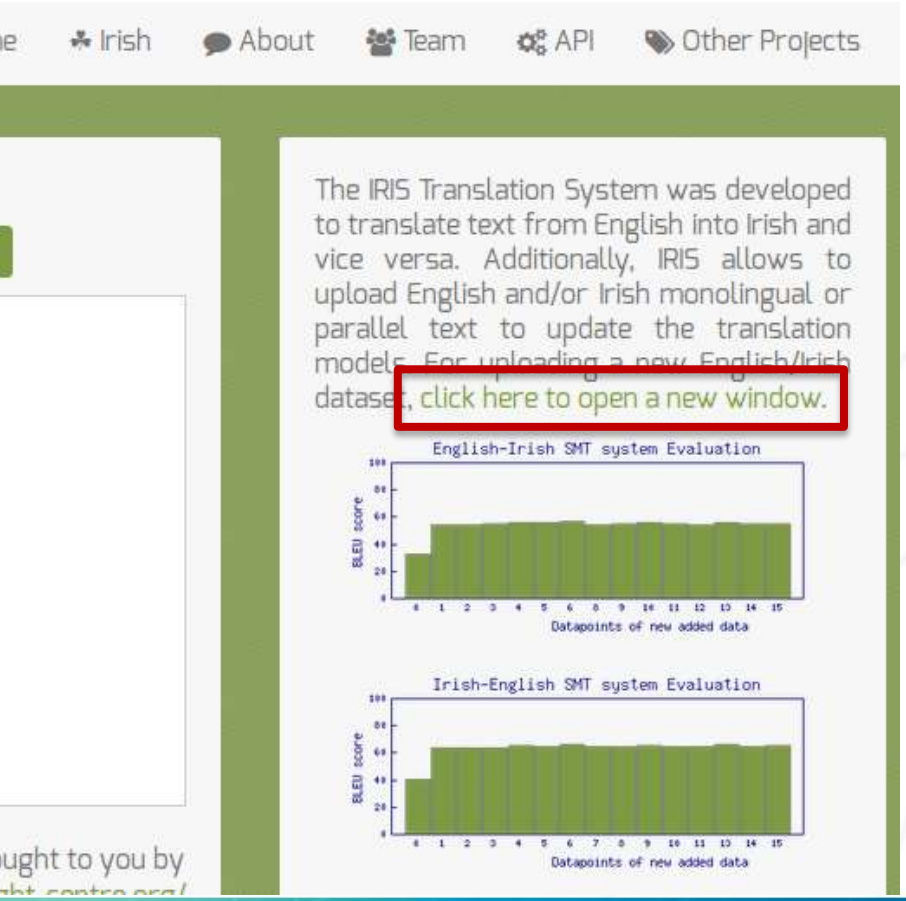
~14/15M source/target words
~ 1M parallel entries

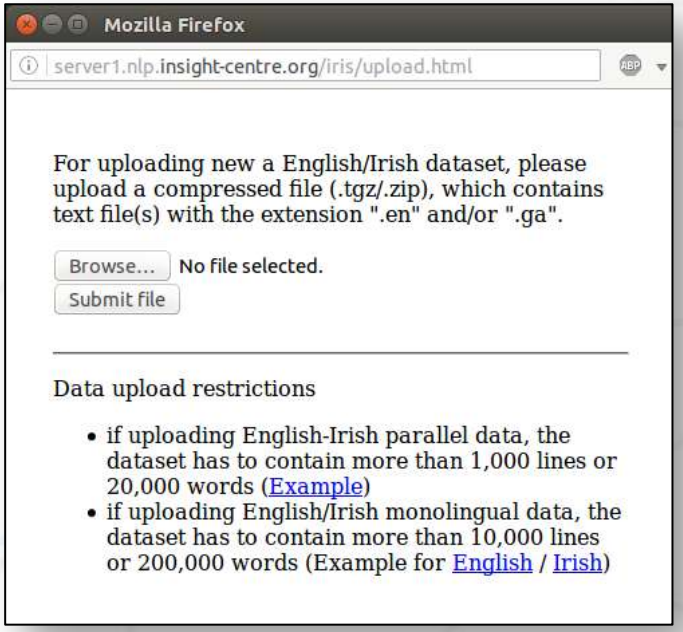


Language Models

- English:
  - News-2007 (3M sentences / 90M words)
    - + 14M words from parallel data, target side
- Irish
  - Wikipedia in Irish (250K sentences, 4M words)
    - + 15M words from parallel data, target side

Adding new data to IRIS





- Data / Language detection (parallel / monolingual / English / Irish)
- Re-training (PT or LM)
- Evaluation
- Eanbling models for the IRIS Interface

IRIS Evaluation

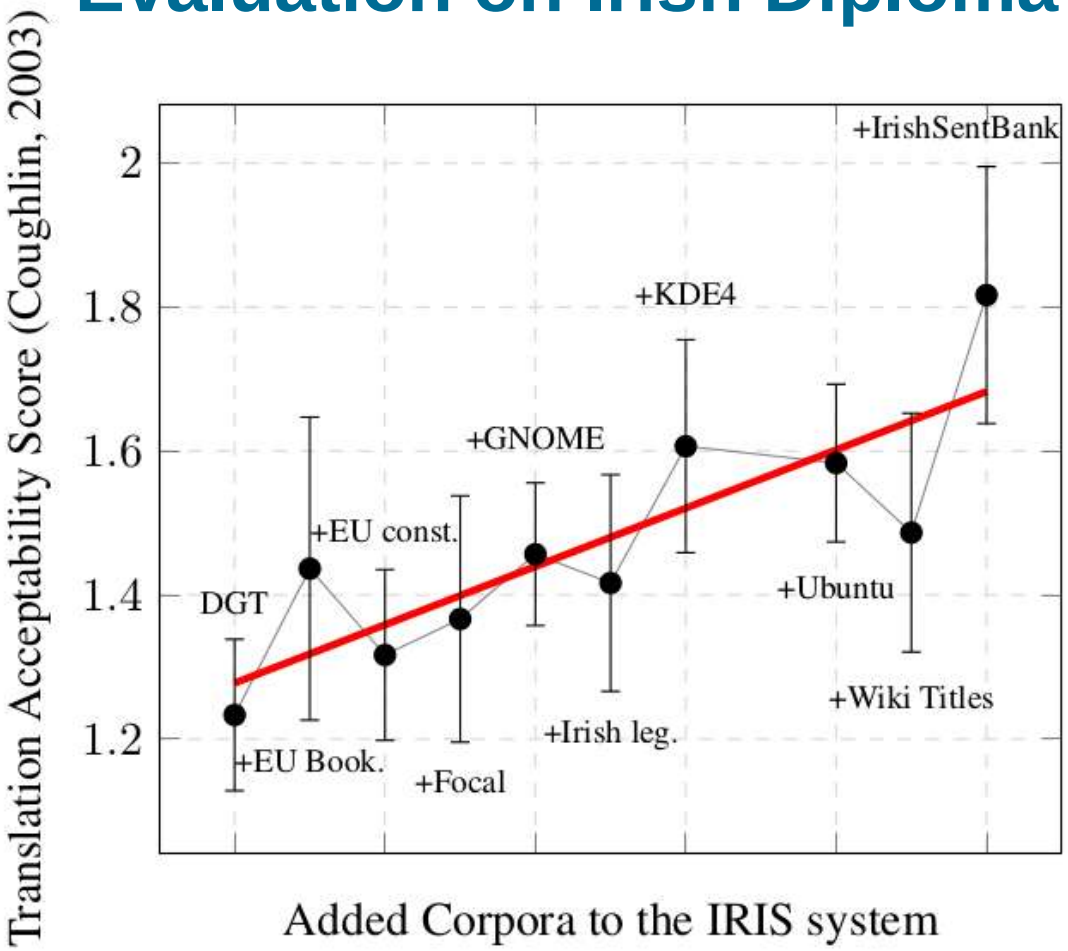
testset link

		English→Irish			Irish→English		
#	Corpus	BLEU	METEOR	chrF	BLEU	METEOR	chrF
0	DGT*	32.39	28.45	56.75	40.50	34.22	56.43
1	+EU Bookshop*	54.54	39.03	67.82	64.05	44.15	69.41
2	+EU constitution*	53.97	38.63	67.21	63.73	44.27	69.89
3	+Focal	54.82	39.30	68.59	64.04	44.65	70.53
4	+GNOME*	55.62	40.11	68.76	65.43	45.24	70.50
5	+Irish legislation	55.77	40.01	68.62	65.02	45.38	70.81
6	+KDE4*	56.62	40.73	69.36	66.28	46.24	71.23
7	+News-2007 (mono. English)	/	/	/	64.45	44.89	69.88
8	+Ubuntu	54.67	39.60	68.06	64.72	45.23	70.24
9	+Wikipedia Titles	55.44	39.99	68.10	65.41	45.52	70.55
10	+Irish sent. bank	55.76	40.23	68.35	64.87	45.72	70.92
11	+Food and Beverages	54.83	39.57	67.41	65.02	45.21	70.05
12	+Wikipedia (mono. Irish)	54.47	39.34	66.88	/	/	/
13	+Textbooks	55.84	40.28	68.61	66.18	46.12	71.21
14	+Apertium	54.85	39.66	67.75	64.68	45.06	70.11
	Google Translate	40.07	33.23	65.93	46.77	39.20	68.83

Evaluation on Irish Diploma Sentences

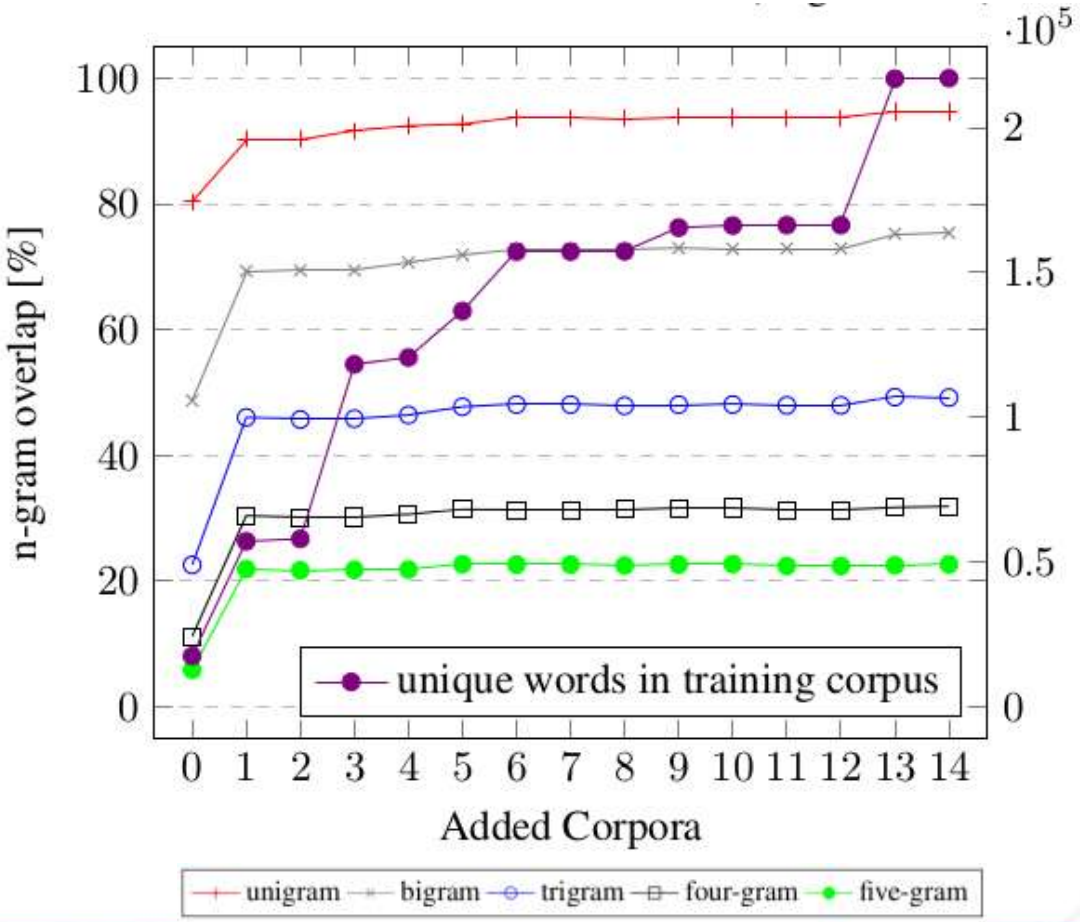
English	Irish
she should be grateful to you for returning her phone which she lost during the music festival .	ba cheart go mbeadh sí buíoch díot as an nguthán a thabhairt ar ais di a chaill sí le linn na féile ceoil
the subcommittee will submit its findings to the county council and the draft report will be published before the end of the tax year .	cuirfidh an fochoiste a thorthaí faoi bhráid na comhairle contae agus foilseofar an dréacht-tuarascáil roimh dheireadh na bliana cánach .
the two red cars , which were stolen in the middle of november , were found on the edge of cork city .	fuarthas an dá charr dhearga , a goideadh i lár mhí na samhna , ar imeall chathair chorcaí .
people like brian and bríd understand the social needs of the local people .	tuigeann leithéidí bhriain agus bhríde riachtanais shóisialta mhuintir na háite .
the annual report of the housing committee stated that there was a significant increase in the number of homeless people in many poor areas .	dúradh / maíodh i dtuarascáil bhliantúil an choiste tithíochta go raibh ardú suntasach sa líon daoine gan dídean in go leor ceantar bocht .
applicants with the necessary experience will be able to apply for the three positions advertised last week in the local paper .	beidh iarratasóirí a bhfuil an taithí chuí acu in ann cur isteach ar na trí phost a fógraíodh an tseachtain seo caite sa nuachtán áitiúil .
a man with a long beard and blue eyes was standing in front of the gates of the town hall .	bhí fear a raibh féasóg fhada air agus súile gorma aige ina sheasamh os comhair gheataí halla an bhaile .
thirteen soldiers were killed during the last days of the war and two civilians were badly injured .	maraíodh trí shaighdiúir déag le linn laethanta deiridh / deireanacha an chogaidh agus gortaíodh beirt sibhialtach go dona .
i put the sound files on the four blue cds and gave them to the office manager ' s personal secretary .	chuir mé na comhaid fuaime ar na ceithre dhlúthdhiosca ghorma agus thug mé iad do rúnaí pearsanta bhainisteoir na hoifige .
she lived for sixteen months opposite the shopping centre , but she now lives on the edge of the city , beside the principal ' s house .	bhí sí ina cónaí ar feadh sé mhí dhéag os comhair an ionaid siopadóireachta , ach tá cónaí uirthi anois ar imeall na cathrach , in aice le teach an phríomhoide

Evaluation on Irish Diploma Sentences



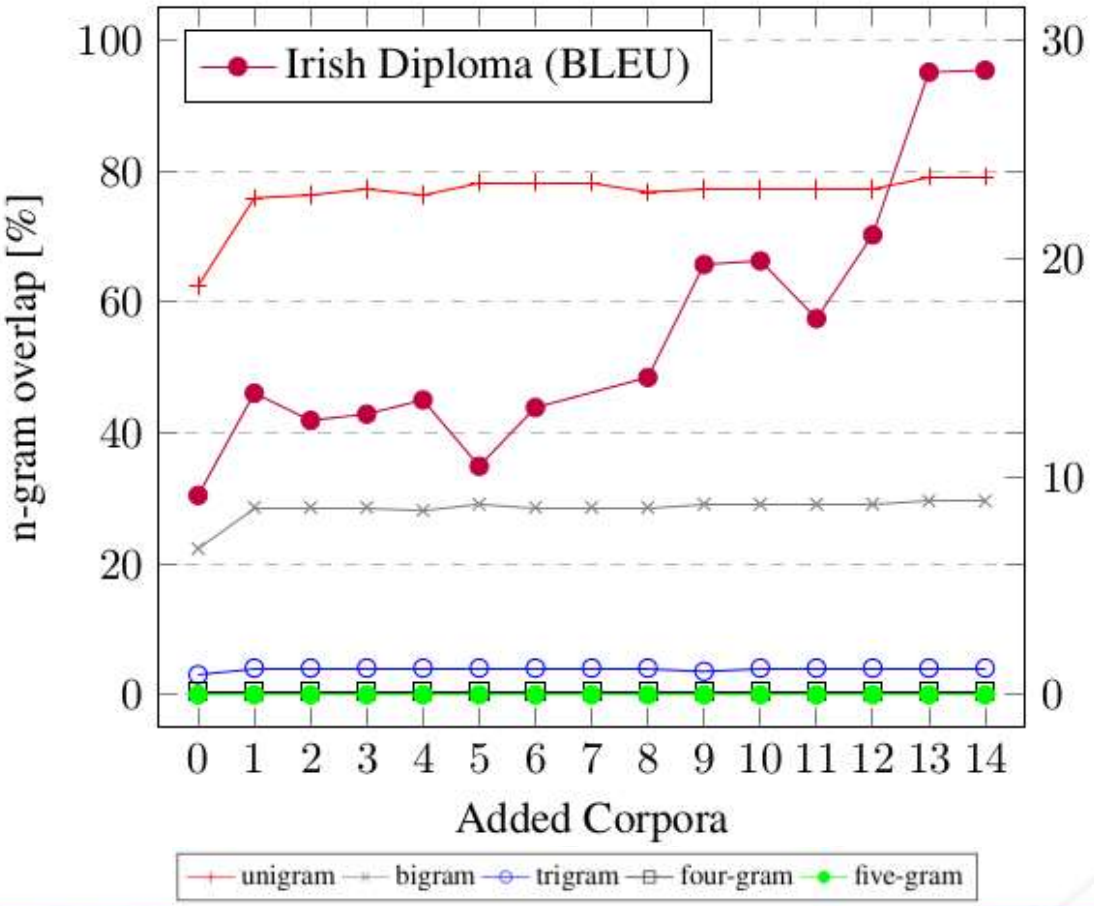
Corpus	Evaluation data set [BLEU] English→Irish	Irish→English
DGT	9.14	8.74
+EU Bookshop	13.83	16.12
+EU constitution	12.58	20.98
+Focal	12.87	16.69
+GNOME	13.52	17.52
+Irish legislation	10.48	19.35
+KDE4	13.17	20.48
+News-2007 (English)	/	25.73
+Ubuntu	14.55	25.60
+Wikipedia Titles	19.73	28.11
+Irish sent. bank	19.90	24.64
+Food and Beverages	17.25	25.98
+Wikipedia (Irish)	21.09	/
+Textbooks	28.55	36.88
+Apertium	28.64	31.46

n-gram overlap between PT and test set



#	Corpus	BLEU
0	DGT*	32.39
1	+EU Bookshop*	54.54
2	+EU constitution*	53.97
3	+Focal	54.82
4	+GNOME*	55.62
5	+Irish legislation	55.77
6	+KDE4*	56.62
7	+News-2007 (mono. English)	/
8	+Ubuntu	54.67
9	+Wikipedia Titles	55.44
10	+Irish sent. bank	55.76
11	+Food and Beverages	54.83
12	+Wikipedia (mono. Irish)	54.47
13	+Textbooks	55.84
14	+Apertium	54.85

N-gram overlap between PT and Diploma test set



#	Corpus
0	DGT*
1	+EU Bookshop*
2	+EU constitution*
3	+Focal
4	+GNOME*
5	+Irish legislation
6	+KDE4*
7	+News-2007 (mono. English)
8	+Ubuntu
9	+Wikipedia Titles
10	+Irish sent. bank
11	+Food and Beverages
12	+Wikipedia (mono. Irish)
13	+Textbooks
14	+Apertium

Web Service API

JSON INPUT


```
{
   "text":"Ireland will use no fossil fuels by the end of the century, according to a new Government energy policy paper.",
   "translation_direction":"en_ga",
   "nbest":10
}
```


JSON OUTPUT


```
{
   "source":"Ireland will use no fossil fuels by the end of the century, according to a new Government energy policy paper.",
   "time":" 2 wallclock secs ( 0.02 usr  0.00 sys +  1.77 cusr  0.13 csys =  1.92 CPU)",
   "text":"Ireland will use no fossil fuels by the end of the century, according to a new Government energy policy paper.",
   "translation_direction":"en_ga",
   "n_best":{
      "Ní úsáidfidh sé gur breoslaí iontaise in Éirinn faoi dheireadh an chéid nua , de réir bheartas fuinnimh an Rialtais . ":"-16.0954"
      "Ní úsáidfidh breoslaí iontaise in Éirinn faoi dheireadh na haoise nua , de réir bheartas fuinnimh an Rialtais . ":"-16.2009",
      "gan aon úsáid a bhaint as breoslaí iontaise in Éirinn faoi dheireadh an chéid nua , de réir bheartas fuinnimh an Rialtais . ":"-16
      "gan aon úsáid a bhaint as breoslaí iontaise in Éirinn faoi dheireadh na haoise nua , de réir bheartas fuinnimh an Rialtais . ":"-1
      "Ní úsáidfidh breoslaí iontaise in Éirinn faoi dheireadh an chéid nua , de réir bheartas fuinnimh an Rialtais . ":"-16.0423",
      "Ní úsáidfear breoslaí iontaise in Éirinn faoi dheireadh na haoise nua , de réir bheartas fuinnimh an Rialtais . ":"-16.3136",
      "Úsáid breoslaí iontaise in Éirinn ar bith faoi dheireadh an chéid nua , de réir bheartas fuinnimh an Rialtais . ":"-16.2788",
      "Ní úsáidfear breoslaí iontaise in Éirinn faoi dheireadh an chéid nua , de réir bheartas fuinnimh an Rialtais . ":"-16.155",
      "Ní úsáidfidh sé gur breoslaí iontaise in Éirinn faoi dheireadh na haoise nua , de réir bheartas fuinnimh an Rialtais . ":"-16.254"
      "Níl aon úsáid a bhaint as breoslaí iontaise in Éirinn faoi dheireadh an chéid nua , de réir bheartas fuinnimh an Rialtais . ":"-16
   },
   "nbest":"10",
   "best_translation":"Ní úsáidfidh breoslaí iontaise in Éirinn faoi dheireadh an chéid nua, de réir bheartas fuinnimh an Rialtais."
}
```


http://server1.nlp.insight-centre.org/iris/rest_service.html

Ongoing Work

Current (version 2)	New (version 3)	
MERT 1x	MERT 5x	✓
standard training data cleaning /handling (clean-corpus-n.perl)	more carefull training data cleaning/handling	✓
evaluation set 1 (60% sentences / 40 terminology)	evaluation set 1 (100 % sentences, larger)	
evaluation set 2 (75 % old / 25 % new data)	evaluation set 2 (100% new data)	✓
finding new data, ...		

IRIS: English-Irish Translation Sytem

Thank You

IRIS Demo: http://server1.nlp.insight-centre.org/iris/

more info: Arcan et al., 2016 – LREC

contact: mihael.arcan@insight-centre.org