Discriminating Between Similar Languages Using Large Web Corpora

Vít Suchomel

xsuchom2@fi.muni.cz

December 7, 2019
RASLAN 2019

Outline

Motivation: Corpus Quality

Query **hjælp** **85,070** (43.54 per million)

gcompris.net	projektet., jeg har brug for din	**hjælp**	for at kunne fortsætte min
altermedia.info	hundrede lig i en afrikansk udørk?	**Hjælp**	dog den forsvarsløse
balder.org	alle tidlige vidner om mord ved	**hjælp**	af klor eller en ildevarslende sort
virksommeord.uib.no	organisk fremskridt, ikke ved	**hjælp**	av revolution, ikke ved hjælp av
virksommeord.uib.no	ved hjælp av revolution, ikke ved	**hjælp**	av brudd paa alt det bestaaende,
copenhagencakes.com	til at tage springet. Med masser af	**hjælp**	og støtte fra min kæreste startede
borgerskolen.no	Venner, der kom mig til	**Hjælp**	, blandt hvilke kan nævnes :
blog.webdanmark.com	tracking på dit website, uden	**hjælp**	fra en programmør. Det er dermed et
helenejuuldesign.blogspot.no	eget design ... naturligvis med den	**hjælp**	fra mig, du ønsker. Kurset går igen
egmont.com	udsatte børn og unge, f.eks. som	**hjælp**	til at håndtere livskriser som sorg
blog.webdanmark.com	Adwords specialist… Har du brug for	**hjælp**	til at oprette din første AdWords
blog.webdanmark.com	til at klikke på disse bannere ved	**hjælp**	af blinkende elementer,
stinehoelgaard.blogspot.no	Og her er det så at jeg behøver jeres	**hjælp**	Jeg kan slet ikke vurderer, om en

Motivation: Corpus Quality

Danish in Norwegian Web Corpus 2015 vs. 2017

Corpus	**Size [tokens]**	**”hjelp”**		**”hjælp”**	
		frq	**frq/mil**	**frq**	**frq/mil**
noTenTen15	1.95 G	269,051	138	85,070	43.5
noTenTen17	3.11 G	562,420	181	519	0.167

Motivation: Text Sources

Questions of web corpora users concerning language variants:

- How much American Spanish is in your corpus?
- Can you give me just Australian English?

Motivation: Text Sources

Questions of web corpora users concerning language variants:

- How much American Spanish is in your corpus?
- Can you give me just Australian English?

TLD helps a lot: `.ar`, `.mx`, `.cu`,…; `.au`.

Motivation: Text Sources

Questions of web corpora users concerning language variants:

- How much American Spanish is in your corpus?
- Can you give me just Australian English?

TLD helps a lot: `.ar`, `.mx`, `.cu`,…; `.au`.

Yet ∼80 % of web pages in enTenTen18 come from generic TLDs (`.com` is the biggest with 53 %).

Web Text Processing Workflow

1. Detect character encoding
2. Extract text from HTML (boilerplate removal)
3. Check document similarity to the target language
   – character n-gram vectors
4. **Eliminate paragraphs containing too much foreign language**
5. **Discerning similar languages on both doc and paragraph level**
6. **Split the document paragraphs by language**

Web Text Processing Workflow – Separate Output Docs

Original	**Lang 1 output**	**Lang 2 output**	**Lang 3 output**
<doc>	<doc lg="L1">	<doc lg="L2">	<doc lg="L3">
L1 paragraph	L1 paragraph		
L1 paragraph	L1 paragraph		
L2 paragraph		L2 paragraph	
L1 paragraph	L1 paragraph		
L1 paragraph	L1 paragraph		
L3 paragraph			L3 paragraph
L3 paragraph			L3 paragraph
L3 paragraph			L3 paragraph
</doc>	</doc>	</doc>	</doc>

Related Work

Lui & Baldwin at ACL 2012 – `langid.py`

- Naive Bayes classifier,
- character 1-4-grams,
- 111 k doc corpus: government, software, newswire, Wikipedia, general web.

Related Work

Tiedemann & Ljubešić at COLING 2012:

- Naive Bayes classifier,
- words,
- blacklisted words,
- parallel/comparable corpora of Bosnian, Croatian and Serbian.

Simple Method, Large Web Corpora?

Web corpora frequency wordlist based method implemented in a Python script:

- Operates on documents and paragraphs,
- shows the contribution of each word to the decision,
- provides a measure of confidence,
- easy to support a new language

Frequency Wordlists From Big Web Corpora

Text from domains .uk and .us

en-UK	corpus freq	en-US	corpus freq
the	232,528,754	the	39,197,118
and	125,125,025	and	20,871,358
to	115,346,475	to	19,843,435
of	113,886,667	of	19,552,038
a	90,778,943	a	15,184,592
in	74,678,079	in	12,939,187
is	46,842,672	is	8,292,931
for	44,833,138	for	8,200,031
that	37,795,087	that	7,261,971
with	32,271,857	you	5,648,462
...		...	
zzzzzd	4	zzzzzzzzzzzs	1
total	3,928,096,030	total	701,279,329

Frequency Wordlists From Big Web Corpora

Text from domains .uk and .us

en-UK	corpus freq	logrf	en-US	corpus freq	logrf
the	232,528,754	7.77	the	39,197,118	7.75
and	125,125,025	7.50	and	20,871,358	7.47
to	115,346,475	7.47	to	19,843,435	7.45
of	113,886,667	7.46	of	19,552,038	7.45
a	90,778,943	7.36	a	15,184,592	7.34
in	74,678,079	7.28	in	12,939,187	7.27
is	46,842,672	7.08	is	8,292,931	7.07
for	44,833,138	7.06	for	8,200,031	7.07
that	37,795,087	6.98	that	7,261,971	7.02
with	32,271,857	6.91	you	5,648,462	6.91
...			...		
zzzzzd	4	0.01	zzzzzzzzzzzs	1	0.15
total	3,928,096,030		total	701,279,329	

Frequency Wordlists From Big Web Corpora

Text from domains .uk and .us and .cz

en-UK	corpus freq	logrf	en-US	corpus freq	logrf	cs-CZ
the	232,528,754	7.77	the	39,197,118	7.75	5.11
and	125,125,025	7.50	and	20,871,358	7.47	4.63
to	115,346,475	7.47	to	19,843,435	7.45	6.97
of	113,886,667	7.46	of	19,552,038	7.45	5.25
a	90,778,943	7.36	a	15,184,592	7.34	7.46
in	74,678,079	7.28	in	12,939,187	7.27	4.73
is	46,842,672	7.08	is	8,292,931	7.07	4.33
for	44,833,138	7.06	for	8,200,031	7.07	4.45
that	37,795,087	6.98	that	7,261,971	7.02	3.48
with	32,271,857	6.91	you	5,648,462	6.91	
...			...			
zzzzzd	4	0.01	zzzzzzzzzzzzs	1	0.15	
total	3,928,096,030		total	701,279,329		

The Method – Word Scoring

$$score(w) = \log_{10}\left(\frac{f(w) \cdot 10^9}{|D|}\right)$$

$f(w)$ ...corpus frequency of the word,
$|D|$ ...corpus size.

The Method – Document Scoring

$$\textit{document score}(\textit{language}) = \sum_{w \in \textit{document}} \textit{language score}(w)$$

$$\textit{paragraph score}(\textit{language}) = \sum_{w \in \textit{paragraph}} \textit{language score}(w)$$

Sample Paragraph From csTenTen17

= word score for	CS	xCS	SK	xSK	EN	DE	PL	SL	HR	FR
Solo	3.35	3.64	3.34	3.56	4.36	3.92	3.84	4.13	4.23	4.23
Pieces	2.29	2.58	2.33	2.54	4.76	2.75	2.42	2.48	2.72	3.56
for	4.49	4.77	4.47	4.69	7.07	4.67	4.64	4.56	4.93	4.62
La	4.62	4.90	4.55	4.77	4.93	4.80	4.63	4.61	4.77	7.42
Naissance	1.12	1.41	1.05	1.27	1.69	1.32	1.78	1.18	0.82	4.87
de	4.96	5.25	4.92	5.14	5.28	5.15	4.95	4.97	4.93	7.69
L	4.95	5.24	5.01	5.23	4.80	4.74	4.89	4.61	4.74	5.39
Amour	2.59	2.87	2.49	2.71	2.83	2.95	2.52	2.56	2.72	5.30
je	7.16	7.45	7.15	7.37	3.55	5.44	5.77	7.51	7.50	6.76
soundtrackové	1.30	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00
album	4.59	4.87	4.76	4.98	4.74	4.83	4.56	4.78	4.90	4.93
velšského	2.13	0.00	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00
multins...isty	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Johna	4.07	4.36	3.93	4.15	1.33	1.13	3.92	3.88	4.05	0.56
Calea	1.66	1.95	1.59	1.81	2.13	1.38	1.46	1.54	2.03	1.35
= sum	49.28	49.29	47.21	48.22	47.47	43.08	45.38	46.81	48.34	56.68

The Method – Confidence Ratio

$$\mathit{confidence\ ratio}(\mathit{document}) = \frac{\mathit{document\ score}(\mathit{top\ language})}{\mathit{document\ score}(\mathit{second\ top\ language})}$$

Confidence Ratio – Sample Sentence

```
<s lang="en-GB" confidence_ratio="1.018" en-GB="122.04" en-US="119.89">
Under      5.74    5.74
the        7.77    7.75
rent       4.70    4.59
deposit    4.56    4.40
bond       4.49    4.63
scheme     5.26    4.41
,          0.00    0.00
the        7.77    7.75
council    5.56    5.20
pays       4.20    4.26
the        7.77    7.75
deposit    4.56    4.40
for        7.06    7.07
a          7.36    7.34
tenant     4.34    3.94
```

```
so         6.34    6.31
they       6.51    6.50
can        6.53    6.54
rent       4.70    4.59
a          7.36    7.34
property   5.38    5.37
privately  4.05    3.99
.          0.00    0.00
```

```
</s>
```

Confidence Ratio – Examples Of Use

Czech vs. Slovak documents:

- CR $\geq 1.05 \Rightarrow$ keep, trust the prediction,
- oterwise throw the document away.

Confidence Ratio – Examples Of Use

Czech vs. Slovak documents:

- CR $\geq 1.05 \Rightarrow$ keep, trust the prediction,
- oterwise throw the document away.

American vs. Peninsular Spanish documents:

- CR $\geq 1.01 \Rightarrow$ keep, trust the prediction,
- oterwise keep, mark the document as in general Spanish.

Wordlist Sources And Sizes

1. Web = TenTens ∪ Aranea (Benko) ∪ WaCs (Reddy, Ljubešić)
   – limited to respective national TLDs
2. DSL Corpus Collection (VarDial 2014 workshop)
3. GloWbE (Davies)

Wordlist Sources And Sizes

1. Web = TenTens ∪ Aranea (Benko) ∪ WaCs (Reddy, Ljubešić) – limited to respective national TLDs
2. DSL Corpus Collection (VarDial 2014 workshop)
3. GloWbE (Davies)

Language	Web wordlist	DSL wordlist	GloWbE wordlist
Bosnian	2,262,136	51,337	
Croatian	6,442,922	50,368	
Serbian	3,510,943	49,370	
Indonesian	860,827	48,824	
Malaysian	1,346,371	34,769	
Czech	26,534,728	109,635	
Slovak	5,333,581	121,550	

Wordlist Sources And Sizes

1. Web = TenTens ∪ Aranea (Benko) ∪ WaCs (Reddy, Ljubešić)
   – limited to respective national TLDs
2. DSL Corpus Collection (VarDial 2014 workshop)
3. GloWbE (Davies)

Language	Web wordlist	DSL wordlist	GloWbE wordlist
Portuguese, Brazilian	9,298,711	52,612	
Portuguese, European	2,495,008	51,185	
Spanish, Argentine	6,376,369	52,179	
Spanish, Peninsular	8,396,533	62,945	
English, Great Britain	6,738,021	42,516	1,222,292
English, United States	2,814,873	42,358	1,245,821

Evaluation – VarDial 2014 Data

Overall accuracy using large web corpus wordlists
and DSL CC v. 1 training data wordlists
on DSL CC v. 1 gold data.

Languages	Wordlist	Accuracy	DSL Best
English GB/US	Large web corpora	**0.6913**	0.6394
English GB/US	GloWbE	**0.6956**	0.6394
English GB/US	DSL training data	0.4706	**0.6394**
Other languages	Large web corpora	0.8565	**0.8800**
Other languages	DSL training data	0.9354	**0.9571**
Bosnian, Croatian, Serbian	DSL training data	0.8883	**0.9360**
Indonesian, Malaysian	DSL training data	**0.9955**	**0.9955**
Czech, Slovak	DSL training data	**1.0000**	**1.0000**
Portuguese BR/PT	DSL training data	0.9345	**0.9560**
Spanish AR/ES	DSL training data	0.8820	**0.9095**

Evaluation – English Variants

Languages:

1. Australia, Canada, Great Britain, Ireland, New Zealand, USA,
2. the above and India, Philippines, Singapore, South Africa,
3. the above and Bangladesh, Ghana, Hong Kong, Jamaica, Kenya, Malaysia, Nigeria, Pakistan, Sri Lanka, Tanzania.

Evaluation – English Variants

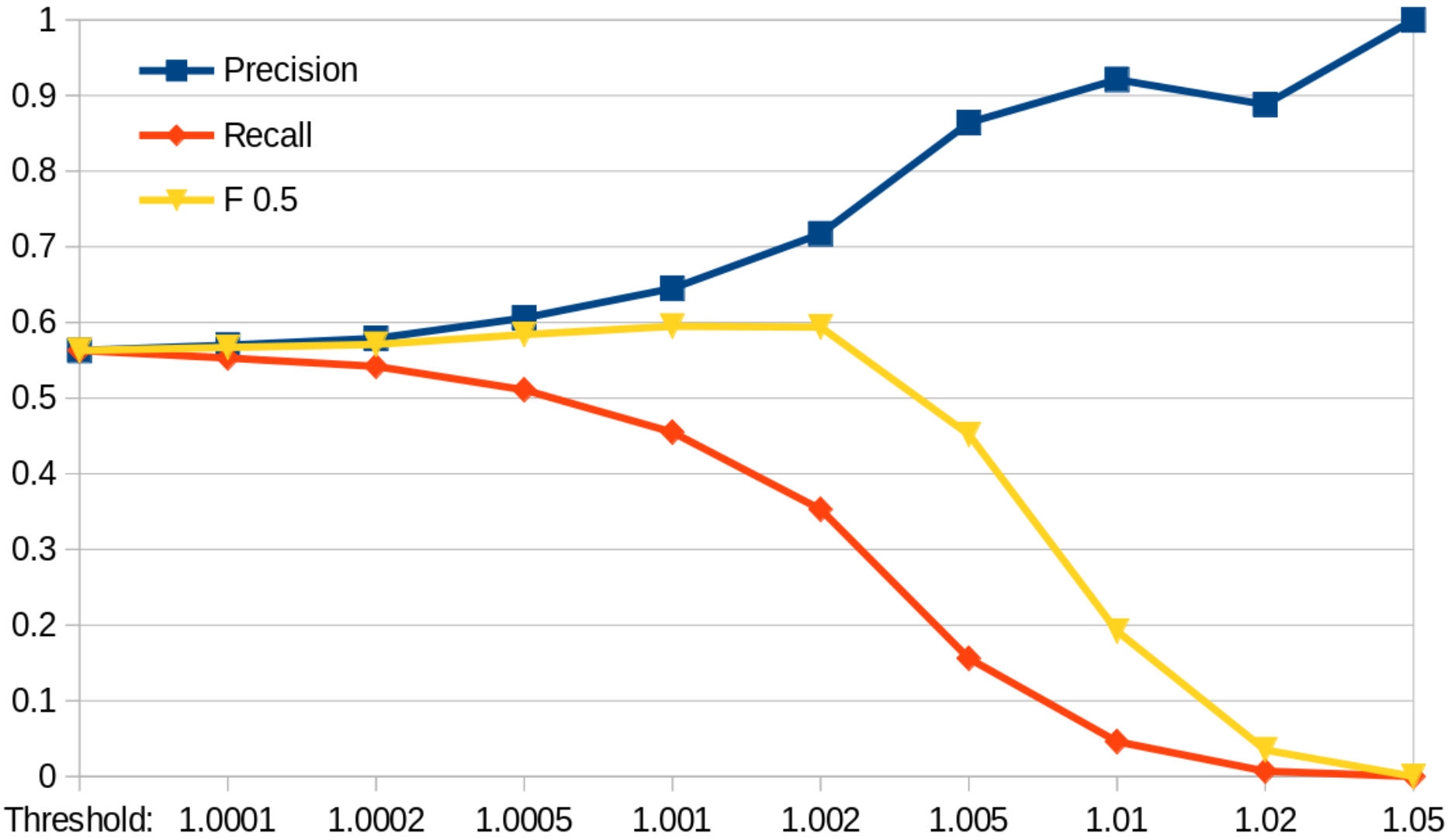
Languages:

1. Australia, Canada, Great Britain, Ireland, New Zealand, USA,
2. the above and India, Philippines, Singapore, South Africa,
3. the above and Bangladesh, Ghana, Hong Kong, Jamaica, Kenya, Malaysia, Nigeria, Pakistan, Sri Lanka, Tanzania.

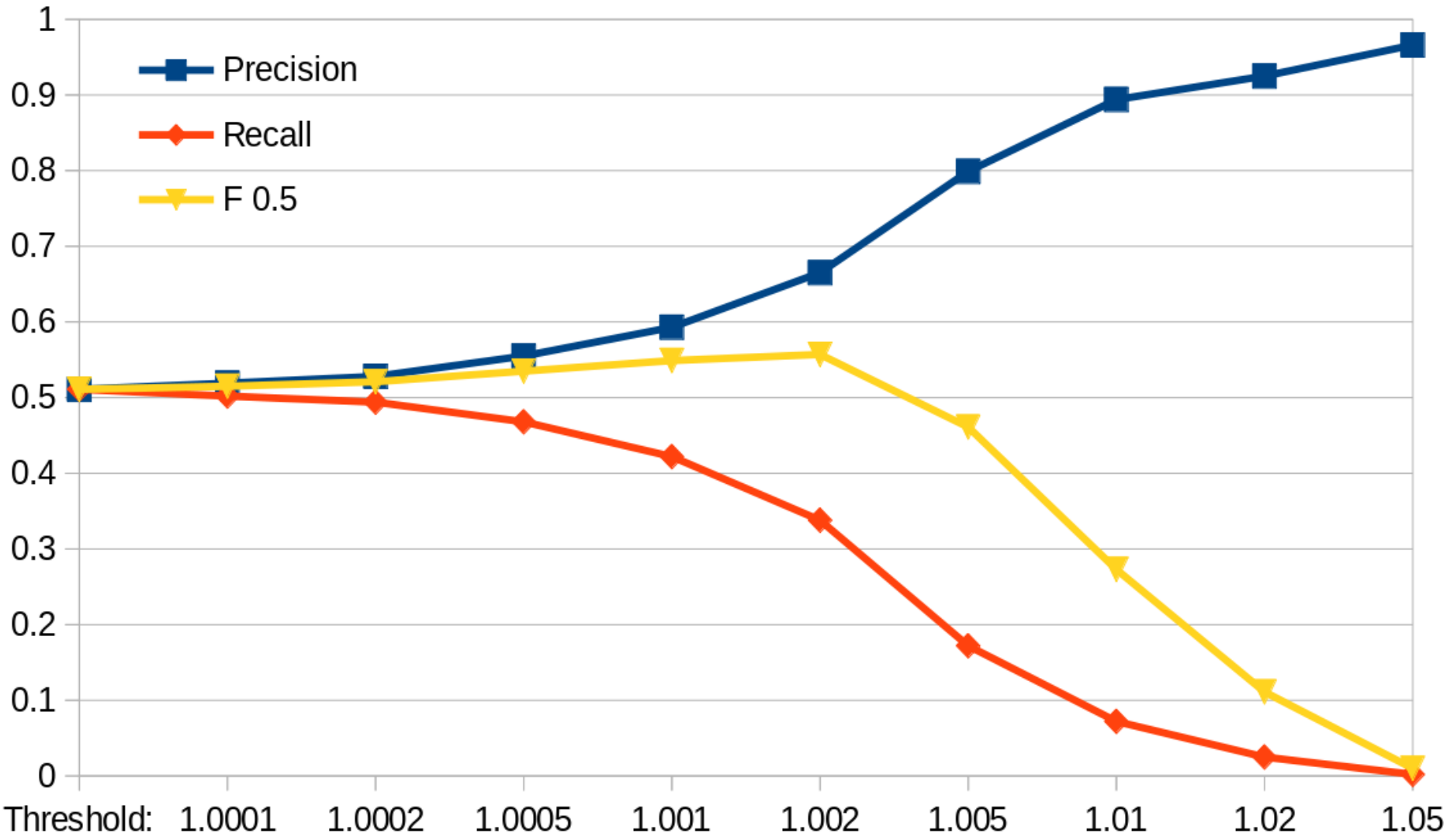
Wordlist sources:

1. enTenTen 08, 12, 13, 15, 18 docs from respective national TLDs,
2. GloWbE docs split to 99 % train set, 1 % evaluation set.

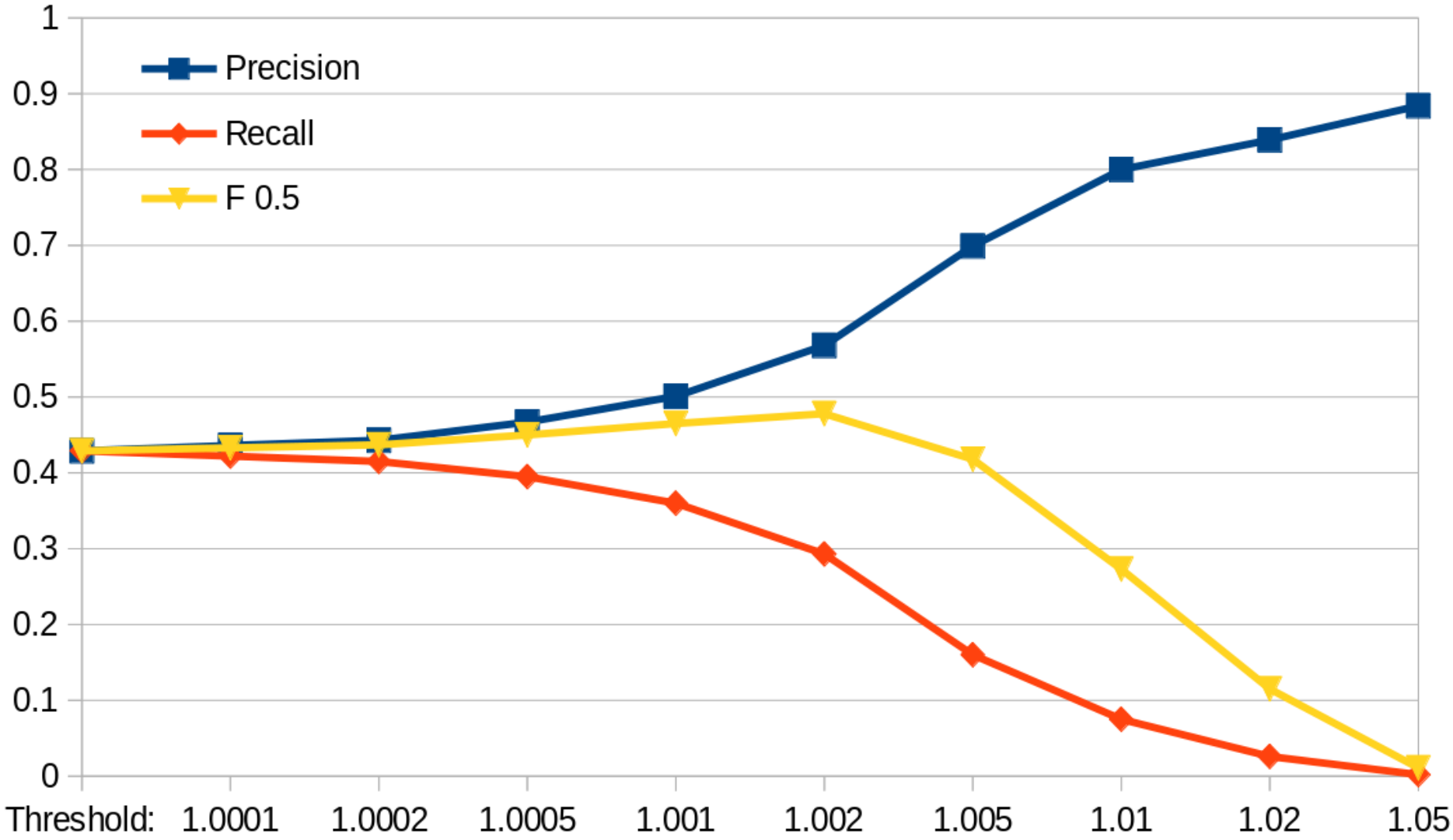
Evaluation – enTenTens on GloWbE – 6 Languages



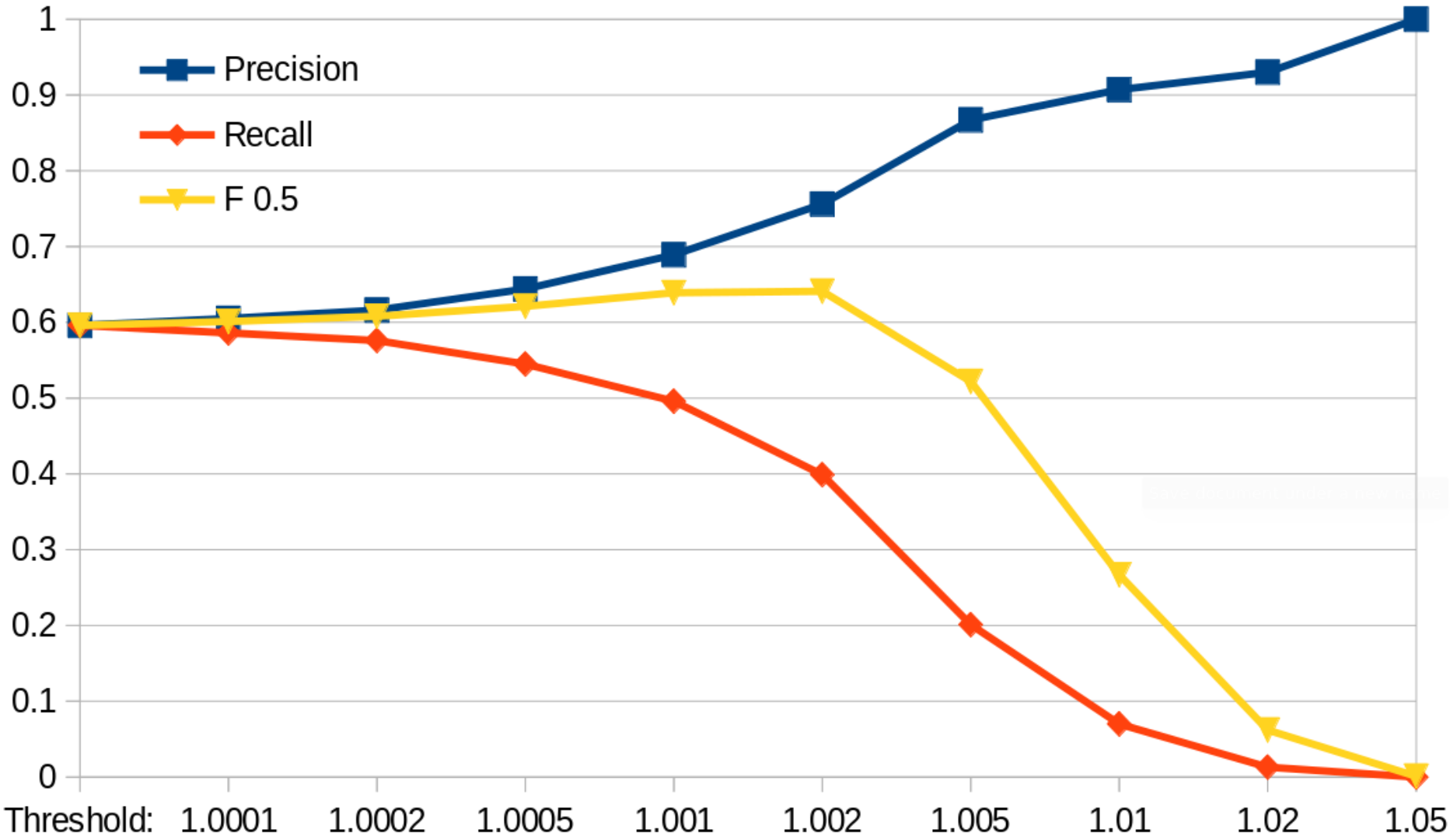
Evaluation – enTenTens on GloWbE – 10 Languages



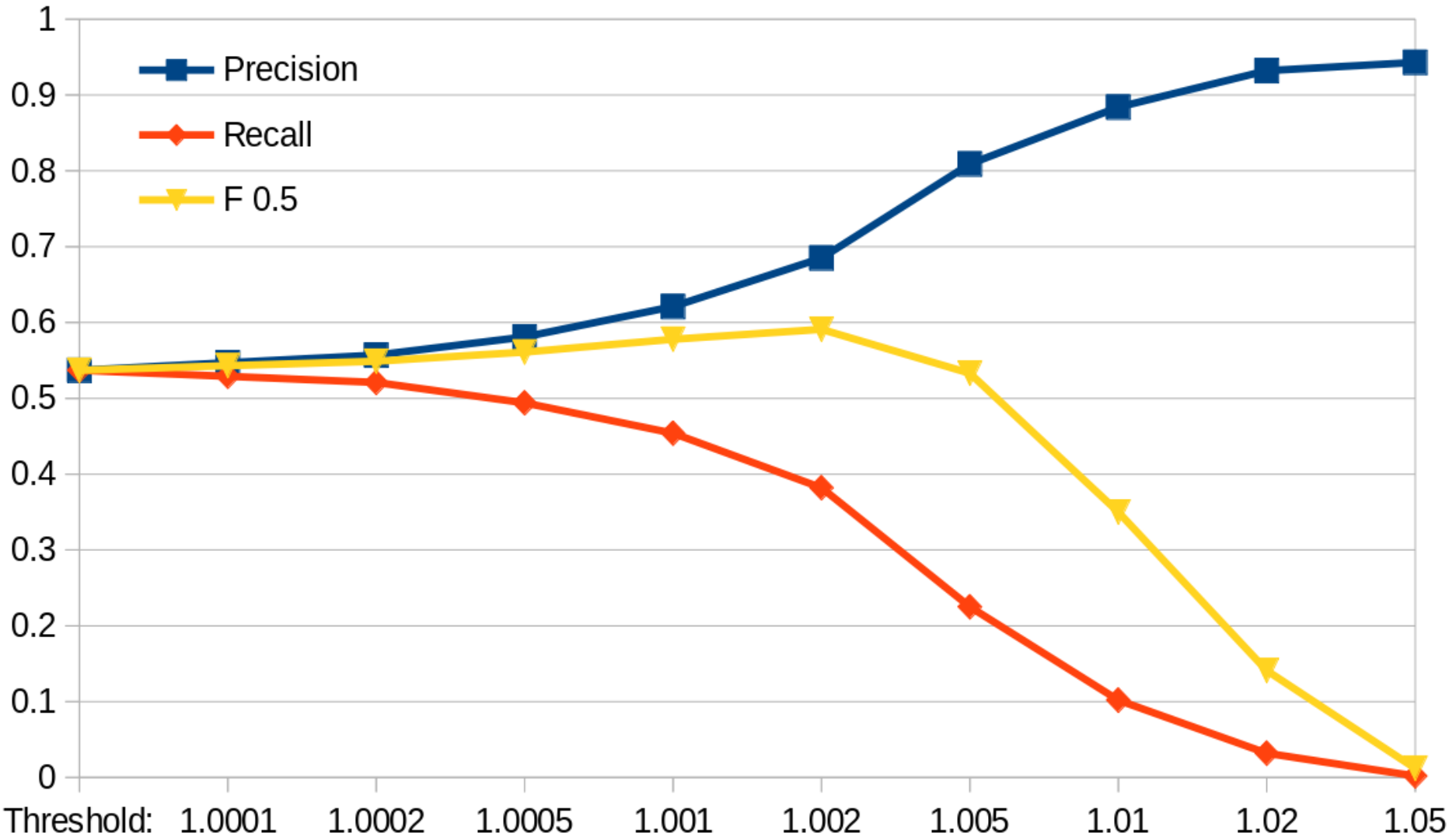
Evaluation – enTenTens on GloWbE – 20 Languages



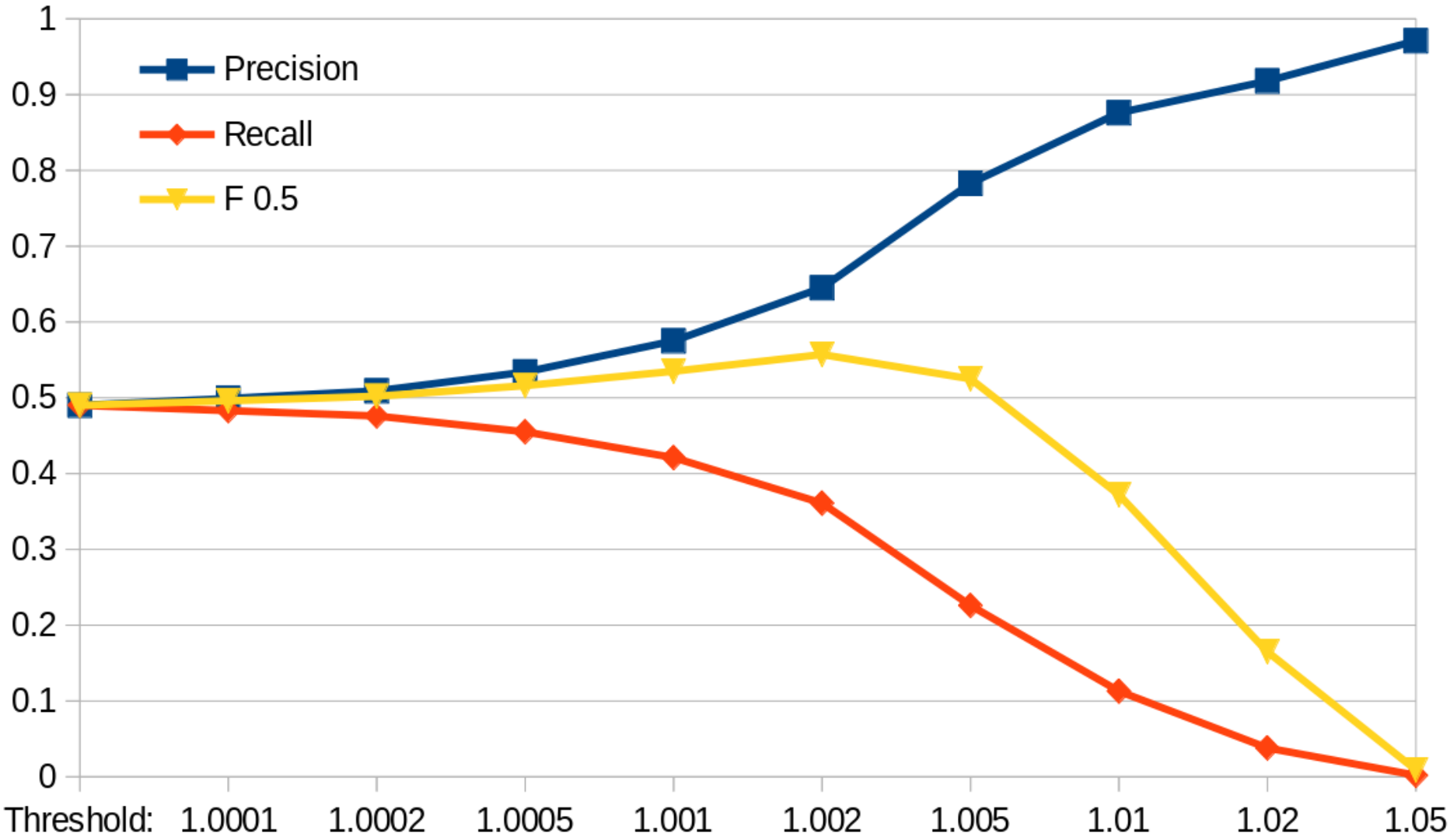
Evaluation – GloWbE on GloWbE – 6 Languages



Evaluation – GloWbE on GloWbE – 10 Languages



Evaluation – GloWbE on GloWbE – 20 Languages



English Variants Evaluation Summary

F 0.5 measure with confidence threshold 1.002

English data set	**enTenTen***	**GloWbE**
6 lang variants	0.594	0.641
10 lang variants	0.557	0.591
16 lang variants	0.478	0.557

- GloWbE train on GloWbE test is better than web corpora on GloWbE test by $\sim 3-8\,\%$ of F 0.5,
- web corpora still close enough to be usable.

Conclusion & Future Work

Achievements:

- Script and sample wordlists at `http://corpus.tools/`,
- easy to understand output and adjustable confidence threshold.

Future Work:

- Compare to langid.py, Google CLD 2/3,
- apply to corpora in Sketch Engine.

Thank you for your attention!

Photo credit: Kathleen & Ryan Rush