A bit of Progress and Stronger n-gram LM Baselines

Ehsan Shareghi*   Daniela Gerz   Ivan Vulić   Anna Korhonen

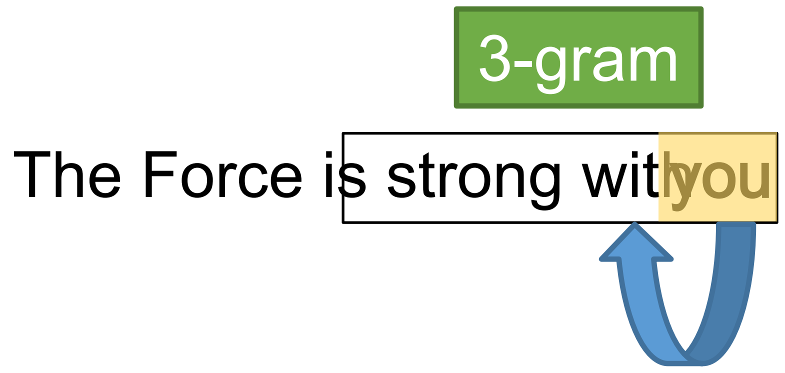
UNIVERSITY OF CAMBRIDGE

Language Technology Lab

NAACL 2019

n-gram LM

$$P(w_1^N) = \prod_{i=1}^{N} P(w_i | w_{i-n+1}^{i-1})$$



Questions:

- Do I know what is Kneser-Ney (KN), and what is Modified KN (MKN)?
- Has there been any progress in n-gram smoothing?
- Are NLM always superior to n-grams? When are they likely to fall short?

Smoothing in general

Donald Trump is a X

$$P(x = \text{politician} \mid \text{Donald Trump is a}) = \frac{\text{count(Donald Trump is a politician)}}{\text{count(Donald Trump is a)}}$$

count(Donald Trump is a politician) = 0
count(Trump is a politician) = 100

P(politician|Donald Trump is a) ≈ P(politician|Trump is a)

Smoothing- something old something new

$$\boxed{\beta(w_i|w_{i-n+1}^{i-1},\Theta)+\gamma(w_{i-n+1}^{i-1},\Theta)P(w_i|w_{i-n+2}^{i-1},\Theta)}$$

Smoothing- something old something new

$$\boxed{\beta(w_i \mid w_{i-n+1}^{i-1}, \Theta) + \gamma(w_{i-n+1}^{i-1}, \Theta) P(w_i \mid w_{i-n+2}^{i-1}, \Theta)}$$

	$\beta(w_i \mid w_{i-n+1}^{i-1}, \Theta)$	Θ
KN	$\dfrac{c(w_{i-n+1}^{i}) - D_n}{c(w_{i-n+1}^{i-1})}$	D_n

Smoothing- something old something new

$$\boxed{\beta(w_i|w_{i-n+1}^{i-1},\Theta)+\gamma(w_{i-n+1}^{i-1},\Theta)P(w_i|w_{i-n+2}^{i-1},\Theta)}$$

	$\beta(w_i\|w_{i-n+1}^{i-1},\Theta)$	Θ
KN	$\frac{c(w_{i-n+1}^{i})-D_n}{c(w_{i-n+1}^{i-1})}$	D_n
MKN	$\frac{c(w_{i-n+1}^{i})-D_n^{c(w_{i-n+1}^{i})}}{c(w_{i-n+1}^{i-1})}$	$D_n^{i\in\{1,2,3+\}}$

Smoothing- something old something new

$$\boxed{\beta(w_i \mid w_{i-n+1}^{i-1}, \Theta) + \gamma(w_{i-n+1}^{i-1}, \Theta) P(w_i \mid w_{i-n+2}^{i-1}, \Theta)}$$

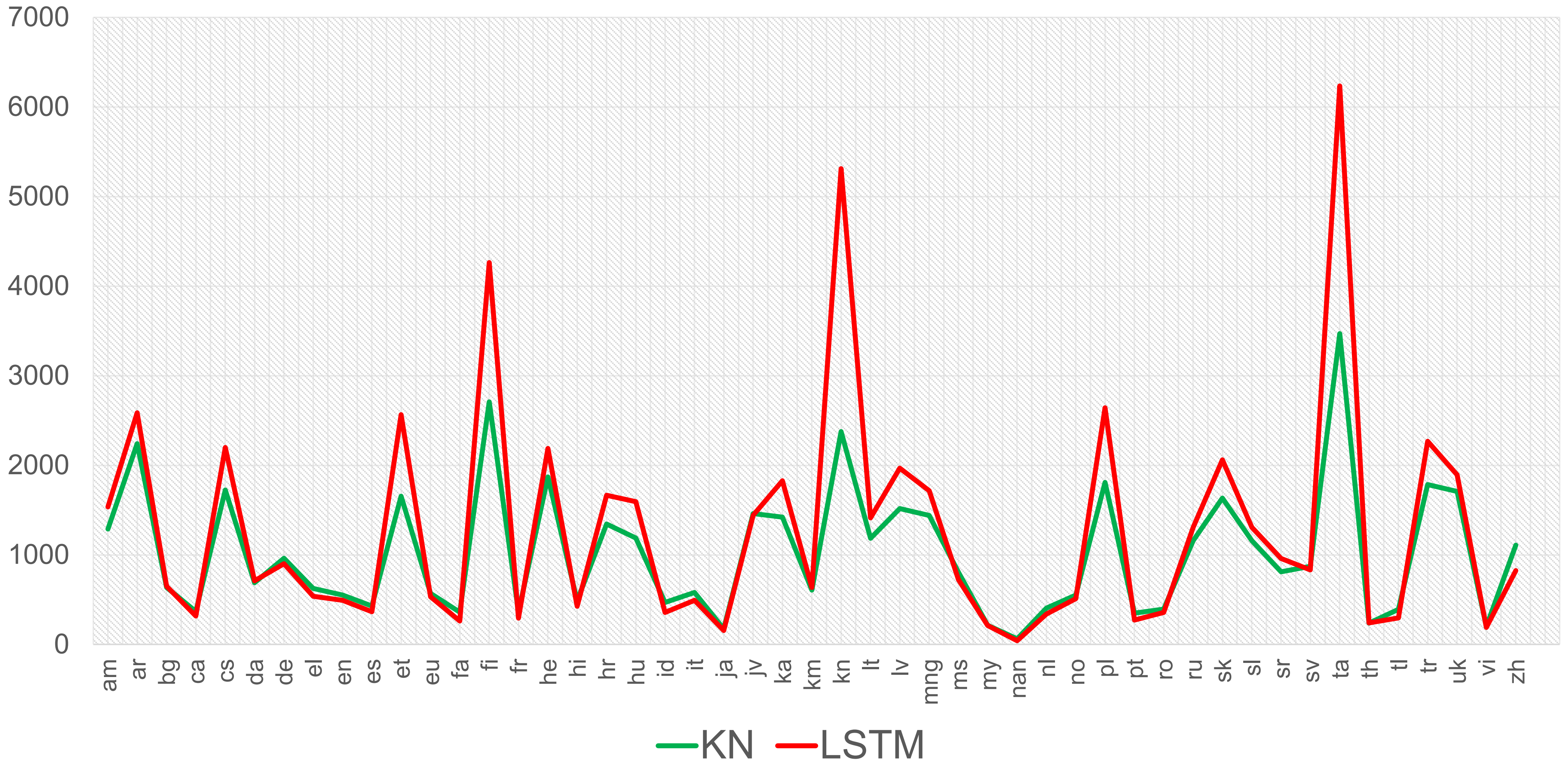
	$\beta(w_i \mid w_{i-n+1}^{i-1}, \Theta)$	Θ
KN	$\frac{c(w_{i-n+1}^{i}) - D_n}{c(w_{i-n+1}^{i-1})}$	D_n
MKN	$\frac{c(w_{i-n+1}^{i}) - D_n^{c(w_{i-n+1}^{i})}}{c(w_{i-n+1}^{i-1})}$	$D_n^{i \in \{1,2,3+\}}$
GKN	$\frac{c(w_{i-n+1}^{i}) - D_n^{c(w_{i-n+1}^{i})}}{c(w_{i-n+1}^{i-1})}$	$D_n^{i \in \{1,\ldots,10+\}}$

Smoothing- something old something new

$$\boxed{\beta(w_i|w_{i-n+1}^{i-1},\Theta)+\gamma(w_{i-n+1}^{i-1},\Theta)P(w_i|w_{i-n+2}^{i-1},\Theta)}$$

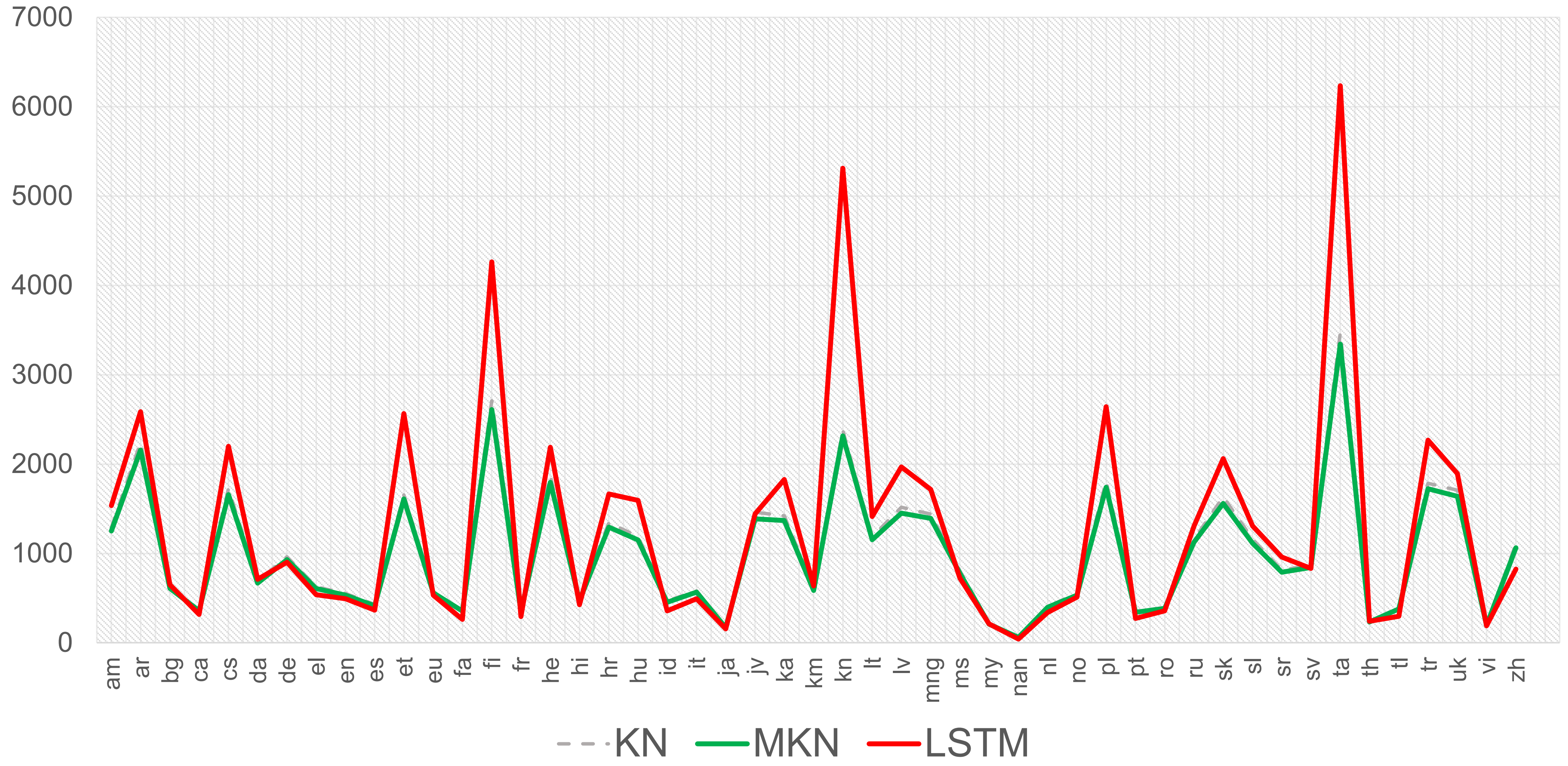
	$\beta(w_i \mid w_{i-n+1}^{i-1},\Theta)$	Θ
KN	$\frac{c(w_{i-n+1}^{i})-D_n}{c(w_{i-n+1}^{i-1})}$	D_n
MKN	$\frac{c(w_{i-n+1}^{i})-D_n^{c(w_{i-n+1}^{i})}}{c(w_{i-n+1}^{i-1})}$	$D_n^{i\in\{1,2,3+\}}$
GKN	$\frac{c(w_{i-n+1}^{i})-D_n^{c(w_{i-n+1}^{i})}}{c(w_{i-n+1}^{i-1})}$	$D_n^{i\in\{1,...,10+\}}$
BKN	$\frac{c(w_{i-n+1}^{i})-D_{w_{i-n+1}}t_{w_{i-n+1}}^{w_i}}{c(w_{i-n+1}^{i-1})+\theta_{w_{i-n+1}}}$	$D_{w_{i-n+1}},\theta_{w_{i-n+1}},t_{w_{i-n+1}}^{w_i}$

Perplexity



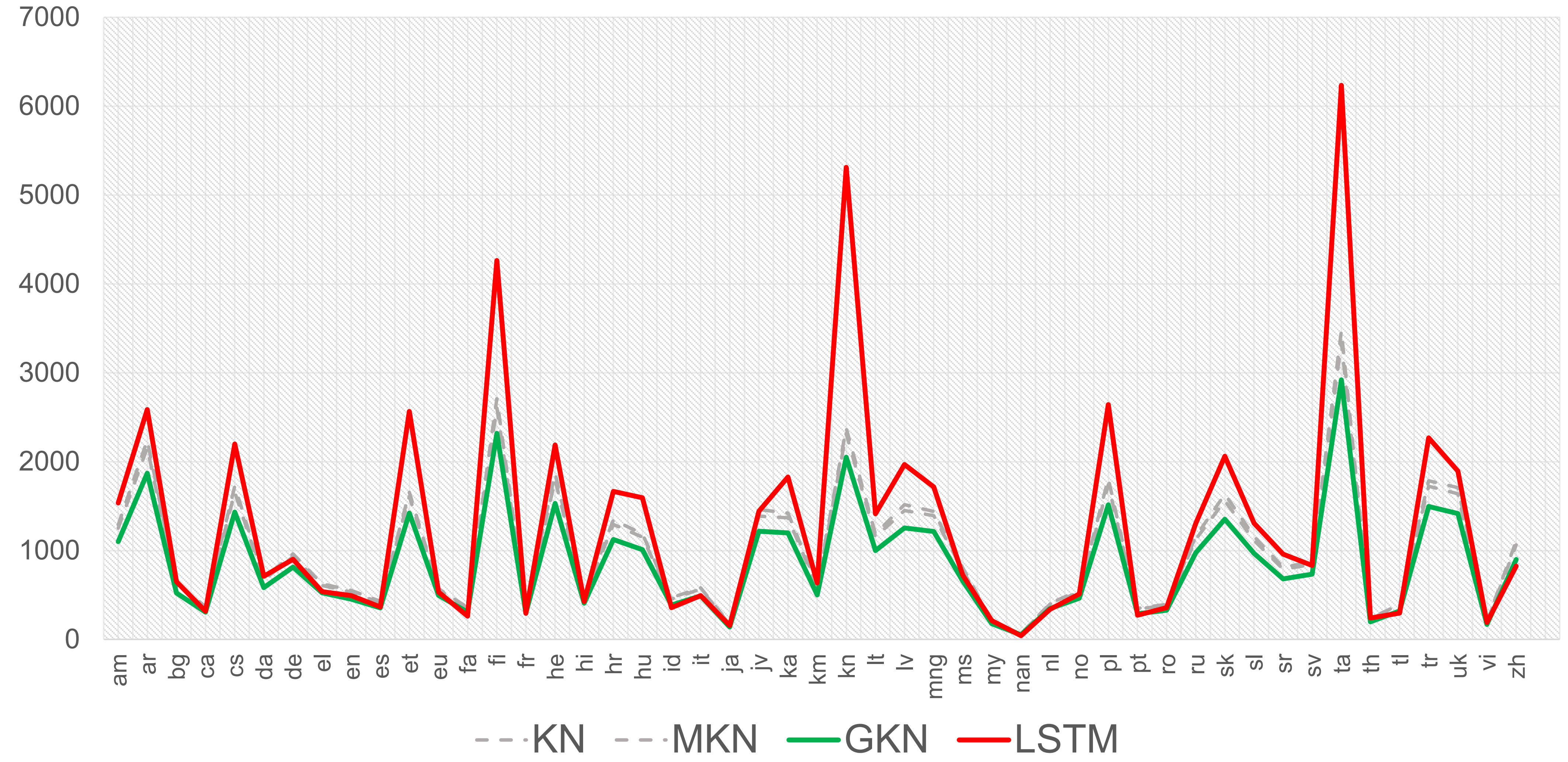
1116 1460

Perplexity



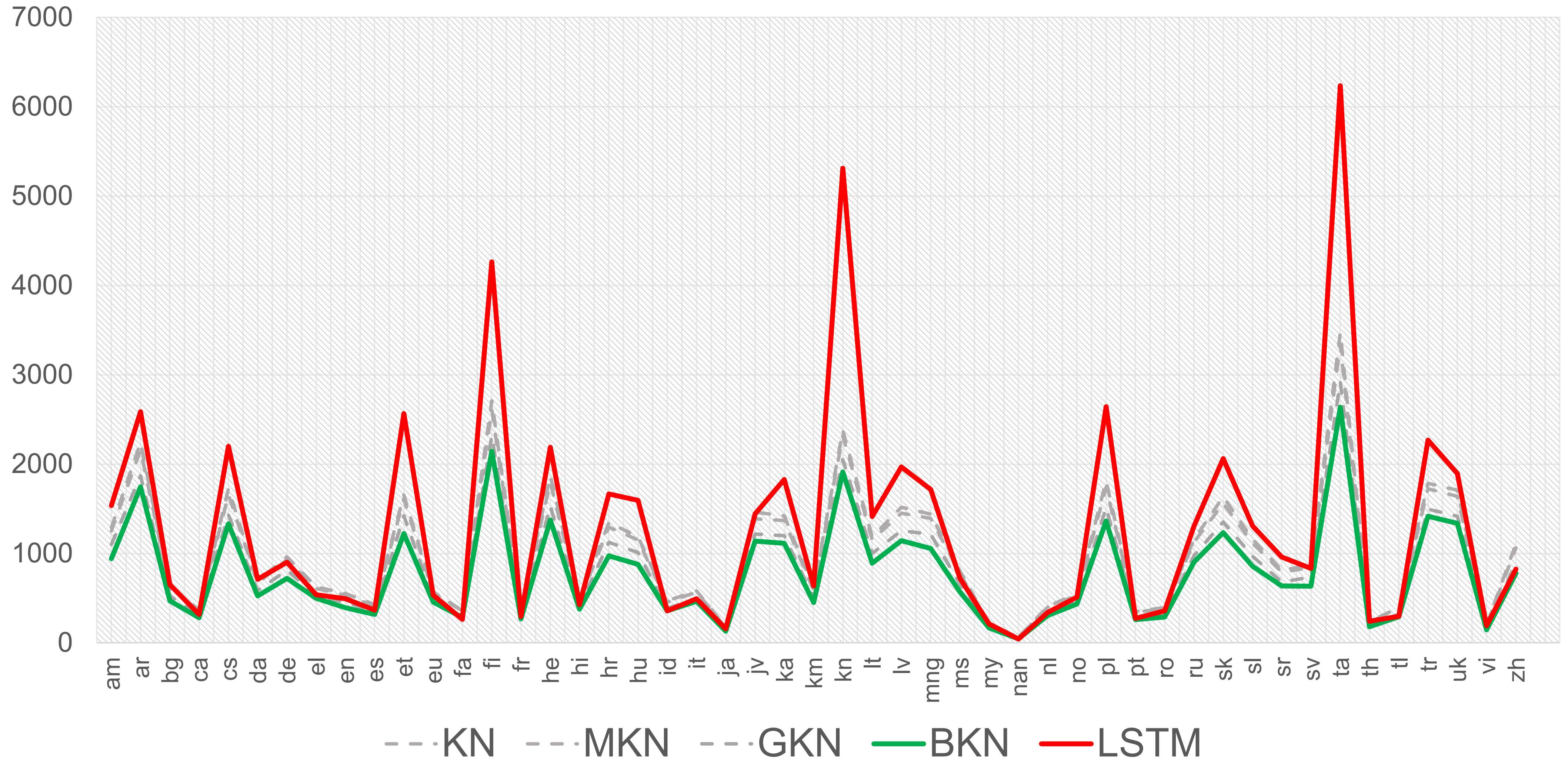
1116 1077 1460

Perplexity



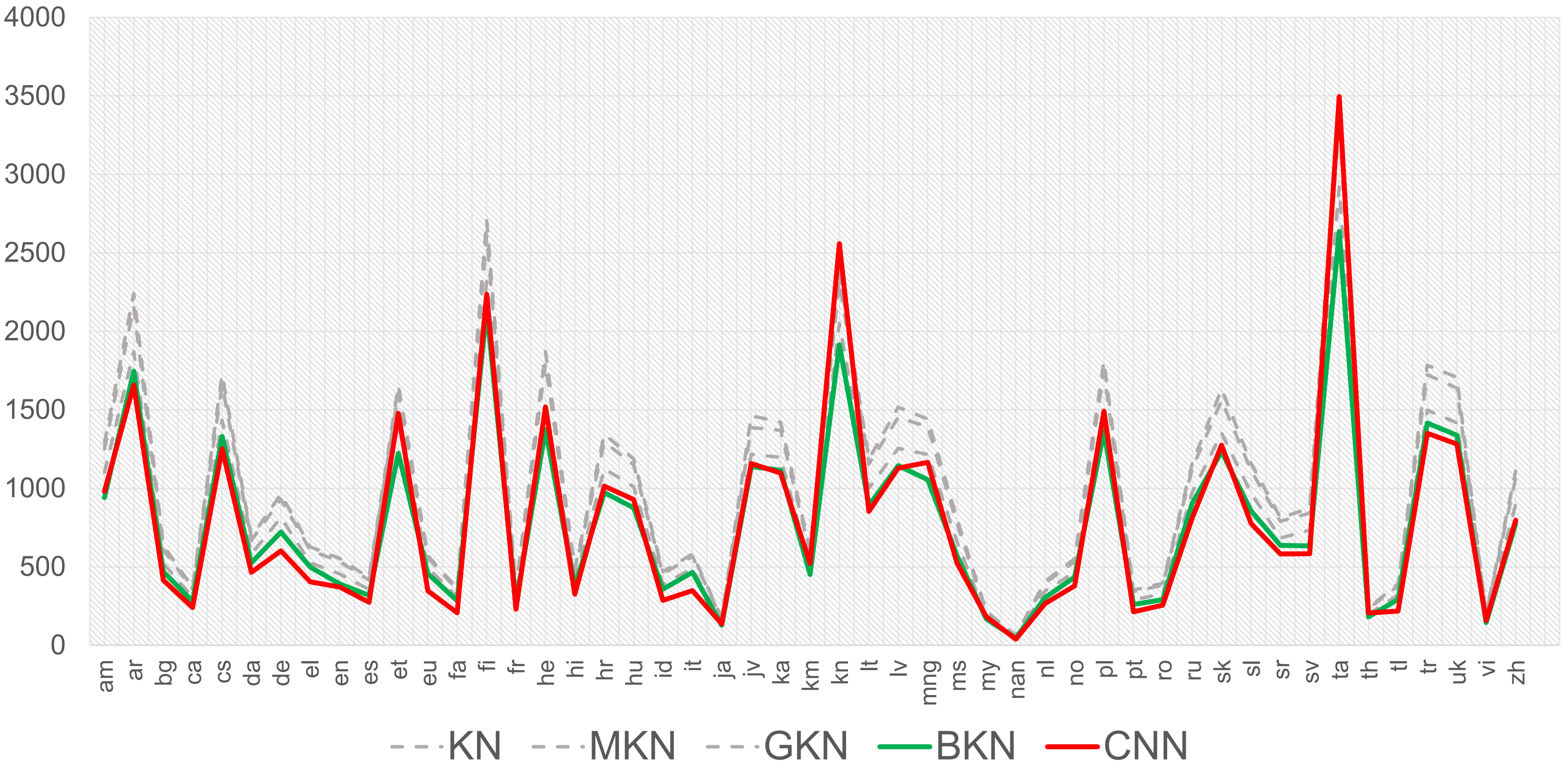
KN	MKN	GKN	LSTM
1116	1077	936	1460

Perplexity



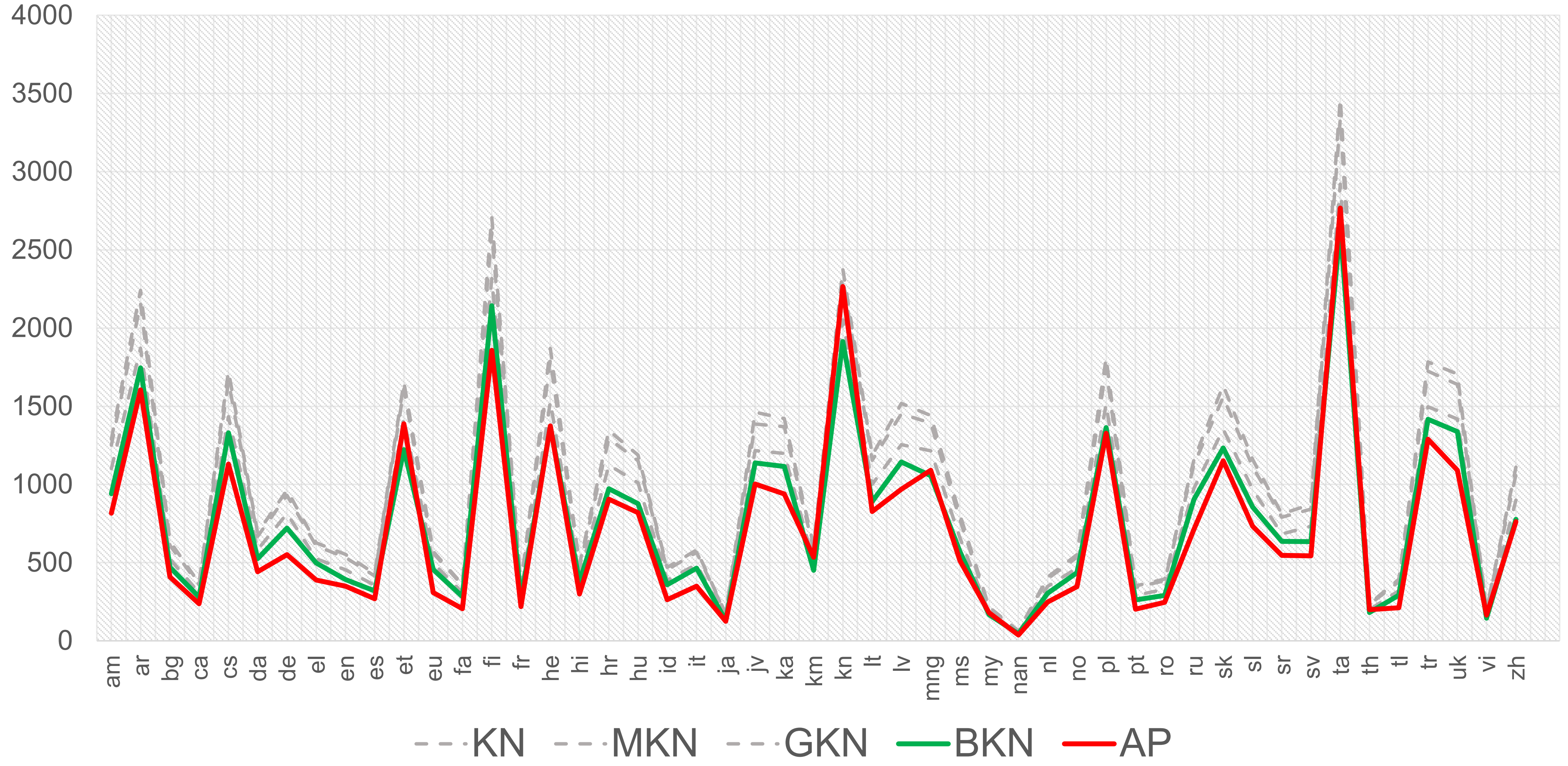
KN	MKN	GKN	BKN	LSTM
1116	1077	936	847	1460

Perplexity

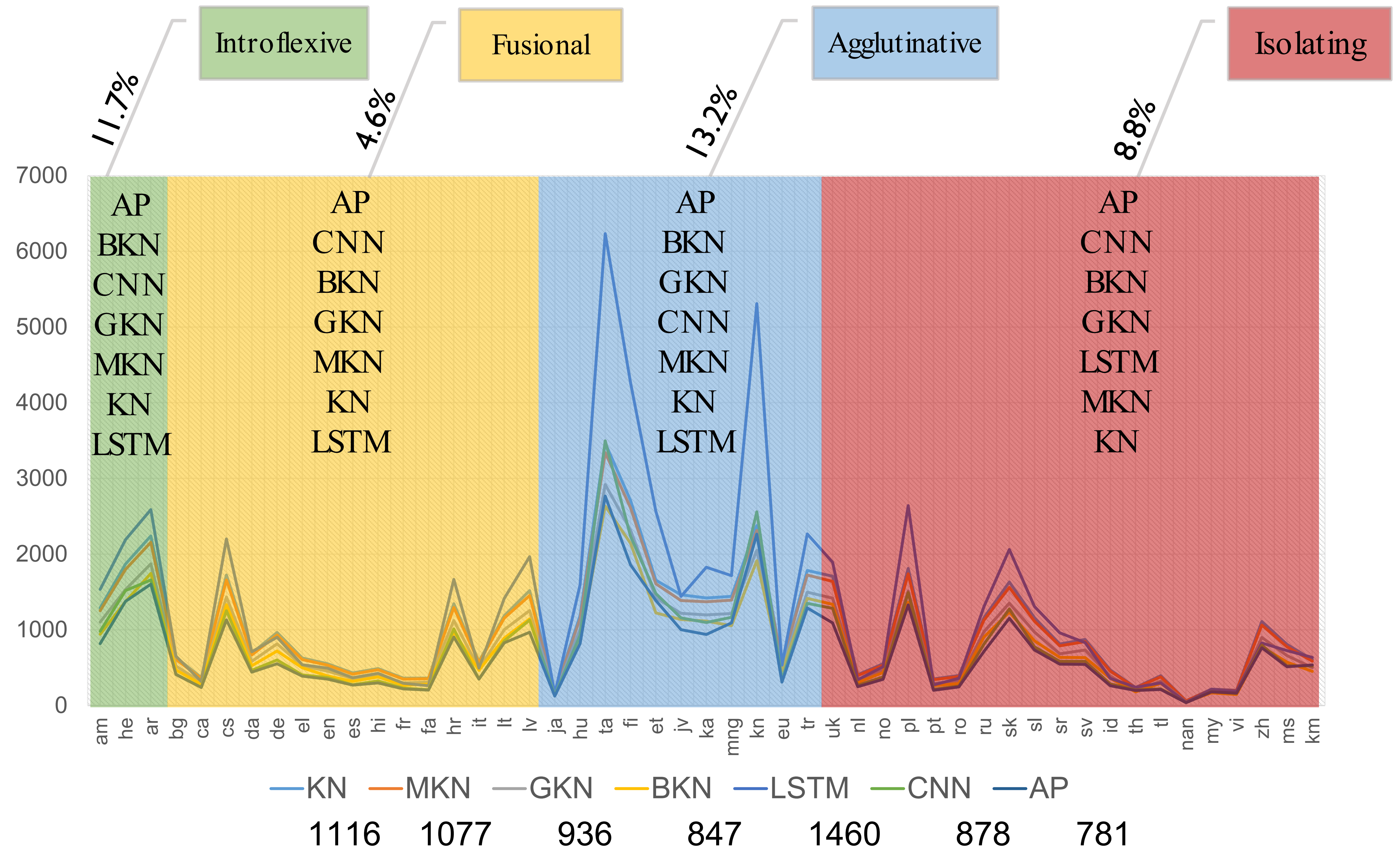


KN	MKN	GKN	BKN	CNN
1116	1077	936	847	878

Perplexity

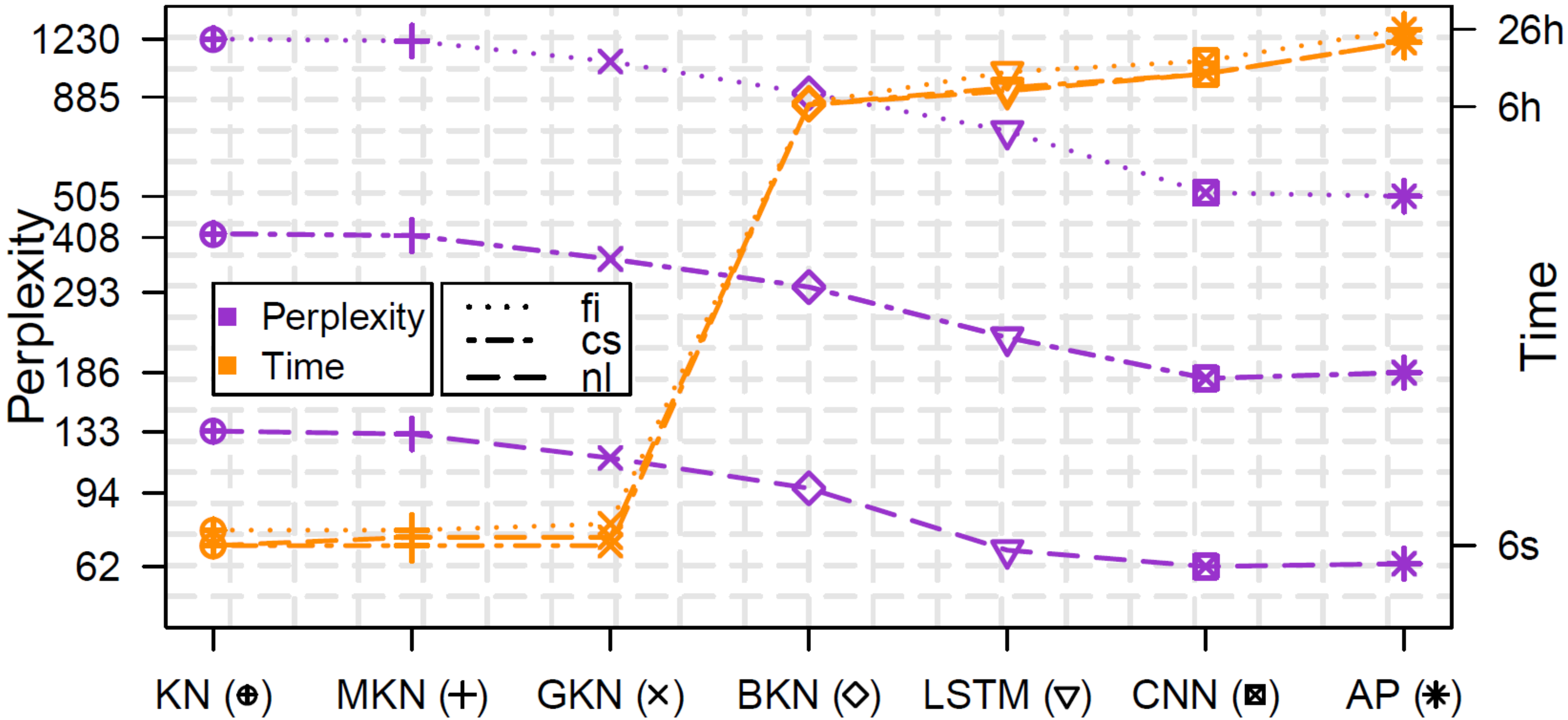


KN	MKN	GKN	BKN	AP
1116	1077	936	847	781



Introflexive
Fusional
Agglutinative
Isolating
11.7%
4.6%
13.2%
8.8%
AP
BKN
CNN
GKN
MKN
KN
LSTM
AP
CNN
BKN
GKN
MKN
KN
LSTM
AP
BKN
GKN
CNN
MKN
KN
LSTM
AP
CNN
BKN
GKN
LSTM
MKN
KN
7000
6000
5000
4000
3000
2000
1000
0
am he ar bg ca cs da de el en es hi fr fa hr it lt lv ja hu ta fi et jv ka mng kn eu tr uk nl no pl pt ro ru sk sl sr sv id th tl nan my vi zh ms km
KN MKN GKN BKN LSTM CNN AP
1116 1077 936 847 1460 878 781

Training Time



Wrapping up …

- n-grams are highly competitive with neural LMs for low-resource setting, high OOV ratio, or for languages with high type-token ratio

- Recent developments in n-gram models permit to lift the finite-order Markov assumption, hence in theory models should be capable of capturing long range dependencies

- The gap between neural and stand-alone n-gram models could be reduced by (somehow) incorporating continuous word representations into n-gram models

- n-gram models have far more attractive computational properties (Memory/Time usage) for both training and inference steps. So invest in improving neural models computational shortcomings.

Thanks! ☺

Contact: Ehsan Shareghi - es776@cam.ac.uk