Twitter Thread by Balaraman Ravindran

Balaraman Ravindran
@ravi_iitm

AI4Bharat/@iitmadras and EkStep Foundation announced the release of Samanantar, the largest publicly available collection of parallel corpora for Indic languages. This work was supported by Tarento Technologies and the @rbc_dsai_iitm Download:

A few Highlights: 1. A total of 46.9M parallel sentences between English and 11 Indic languages
2. Of these, 34.6M parallel sentences were newly mined as a part of this work (~33M from IndicCorp that we released last year)

3. A total of 82.7M parallel sentences between 11C2 Indic language pairs extracted from the above English-centric parallel corpus
4. A new single script joint model for En-Indic and Indic-En translation which outperforms all commercial and publicly available systems.

5. A cross lingual semantic similarity benchmark containing ~30K annotations for ~9.5K sentence pairs from Samanantar certifying the quality of the mined data.

This is a tremendous effort that will have a great impact on Indic NLP research! @OfficialIndiaAI

Work spearheaded by my wonderful colleagues Mitesh Khapra and Pratyush Kumar!