Evaluation of Machine Translation in Languages

Stefan Liemawan Adji

July 5, 2024

1 Objectives

This paper aims to evaluate existing techniques and circumstances around the task of machine translation across different languages.

2 Related Work

3 Methodology

Take the list of most spoken language by population according to Wikipedia [1], then take one for each branch.

No.	Language	Native speakers (in millions)	Language family	Branch
1	Egyptian Arabic	78	Afroasiatic	Semitic
2	German	76	Indo-European	Germanic
3	Hausa	54	Afroasiatic	Chadic
4	Hindi	345	Indo-European	Indo-Aryan
6	Japanese	123	Japonic	Japanese
7	Javanese	68	Austronesian	Malayo-Polynesian
8	Korean	81	Koreanic	—
9	Mandarin Chinese	941	Sino-Tibetan	Sinitic
5	Persian	62	Indo-European	Iranian
10	Russian	148	Indo-European	Balto-Slavic
11	Spanish	486	Indo-European	Romance
13	Telugu	83	Dravidian	South-Central
12	Turkish	84	Turkic	Oghuz
14	Vietnamese	85	Austroasiatic	Vietic

Table 1: List of most spoken languages per branch

Egyptian Arabic, Javanese, and Telugu might be discarded due to low examples in dataset.

References

[1] Wikipedia contributors. *List of languages by number of native speakers — Wikipedia, The Free Encyclopedia*. [Online; accessed 5-July-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=List_of_languages_by_number_of_native_speakers&oldid=1231985127.