Pre-training Data Quality for Low-Resource Languages

New Corpus and BERT Models for Maltese

Micallef, K., Gatt, A., Tanti, M., van der Plas, L., & Borg, C. (2022)

Etat de l'art

Chau et al. (2020) et Muller et al. (2021) :

Pré-entraînement de mBERT sur une langue spécifique

—> amélioration considérable des performances

Korpus Malti v4.0

data subset	documents	sentences	tokens	size
belles_lettres	195	299 762	4 454 906	21.82MB
blogs	25 436	807 628	14 562 039	74.45MB
comics	62	2 413	44 768	233.22KB
court	2 663	694 227	11 881 638	61.91MB
eu_docs	2 974	5 099 564	135 811 945	773.25MB
government_gazette	2 974	1 881 034	39 771 556	203.61MB
gov_docs	272	120 209	1 900 842	10.79MB
law_eu	71	4 433 235	98 582 031	541.13MB
law_mt	2 596	401 118	7 631 651	38.84MB
legal	3	4 784	83 581	490.67MB
nonfiction	2 177	208 763	3 902 436	20.01MB
parliament	6 198	3 935 906	82 294 520	433.09MB
press_eu	5 483	413 317	9 774 919	55.73MB
press_mt	46 782	713 886	17 679 904	93.15MB
speeches	62	2 067	51 259	286.63MB
theses	19	11 545	310 243	1.63MB
umlib_oar	11 688	963 606	21 235 949	106.11MB
web_general	2	685 873	14 741 525	75.22MB
wiki	3 469	79 134	1 885 661	9.73MB
all	131 429	20 758 071	466 601 373	2.52GB

Korpus Malti v4.0, BERTu et mBERTu

Korpus Malti v4.0

mBERT

BERTu

- monolingue
- entraîné à partir de zéro sur KM v4.0.

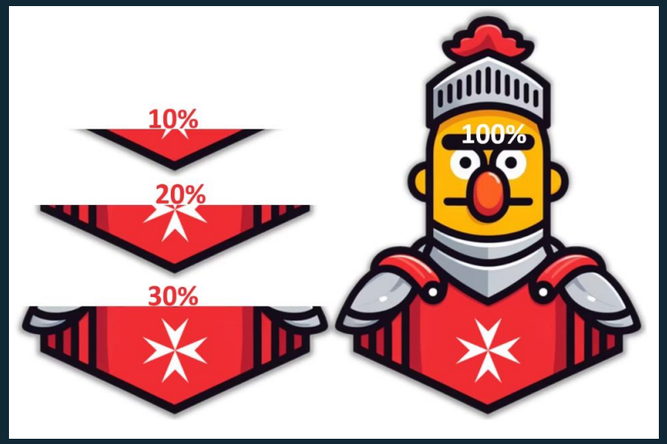
mBERTu

- multilingue
- Modèle mBERT + affinage sur KM v4.0.

Facteurs clé du pré-entraînement

Impact du domaine du corpus

- XXX



Impact de la taille du corpus

- Simulation de divers niveaux de richesse linguistique
- Nombre d'époques identiques

Tâches - syntaxiques et sémantiques - de référence

Pour mesurer différents niveaux de compréhension du langage

1

Etiquetage morphosyntaxique

Part-of-Speech tagging

2

Analyse syntaxique des dépendances

Dependency parsing

3

Reconnaissance d'entités nommées

Named entity Recognition

4

Analyse de sentiments

Sentiment analysis

Evaluation des résultats

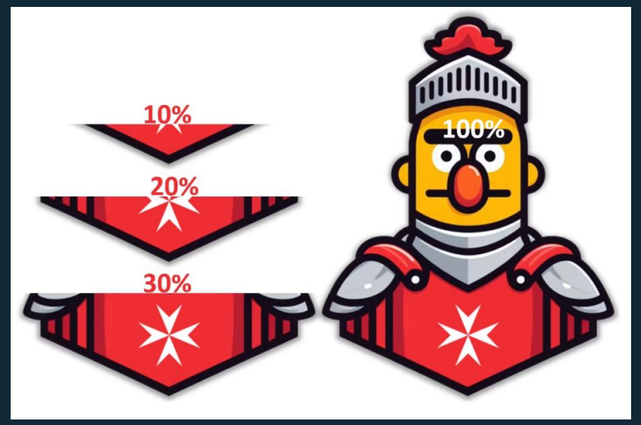
1. Réalisation par chaque modèle (BERTu et mBERTu) de chaque tâche de référence (1, 2, 3, 4)
2. Comparaison des résultats à des étiquettes de référence produites manuellement
3. Obtention d'un score pour chaque tâche
4. Moyenne sur 5 exécutions pour réduire les effets aléatoires —> fiabilité

Résultats obtenus & conclusions

Impact du domaine du corpus

- corpus mixte > Wikipédia
  - Exception : REN

Impact de la taille du corpus



- Influence non linéaire de la quantité de données sur les performances :
  - Les modèles entraînés sur 10% KM v4.0 surpassent déjà mBERT
  - Nette amélioration entre 10 et 50%
  - Gains marginaux au-delà de 50% (seuil de qualité)

Complémentarité BERTu / mBERTu

BERTu

- Monolingue
- Entraîné de zéro sur Korpus Malti v4.0

- Meilleur sur les tâches sémantiques
- Sensible aux nuances lexicales et particularités morpho

—> Compréhension linguistique fine et spécifique à la langue

mBERTu

- Multilingue
- Affiné sur Korpus Malti v4.0

- Légèrement meilleur sur les tâches syntaxiques
- Transfert interlinguistique issu du pré-traîtement multilingue initial avec mBERT

—> Généralisation des structures grammaticales communes à plusieurs langues

Conclusions

- Un pré-entraînement ciblé sur un corpus monolingue de qualité, même de petite taille, est plus bénéfique qu'un modèle multilingue pré-entraîné sur des données massives mais hétérogènes.
- La qualité du pré-entraînement joue un rôle déterminant dans la performance des modèles de langue (cf Chau et al. (2020) et Muller et al. (2021)).

Perspectives pour les langues peu dotées

Des modèles performants peuvent être réalisés à un coût relativement faible avec des corpus propres et bien structurés, même de taille réduite.

—> Ouvre des voies à la création de ressources locales de qualité pour d'autres langues peu dotées.