Data Quality for Deep Learning

DL学習とデータ品質

October 24, 2018

Carl Hoffman, CEO
Isao Tanner, Technical Consultant
Basis Technology

Ubiquitous Deep Learning

Deep learning is being applied to many fields

DLは多くの分野に適用されている

- Image recognition 画像認識
- Speech スピーチ

AND

Text Analytics テキスト分析

Image source: www.markable.ai

Deep Learning Applications

BASIS TECHNOLOGY

Product sentiment analysis, business intelligence, predicting civil unrest,etc.

製品への感情分析、ビジネスインテリジェンスの向上、暴動の予測

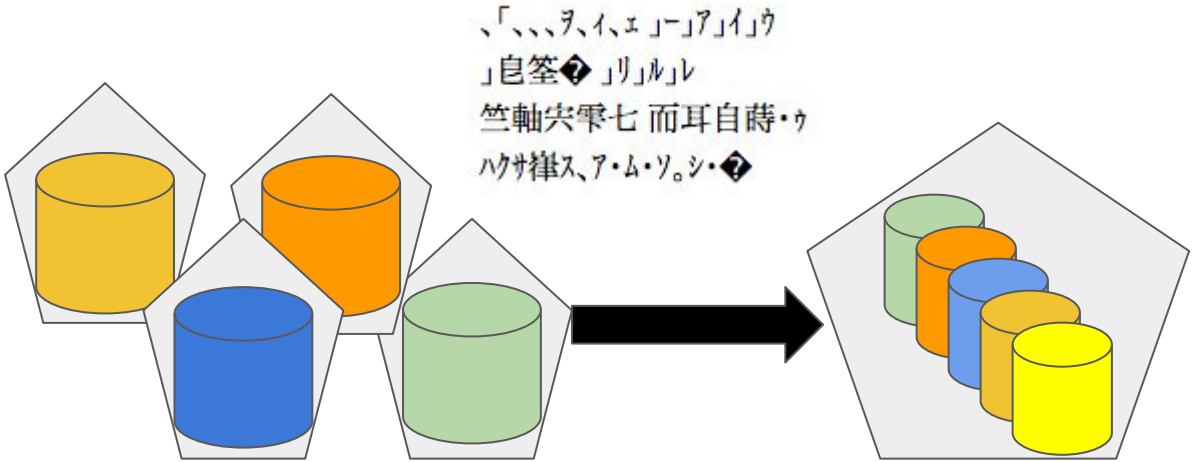
Photo by rawpixel on Unsplash

Photo by rawpixel on Unsplash

Photo by Jonathan Harrison on Unsplash

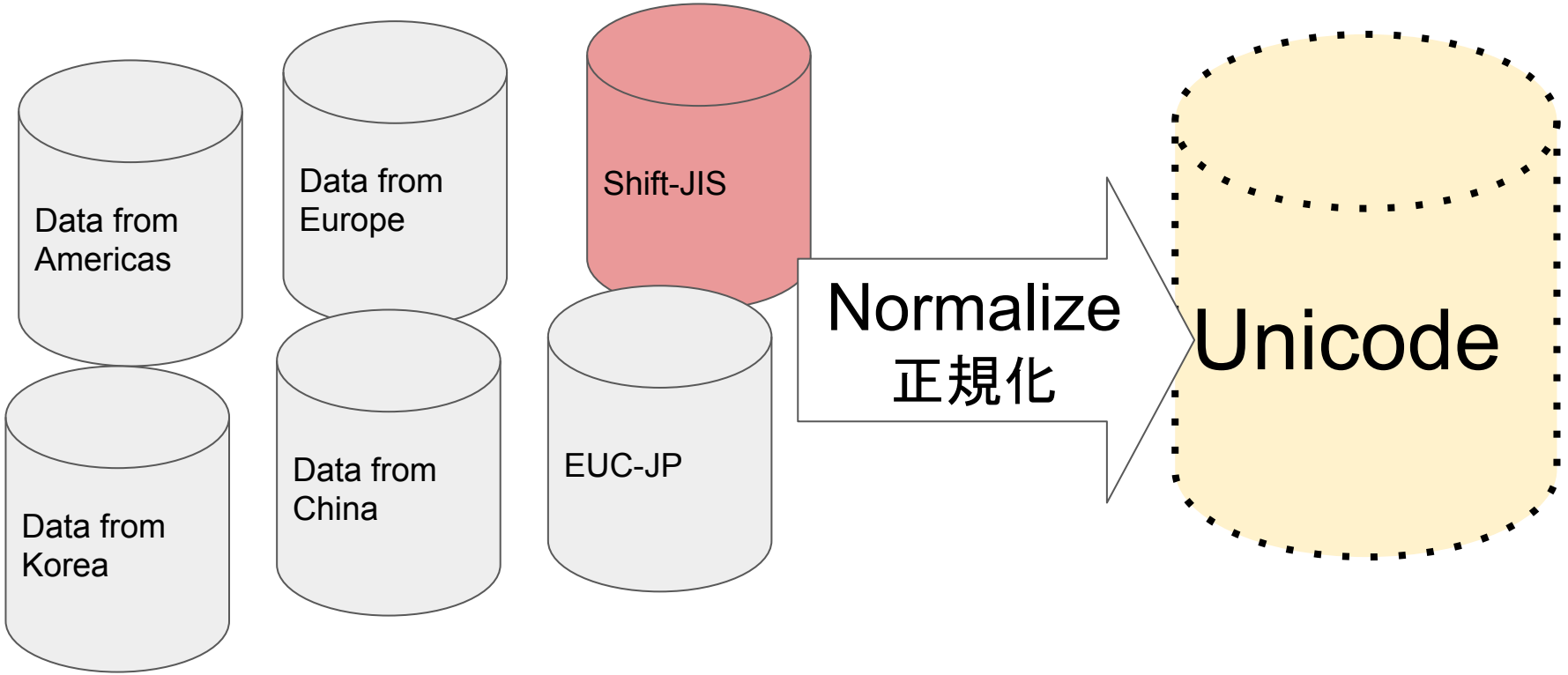
Deep Learning & the Real-World
DLと現実

- Deep learning needs “clean data” to train on
  DLのトレーニングにはクリーンなデータが必要
- Real world data is “dirty” でも実際には不良なデータが混在する



Japanese-Specific Data Quality Issues

日本固有のデータ問題：漢字・ひらがな・カタカナ



Why Is “Dirty Data” a Problem for DL?
なぜ正規化が重要か

バーガーキングのワッパーが美味しい。
バーガーキングのワッパーが美味しい。
ﾊﾞｰｶﾞｰｷﾝｸﾞのワッパーが美味しい。

Same sentence will be “learned” as separate examples

同じ文書が違うものと学習される。

ガ

- ガ u30AC 全角
- カ＋゛ u30AB+u3099
- ｶ+ﾞ uFF76+uFF9E 半角

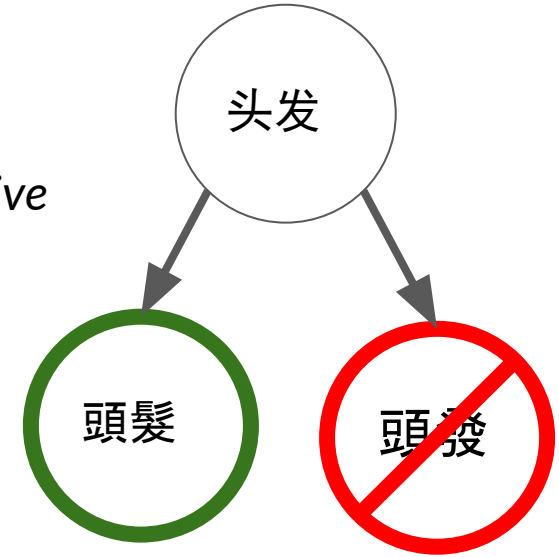
- が u304C
- か＋゛ u304B+u3099

Normalization for Other Langauges
他の言語での正規化

Chinese

Context sensitive Simplified and Traditional script Conversion

中国語:
繁体字と簡体字の対応

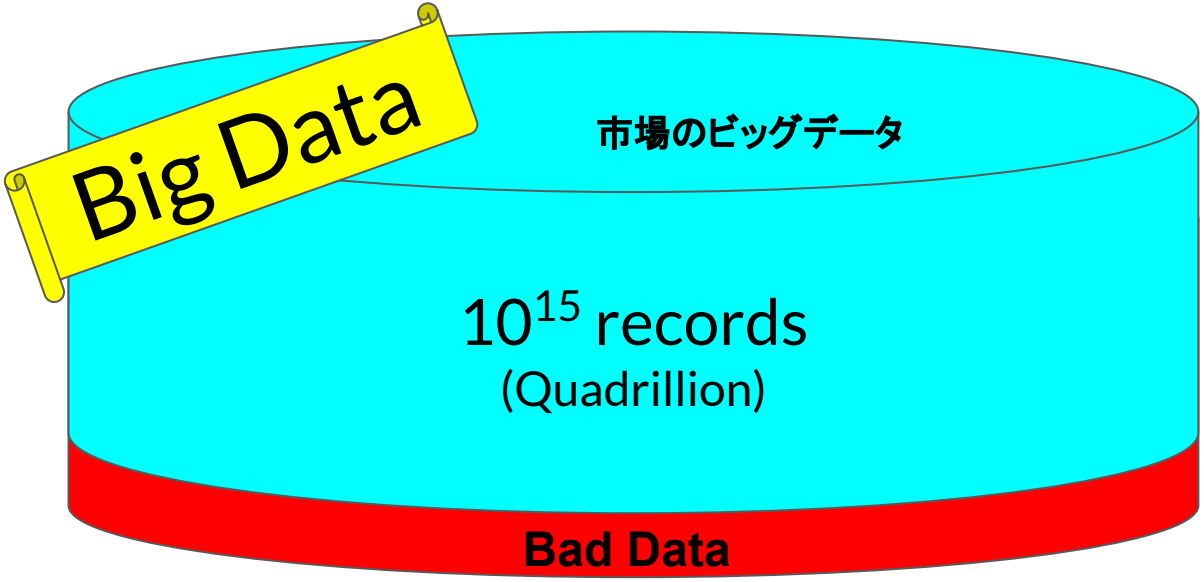


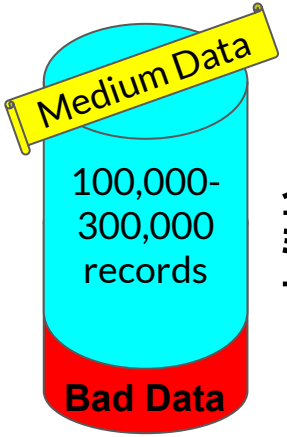
Arabic アラビア語

Yeh with hamza above: The following combinations are converted to ئ (U+0626).
ى (U+06CC) combined with hamza above (U+0654)
ى (U+0649) combined with hamza above (U+0654)
ي (U+064A) combined with hamza above (U+0654)

“Big Data” vs. “Medium Data”

- Many companies only have “medium data”
- Effect of “bad data” is significant
  ミディアムデータでは不良データの影響は重大





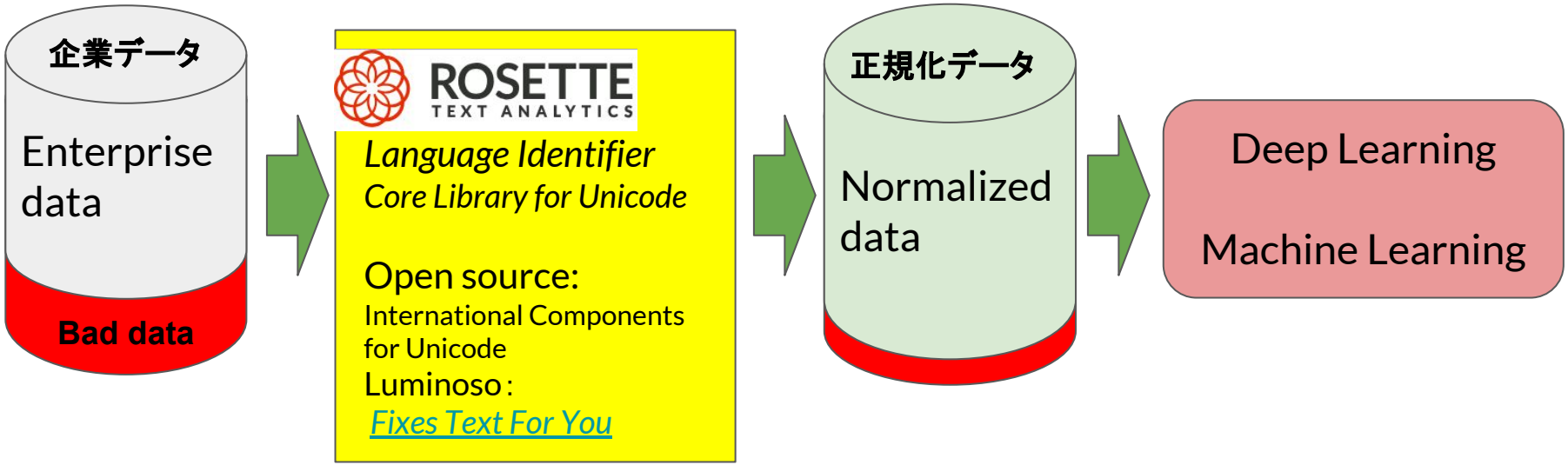
企業では
ミディアム
データ

Better DL Results With Cleaned Data

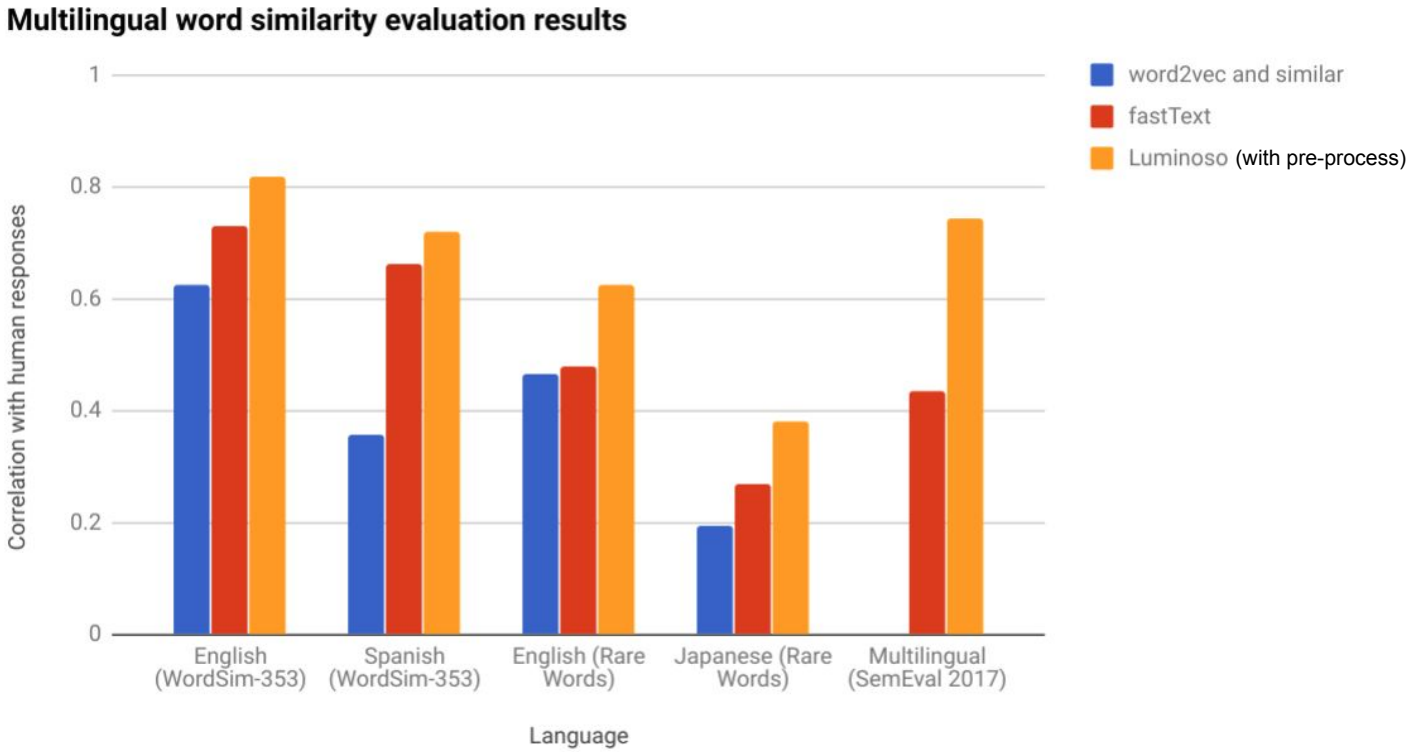
不良データを削減すればDLの結果も良くなる

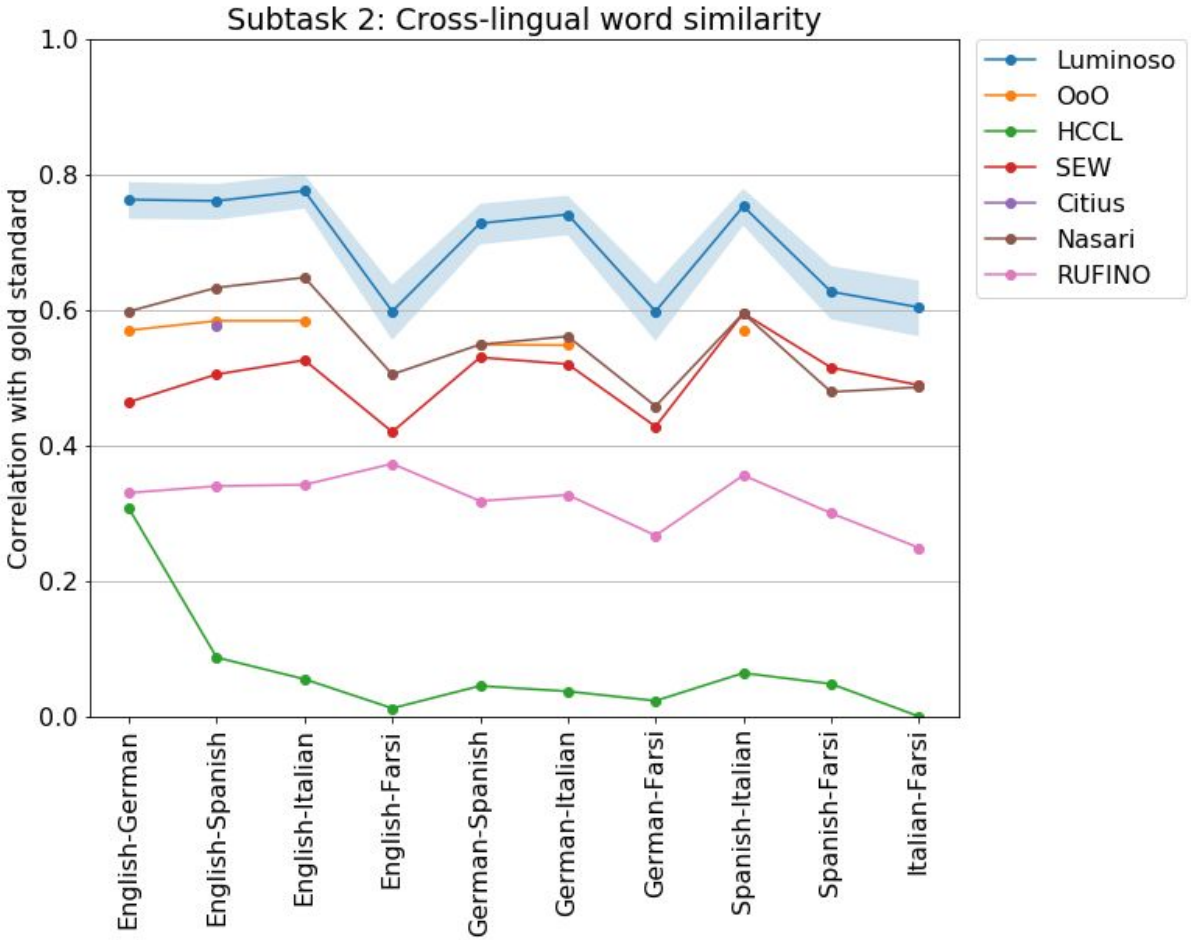
Include pre-process step to DL strategy

DL戦略には前処理ステップを含めよう



Luminoso Test Results





BASIS TECHNOLOGY

Luminoso with pre-processing does much better than the competition.

前処理機能を備えたLuminosoは競合よりもはるかに良好な結果になりました。

Thank you、ありがとうございました。

Go forth and
Deep Learn!

But don’t forget
to clean before
use.

さあ、もっとDLしましょう。

だだし、綺麗なデータで。

Data will continue to grow rapidly

データは増え続けている

Less than 10% of the data is being analyzed today.

現在分析されているデータは10%未満