

Generating Meaningful Business Descriptions for Sidecar

SIT Academy, 18.02.2022

**Dae-Jin Rhee and
Marlies Monch**



Introductions



sidecar



Dae-Jin Rhee

Data Scientist



Marlies Monch

Data Scientist

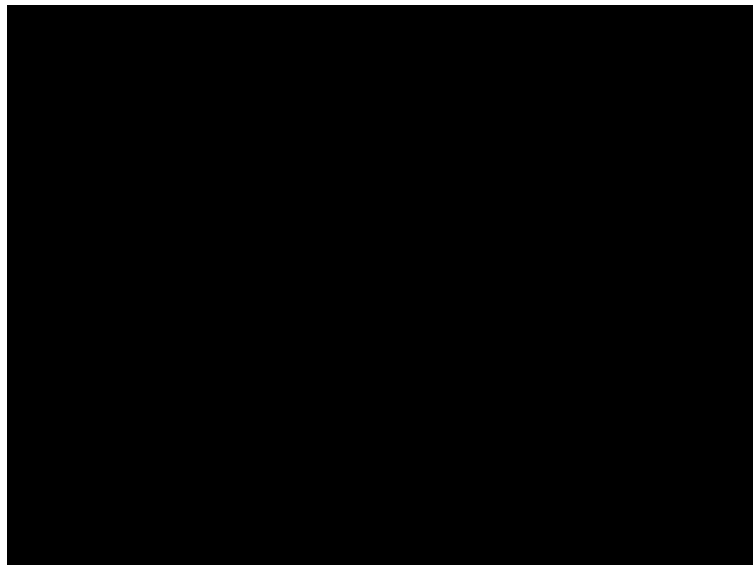
Background in Psychology

Business Objective



sidecar

Technical Name	"MOTHRM_MDN_NM"
Business Name	"Name before marriage"
Business Description	"The family name under which the mother was born. Applicable when ..."



Data Stewards time is costly

Use NLP for data entry?



Utilizing NLP for Data Entry? – Our Approach

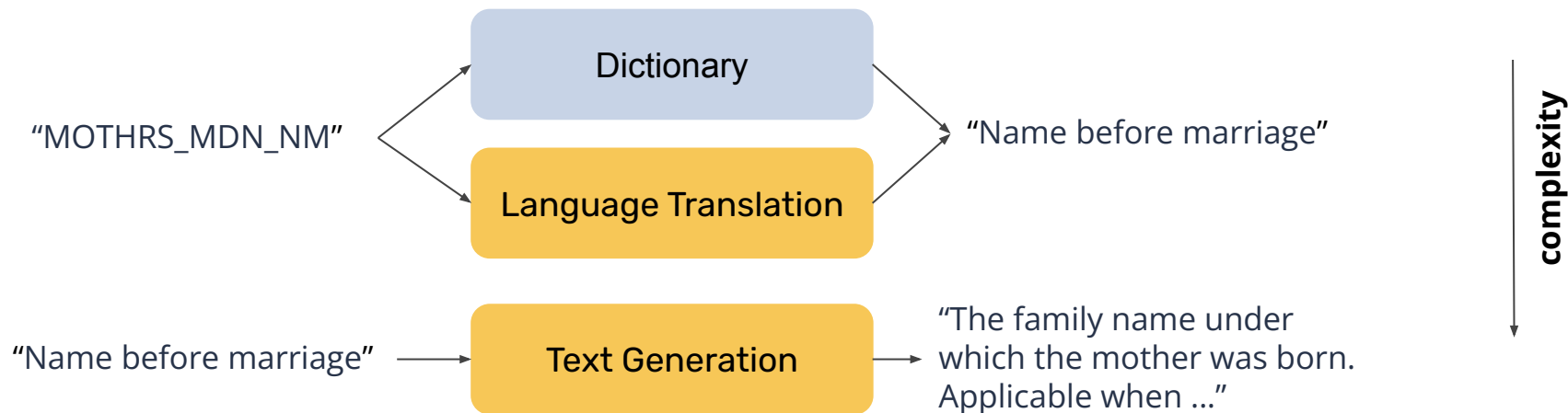


sidecar

Input

Method

Desired Result



Dictionary Similarity Approach



sidecar

Create a dictionary from
pre-processed data

Match Technical Name in the
dictionary

Output Business name

Technical Name	Business Name
id	id
gndr	gender
date of return	return date
unit price discount	discount amount
mothrs mdn nm	Name before marriage



Search term

mother



mothrs mdn nm

dictionary match



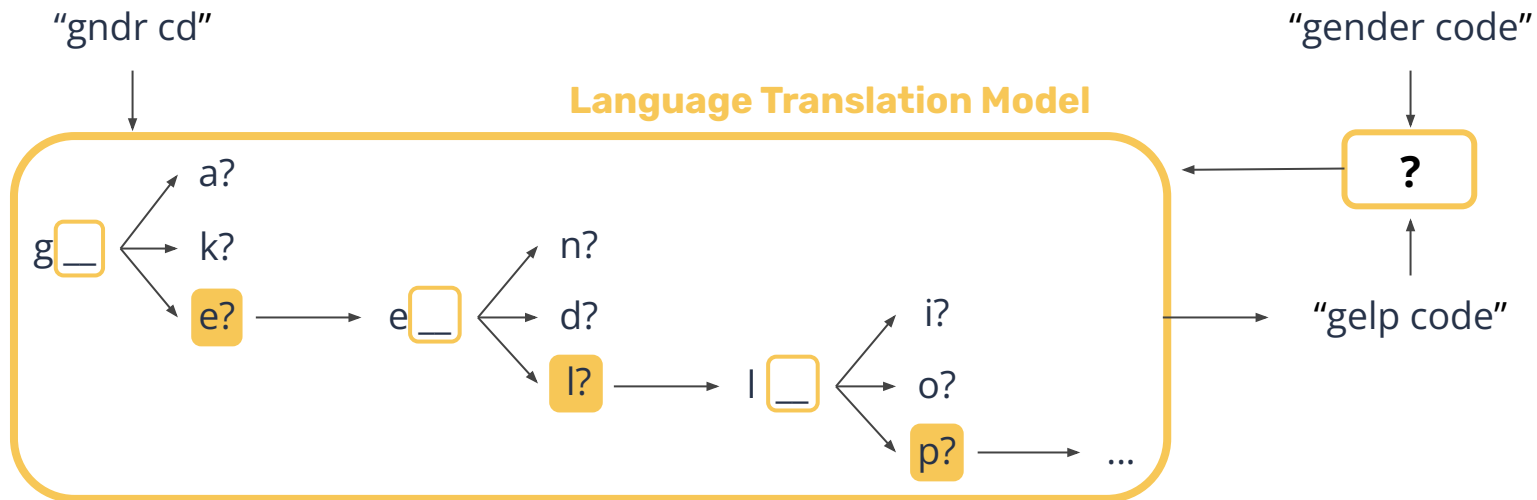
Business Name

Name before
marriage

How Does Language Translation Work?



sidecar

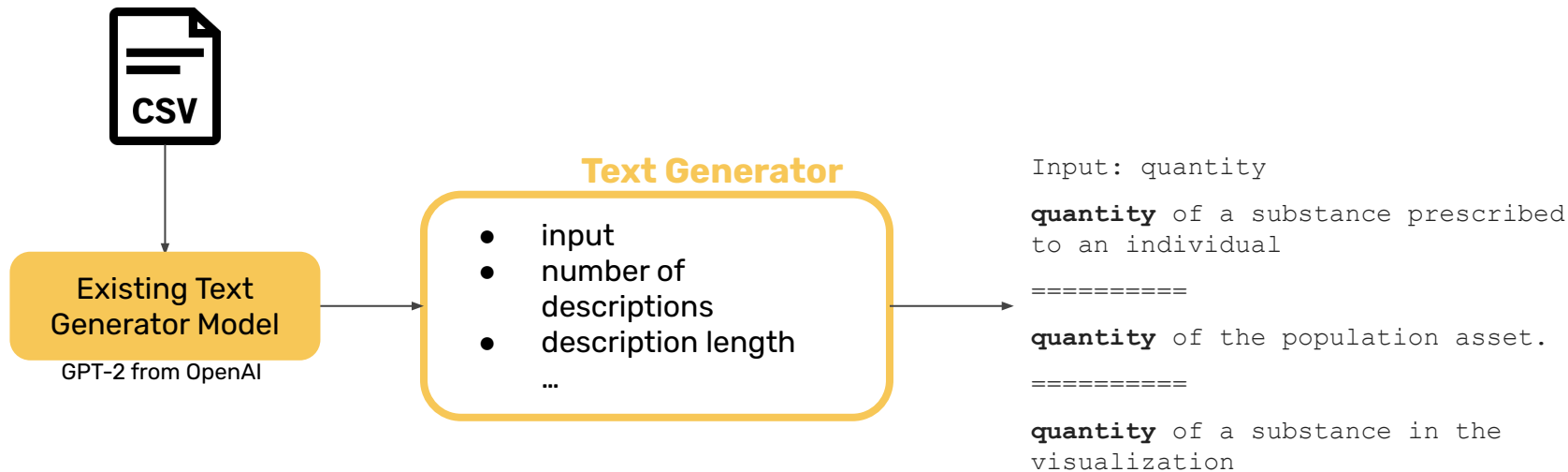


process continues until threshold for match is reached

How does Text Generation work?



sidecar



Using NLP for Data Entry – Conclusions



sidecar

- Translation model generates meaningful business labels
 - Quick and easy to implement
 - Improved performance over dictionary approach
- Generation of Business Descriptions → needs improvement
 - Better models available
 - Question of hardware and scope
- More Data improves performance
 - Utilize properties of the original datasets



NLP offers a promising approach to save resources for data entry!

Connect with us on LinkedIn



sidecar



Marlies Monch
Data Scientist



Dae-Jin Rhee
Data Scientist



Extras (not to be presented)



sidecar



1. Levenshtein distance

- Insertion: **baby** → **bae**by
- Deletion: **baby** → **bay**
- Substitution: **baby** → **buby**

2. Number of Characters

Levenshtein distance

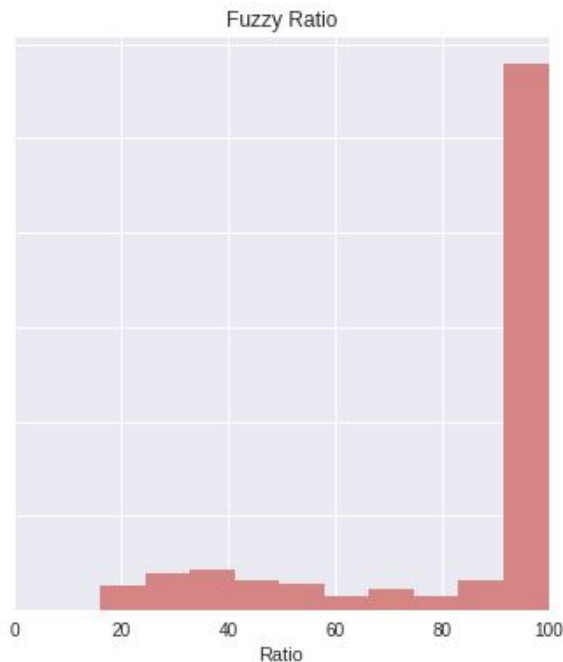
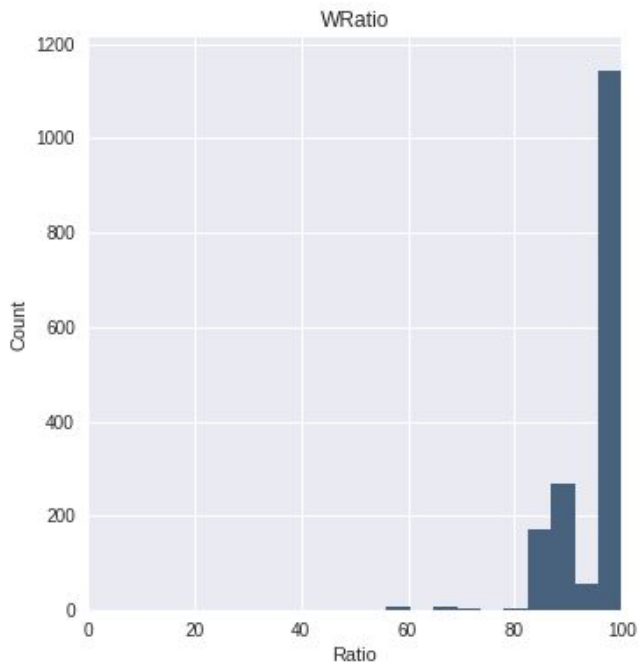
combined length of strings

Dictionary Similarity Results



sidecar

Fuzzy Ratios for Test Data

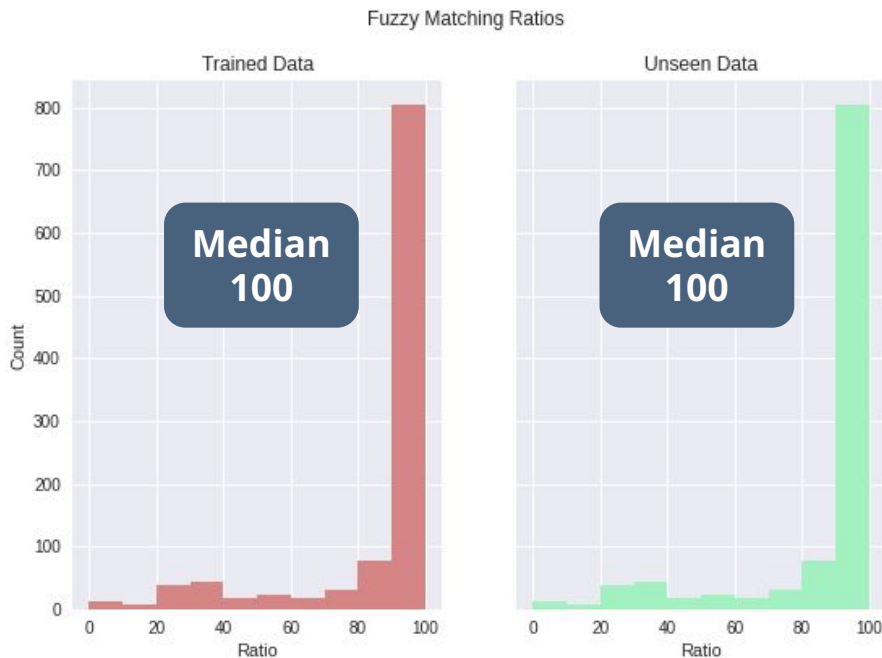


- WRatio as main value for matching
- Fuzzy Ratio/ Levenshtein as a secondary value.

Fuzzy Ratios for Sequence to Sequence Model



sidecar







Match Ratio from 0 to 100

Sidecar corporate colours



sidecar

	#48617cff
	#d68585ff
	#f7c757ff
	#a3f0c0ff

<https://www.sidecar-data.ch/>

Font:
Rubik
Rubik

Font:
Open Sans
Open Sans