DETECTION OF DUPLICATES AMONG NON-STRUCTURED DATA FROM DIFFERENT DATA SOURCES

SITUATION

When assessing a commercial risk, an insurer needs to gather various information about the risk. This long and complex process implies numerous questions. Thus an insurer is prompt to use an external source to help reduce the number of questions.

Thus, we want to detect duplicate of commercial risk in another data source using as **little** as possible information [Beauchemin and Lamontagne, 2020].

SITUATION

When assessing a commercial risk, an insurer needs to gather various information about the risk. This long and complex process implies numerous questions. Thus an insurer is prompt to use an external source to help reduce the number of questions.

Thus, we want to detect duplicate of commercial risk in another data source using as **little** as possible information [Beauchemin and Lamontagne, 2020].

Example

David Beauchemin, the owner of "Beauchemin inc.", calls for insurance. Using minimal information, we want to retrieve as much as possible from an external source to ask him as less than the necessary number of questions.

HOW DO WE DETECT DUPLICATE?

| Database 1 | → | Duplicate Detection Model | ← | Database 2 |

Laval University
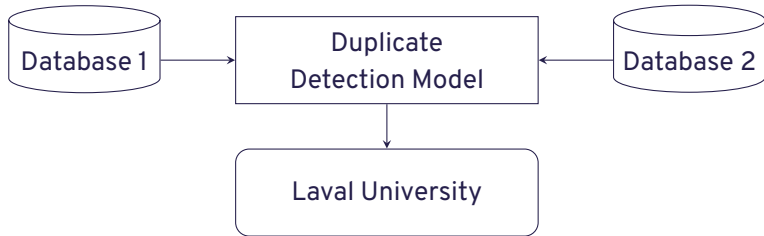
DATABASES

- *Registre des Entreprises du Québec* ($\sim$3.5 million entries)
    - Names
    - Addresses
    - Economic activities
    - Administrative informations
- **Private dataset** (21,444 enterprises)
    - Name
    - Address
    - Economic activity

PRIVATE DATASET AND THE REQ

11,649 commercial risk are in the province of Quebec, 1706 were annotated

- 1418 *(commercial risk, REQ entity)*
- 288 *(commercial risk, None)*.
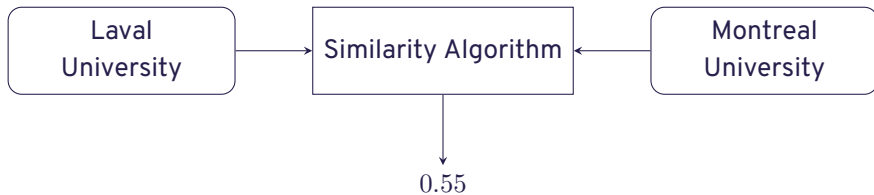
We only use the names and the addresses[1].

---

1. Addresses and names are normalize (not included in the presentation).

Similarity Between Two Entities

SIMILARITY ALGORITHM

SIMILARITY ALGORITHM

Similarities algorithms are one that measures the resemblance between two string base on the distance between their tokens.

NAME INTERESTING RESULTS

	StoS	Jaro-Winkler	Jaccard
Recall (%)	44.15	64.95	**65.87**

ADDRESS INTERESTING RESULTS

	StoS	Jaro-Winkler	CSS
Recall (%)	13.19	49.79	**50.36**

MISSING ADDRESSES

Since 16% (270) of the pair *(commercial risk, REQ entity)* are missing a address, the results are under-evaluated. For example, the CSS accuracy without those pairs is at 62.20% near 12% higher.

Those missing addresses are due to the confidentiality policy of the REQ [Registraire des entreprises du Québec, 2017].

SIMILARITY ALGORITHM SUMMARY

Using either name or address, the similarities algorithms allows us to find the matched entity around 50% of the time.
The leading approach been Jaccard using the name at near 67%.

MACHINE LEARNING SIMILARITY ALGORITHM
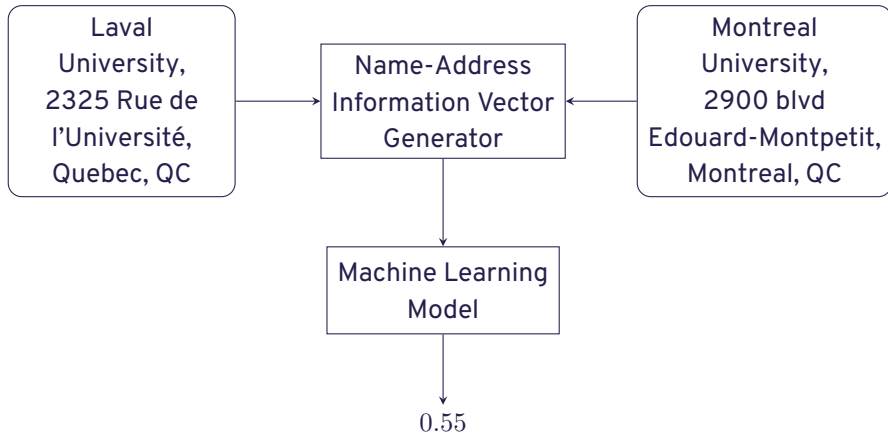
NAME-ADDRESS INFORMATION VECTOR GENERATOR

We used the previous similarity algorithm to generate an information vector between two entities using the name and address.

Example of an information vector

StoS	Levenshtein	Jaro-Winkler	LCSP	Jaccard	Cosinus	-
0.00	0.15	0.25	0.35	0.15	0.15	-

StoS	Levenshtein	Jaro	LCSP	Jaccard	Cosinus	CSS
0.00	0.16	0.55	0.15	0.45	0.37	0.48

RESULTS

	Logistic Regression	Random Forest	Multilayer perceptron	Jaccard
Recall (%)	66,67	73,54	**79,73**	72,51

IMPROVING THE RESULTS - *N* MOST SIMILAR

We consider a matching is good when the pair *(commercial risk, REQ entity)* is included in the *N* most similar.

RESULTS

Recall (%) by model (LR, RF, MP, Jaccard) for Top 1, Top 5, and Top 10.

Model	Top 1	Top 5	Top 10
LR	66.67	79.04	79.73
RF	73.54	81.44	83.85
MP	79.73	89.35	91.07
Jaccard	72.51	79.38	81.1
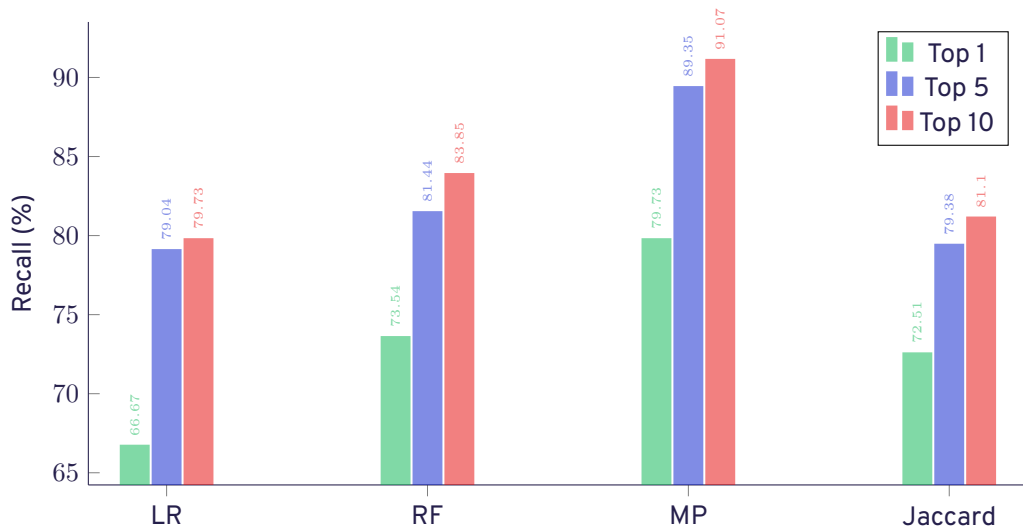
MACHINE LEARNING ALGORITHM SUMMARY

The used of name and address simultaneously with a machine learning algorithm improved the results.
Using a top N approach helps achieve better results when N is greater than 1.

FUTURE WORKS

- Word embeddings [Mikolov et al., 2013, Pennington et al., 2014, Wu et al., 2017, Foxcroft et al., 2019, Singh et al., 2019]
- Siamese Network [Godbole et al., 2018, Imtiaz et al., 2020]
- Uses of spatial data [Sehgal et al., 2006]
- Removal of more specitif stop words using a TF-IDF approach [Sammut and Webb, 2010]

CONCLUSION

- We have shown that using a similarity algorithm can achieve good results.
- Uses of machine learning algorithm (such as multilayer perceptron) can achieve greater results.
- Using a *N* most similar approach, where *N* is greater than one, help improved the results, achieving almost the max recall value.

ACKNOWLEDGMENTS

THANK YOU FOR LISTENING!

REFERENCES i

📄 Beauchemin, D. and Lamontagne, L. (2020).
Détection de doublons parmi des informations non structurées provenant de sources de données différentes.
Master's thesis, Université Laval.

📄 Foxcroft, J., d'Alessandro, A., and Antonie, L. (2019).
Name2vec : Personal names embeddings.
In *Advances in Artificial Intelligence*, pages 505–510.

📄 Godbole, A., Dalmia, A., and Sahu, S. K. (2018).
Siamese neural networks with random forest for detecting duplicate question pairs.

REFERENCES ii

Imtiaz, Z., Umer, M., Ahmad, M., Ullah, S., Choi, G. S., and Mehmood, A. (2020).
Duplicate Questions Pair Detection Using Siamese MaLSTM.
IEEE Access, 8 :21932–21942.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient Estimation of Word Representations in Vector Space.

Pennington, J., Socher, R., and Manning, C. (2014).
GloVe : Global vectors for word representation.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*,
pages 1532–1543.

REFERENCES iii

📄 Registraire des entreprises du Québec (2017).
Banque de données publique sur les entreprises au Québec - Guide d'intégration fonctionnelle.
Québec.

📄 Sammut, C. and Webb, G. I., editors (2010).
TF-IDF, pages 986–987.
Springer US.

📄 Sehgal, V., Getoor, L., and Viechnicki, P. D. (2006).
Entity resolution in geospatial data integration.
In *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems*, page 83–90. Association for Computing Machinery.

REFERENCES iv

Singh, L., Singh, S., Arora, S., and Borar, S. (2019).
One embedding to do them all.

Wu, L., Fisch, A., Chopra, S., Adams, K., Bordes, A., and Weston, J. (2017).
Starspace : Embed all the things !