Generic natural language distance via online semantic volumetric inference

Alex-Pauline Poudade, Pascal Rabier, Neau-Monier Sarah, Olivier Poudade, Grimault Valérie, Emmanuel Martins, Ludwig De Sousa

Artificial Intelligence Lab / Ministry of Foreign Affairs

With the generous help of Google and Alphabet Inc. and Benoît Dardelet.

November 8th, 2022, rev. 1.0

Abstract: This paper discusses the approach of creating semantic meaning ad hoc through direct explicit volumetric adherence or relative intersection, from online databases, such as Wikipedia or Google. We demonstrate this approach through use of correlation, between a dictionary index – a lexicon - and an import/export industry ISO A129 standard used by the Ministry of Finances, in the French language. We conclude, this approach by giving the most and least meaningful industrial results, for the French language. This questions whereas online apparent generic Natural language processing (NLP) pivot Chomsky Universal grammar (UG) representation, could inherit implicit initial national culture.

Keywords: Natural language processing (NLP), Universal grammar (UG), Noam Chomsky, online Querying, Set theory, Inclusion-Exclusion, Big Query, logarithmic scale comparison, Industry,

JEL Classifications Numbers: L60; C38; C52; C65; G21; G32; O33; A14; A19; A22; B55; C00; C01; C02; C10; C12; C19; C26; C45; H10; H19; Z30; Z31.

REFERENCES: doi.org/10.7910/DVN/WKLWF8

Introduction

In this paper, we present a novel way to create generic strong semantic inference, in the domain of Natural language processing (NLP), through use of Internet references.

We start with 2 sets of data, a natural language dictionary index lexicon of 14294 expanded entries - not requiring further grammar processing - and a category index lexicon comprising 76 international standardized industry import/export domains. The former dataset is referenced alternatively as S_1 or as $L382$, and the latter dataset is referenced as S_2 or as $A129$.

Deriving the two datasets S_1 and S_2, we obtain the set S_3:

$$S_3 = S_1 \otimes S_2$$

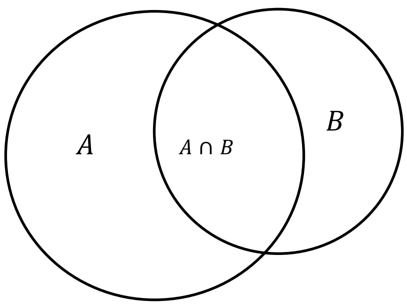
$$\left|\sum_{k=0}^{10844743} S_{3_k}\right| = \left|\left(\sum_{i=0}^{142695} S_{1_i}\right) * \left(\sum_{j=0}^{75} S_{2_j}\right)\right|$$

The S_3 set has cardinality of nearly 11 million elements.

Each element $S_{3_{i \otimes j}}$ of the 10844744 elements of S_3 belongs to a the semantic adherence subset $A129$ category.

Let A equal the set of returned results of internet references for an element of S_1, and let B equal the set returned results of internet references for an element of S_2. Both represent volumetric internet existence.

By the sets property of inclusion-exclusion, we obtain a semantic adherence S_3 :



$$\boxed{S_3 = Max\left(\frac{|A \cap B|}{|A \cup B|}\right) = Max\left(\frac{|A \cap B|}{|A| + |B| - |A \cap B|}\right) = Max\left(\frac{|A \cap B|}{|A + B|} - 1\right)}$$

The $A \cap B$ subset represents the set of returned results of internet references for an element of S_3: an element of both (S_1 AND S_2). We note the total theoric number T of references, required through Internet searches, in order to establish strong semantic natural language inference, is thus :

$$T = S_1 * S_2 + S_1 + S_2 = S_1 * (S_2 + 1) + S_2 = S_2 * (S_1 + 1) + S_1$$

And in our case, $T = |S_2| * (|S_1| + 1) = 76 * (142696 + 1) + 142696 = 10987668$ queries

Practicality

We create a special new element category ZZZZ for set S_2 when weak adherence:

$$k * S_3 < semantic\ threshold$$

To create a superset S_4 of $S_1$of 142694 common words infering one of 77 industry categories (including the added ZZZZ category), we reassemble 3 different datasets d_1, d_2 and d_3 from 3 different sources:

d_1 is the result of querying externally Google database with XMLHTTP client requests and calculating S_3 in runtime with $d_1 \subset L382$ / $|d_1|$ =25786.

d_2 is the result of querying externally Wikipedia database with Google Cloud BigQuery console and calculating S_3 post runtime with $d_2 \subset L382$ / $|d_2|$ =94987.

d_3 is the result of querying internally Google database with Google Cloud BigQuery console and calculating S_3 post runtime with $d_3 \subset L382$ / $|d_2|$ =94987.

Finally, we reconstruct 2 main datasets $D_1 \subset L382$ and $D_2 \subset L382$, labelled respectively $L382TOA129W$ and $L382TOA129G$ with both predominances, as so:

	$L382TOA129W$ *Inference subsets*	$L382TOA129W$ *Inference typing*	$L382TOA129G$ *Inference subsets*	$L382TOA129G$ *Inference typing*
	d_1 (18,07% of S_1)	$Max\left(\frac{\lvert A \cap B\rvert}{\lvert A + B\rvert} - 1\right)$	d_1 (18,07% of S_1)	$Max\left(\frac{\lvert A \cap B\rvert}{\lvert A + B\rvert} - 1\right)$
S_1	d_2 (66,56% of S_1)	$Max\left(\frac{\lvert A\cap B\rvert}{\lvert A+B\rvert} - 1\right)$ *or* $Max(A \cap B)$	d_3 (66,56% of S_1)	$Max\left(\frac{\lvert A\cap B\rvert}{\lvert A+B\rvert} - 1\right)$ *or* $Max(A \cap B)$
	unprocessed	*N/a*	*Unprocessed*	*N/a*
	120773 entries (84,5% of S_1)		*120773 entries (84,5% of S_1)*	

These 11 datasets, along with initial datasets are provided for the reader.

Categories used

Translated import/export A129[1] categories id;category_name;blk;blk2;blk3

A01Z;Products of cultivation and breeding	C14Z;Articles of clothing	C25B;Boilermaking products	C30B;Railway rolling stock
A02Z;Forestry products	C15Z;Leather, luggage and footwear	C25C;Arms and ammunition	C30C; Aeronautical and space construction products
A03Z;Fishery and aquaculture products	C16Z;Wood, articles of wood	C25E;Cutlery, tools, hardware and miscellaneous metal articles	C30D;Military combat vehicles
B05Z;Coal	C17A; Pulp, paper and paperboard	C26A;Components and electronic cards	C30E;Cycles and motorcycles
B06Z;Natural hydrocarbons	C17B;Paper or paperboard articles	C26B;Computers and peripheral equipment	C31Z;Furniture
B07Z;Metal Ores	C18Z;Printing and Reproducing Material	C26C;Telephones and communication equipment	C32A;Jewellery and jewellery, musical instruments
B08Z;Miscellaneous extractive industry products	C19Z;Refined petroleum products and coke	C26D;Consumer electronics	C32B;Instruments for medical, optical and dental purposes
C10A;Meat and meat products	C20A;Basic chemicals, nitrogen products, plastics and synthetic rubber	C26E; Measuring, testing and navigating apparatus and horological articles	C32C; Sporting goods, games and toys, miscellaneous manufactured goods
C10B;Prepared and preserved fish and fish products	C20B;Perfumes, cosmetics and cleaning products	C26F;electromedical equipment for diagnosis and treatment	D35A;electricity
C10C;Fruit and pulse products, including juices	C20C;Miscellaneous chemicals	C26G; Optical and photographic materials and magnetic and optical media	D35B;Manufactured gas
C10D;Vegetable and animal oils and fats, meal	C21Z;Pharmaceuticals	C27A;Household appliances	E37Z;Sewage sludge and household waste
C10E;Dairy and frozen products	C22A;Rubber Products	C27B;Electrical material	E38Z;Industrial waste
C10F;Products of grain processing and starch products	C22B;Plastic products	C28A;Machinery and equipment for general use	J58Z;Publishing products, software
C10G;Bakery and pastry products	C23A;Glass and glassware	C28B;Agricultural and forestry machinery	J59Z;Recorded CDs and DVDs
C10H;Miscellaneous food products	C23B;Construction materials and miscellaneous mineral products	C28C;Machine tools	M71Z;Plans and technical drawings
C10K;Animal feed	C24A;Steel and primary steel products	C28D;Miscellaneous machines for specific use	M74Z;Exposed photographic plates and films
C11Z;Beverages	C24B;Non-ferrous metals	C29A;Automotive products	R90Z;Paintings, engravings, sculptures
C12Z;Tobacco factory	C24C;Foundry Products	C29B;automotive equipment	R91Z;Antiques and collectibles
C13Z;Products of the textile industry	C25A;metal elements for construction	C30A;Ships and Boats	S96Z;Raw Hair

[1] cf. https://lekiosque.finances.gouv.fr/fichiers/guide/Table_AGREG.pdf

Observed result

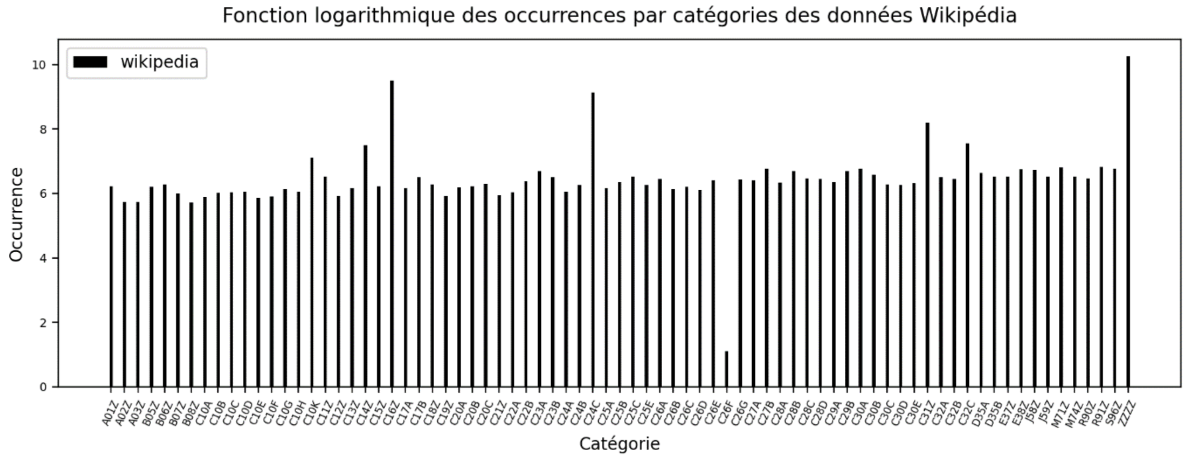


Figure 1. Logarithmic sparse Wikipedia data occurrence for specific 77 industrial categories

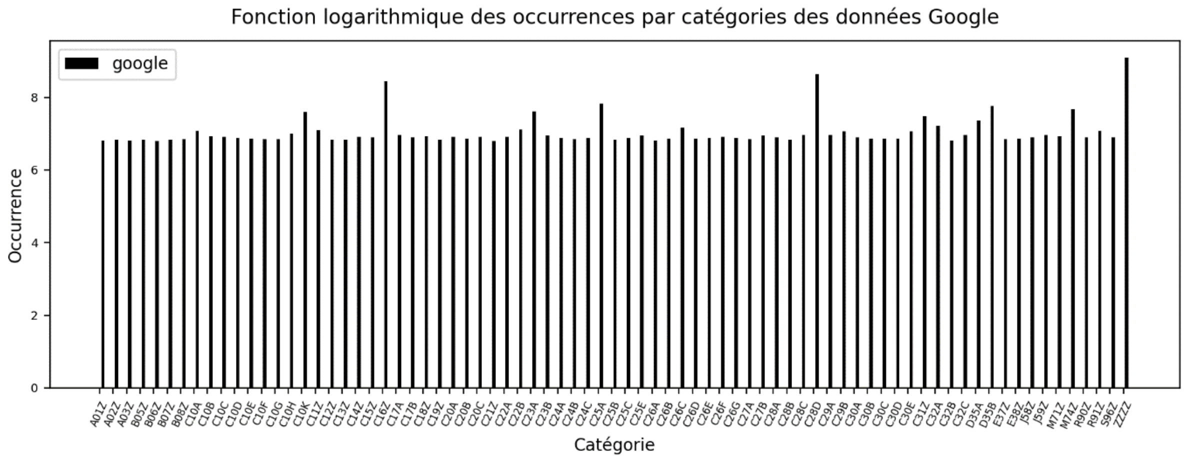


Figure 2. Logarithmic sparse Google data occurrence for specific 77 industrial categories

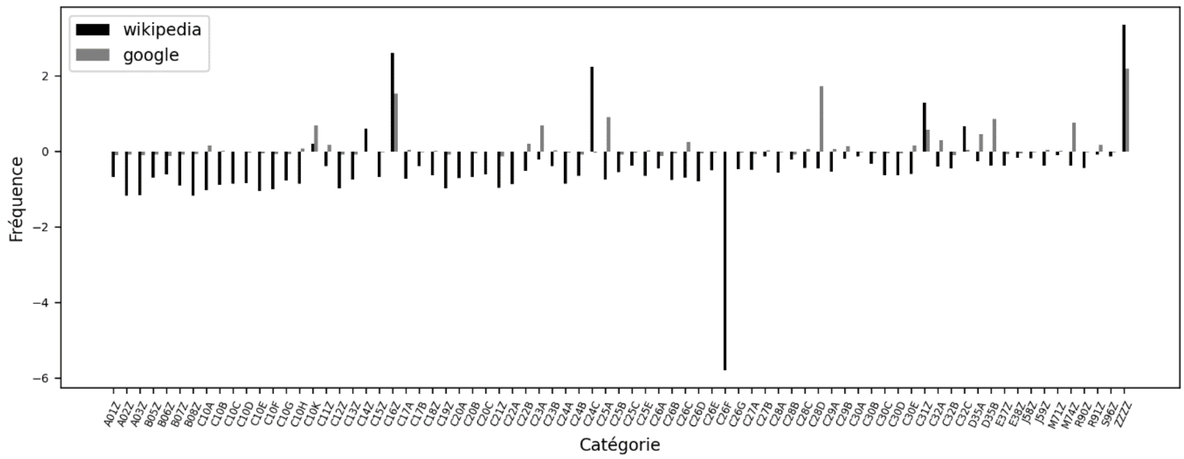


Figure 3. Raw comparison of sparse Wikipedia and Google data frequency for specific 77 industrial categories

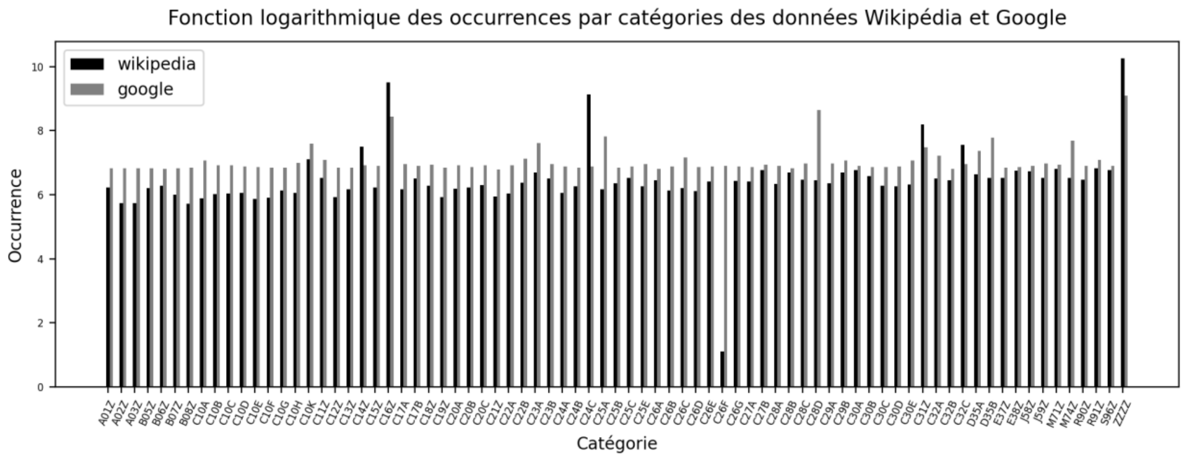


Figure 4. Logarithmic comparison of sparse Wikipedia and Google data occurrence for specific 77 industrial categories

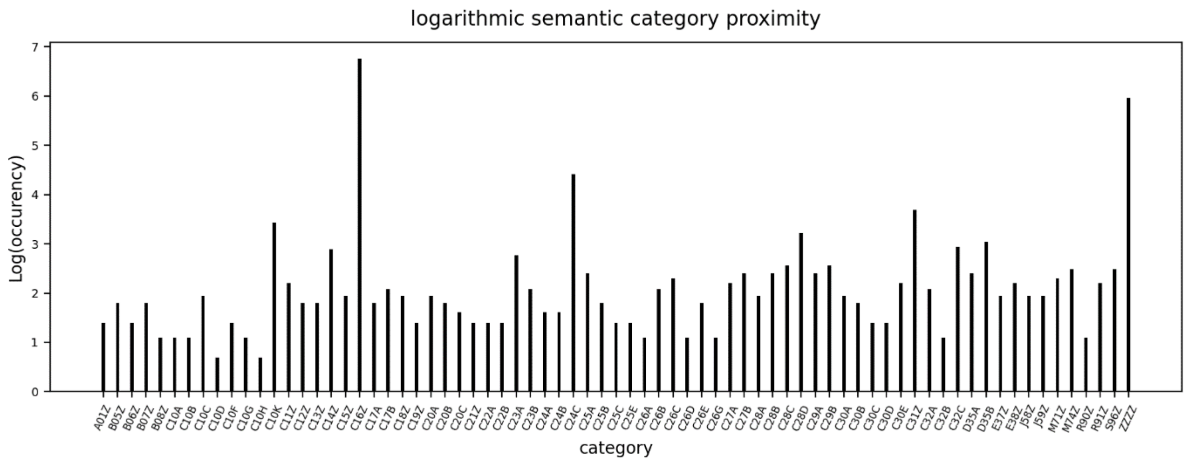


Figure 5. Logarithmic merged sparse Wikipedia and Google data occurrence for specific 77 industrial categories

Five most important semantic hyper-adherence industrial import/export categories bearing online over-significance for the French language	
1	Wood, articles of wood
2	Foundry Products
3	Animal feed
4	Furniture
5	Agricultural and forestry machinery

Five most important semantic hypo-adherence industrial import/export categories bearing online under-significance for the French language	
1	Vegetable and animal oils and fats, meal
2	Miscellaneous food products
3	Components and electronic cards
4	Consumer electronics
5	Optical and photographic materials and magnetic and optical media

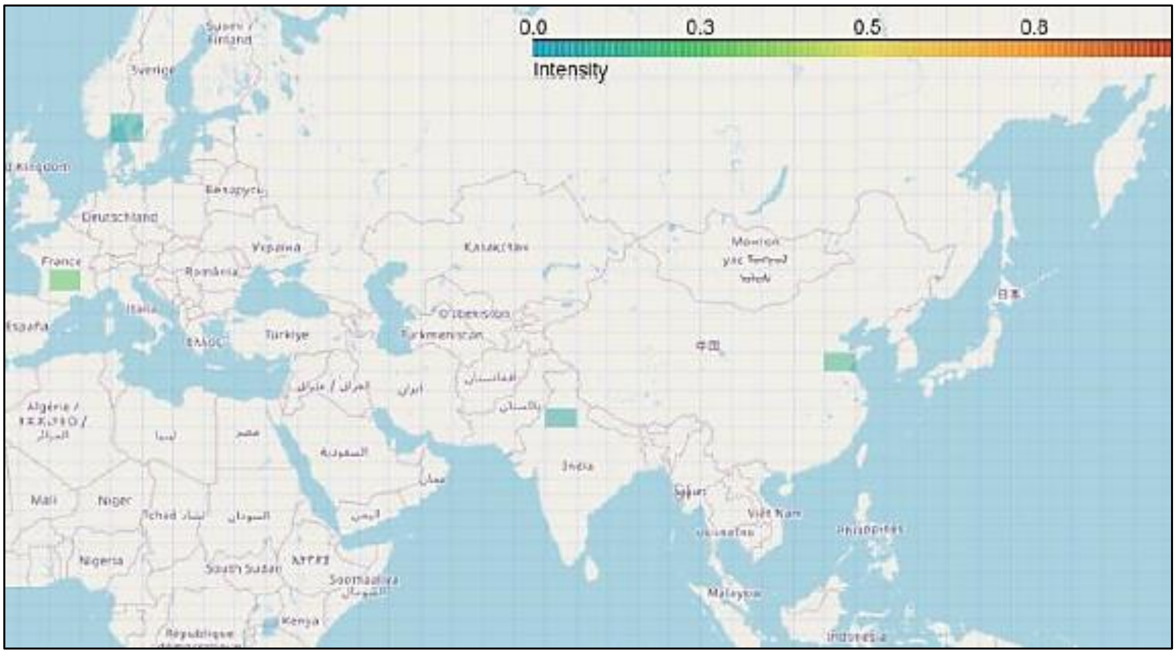


Figure 7. Tool displaying 5 varying regional intensities for specific industrial category during specific timespan interval

<XML>

<?xml version='1.0' encoding='UTF-8'?><xml><records><record><ref-type name="Dataset">59</ref-type><contributors><authors><author>Poudade, Alex-Pauline & Al.</author></authors></contributors><titles><title>Generic natural language distance via online semantic volumetric inference</title></titles><section>2022-11-18</section><dates><year>2022</year></dates><edition>DRAFT VERSION</edition><keywords><keyword>Natural language processing (NLP), Universal grammar (UG), Noam Chomsky, online Querying, Set theory, Inclusion-Exclusion, Big Query, logarithmic scale comparison, Industry</keyword></keywords><publisher>Harvard Dataverse</publisher><urls><related-urls><url>https://doi.org/10.7910/DVN/WKLWF8</url></related-urls></urls><electronic-resource-num>doi/10.7910/DVN/WKLWF8</electronic-resource-num></record></records></xml>

<RIS>

Provider: Harvard Dataverse
Content: text/plain; charset="utf-8"
TY - DATA
T1 - Generic natural language distance via online semantic volumetric inference
AU - Poudade, Alex-Pauline & Al.
DO - doi:10.7910/DVN/WKLWF8
ET - DRAFT VERSION
KW - Natural language processing (NLP), Universal grammar (UG), Noam Chomsky, online Querying, Set theory, Inclusion-Exclusion, Big Query, logarithmic scale comparison, Industry
PY - 2022
SE - Fri Nov 18 11:38:23 EST 2022
UR - https://doi.org/10.7910/DVN/WKLWF8
PB - Harvard Dataverse
ER -

<BibTeX>

@data{DVN/WKLWF8_2022,
author = {Poudade, Alex-Pauline & Al.},
publisher = {Harvard Dataverse},
title = {{Generic natural language distance via online semantic volumetric inference}},
year = {2022},
version = {DRAFT VERSION},
doi = {10.7910/DVN/WKLWF8},
url = {https://doi.org/10.7910/DVN/WKLWF8}
}