onepoint.
beyond the obvious

CUTIE : apprendre à comprendre les documents non-structurés

3 février 2023

Au menu

- Données visuellement riches
- CUTIE
- Base de connaissance
- Causalité

QUELLES DONNÉES ?

Les factures, des données visuellement riche

Deux sources d'informations :

- ✓ Le texte (information sémantique) ;
- ✓ L'organisation générale du document (information spatiale) ;

Empfangsschein

Konto / Zahlbar an
CH12 3456 7890 1234 5678 9
Steuerverwaltung der Stadt Bern
Bundesgasse 33
3011 Bern

Referenz
12 34567 89012 34567 89012 34567

Zahlbar durch
Martina Muster
Bubenbergplatz 1
3011 Bern

Währung CHF **Betrag** 2 500.00

Annahmestelle

Zahlteil

Konto / Zahlbar an
CH12 3456 7890 1234 5678 9
Steuerverwaltung der Stadt Bern
Bundesgasse 33
3011 Bern

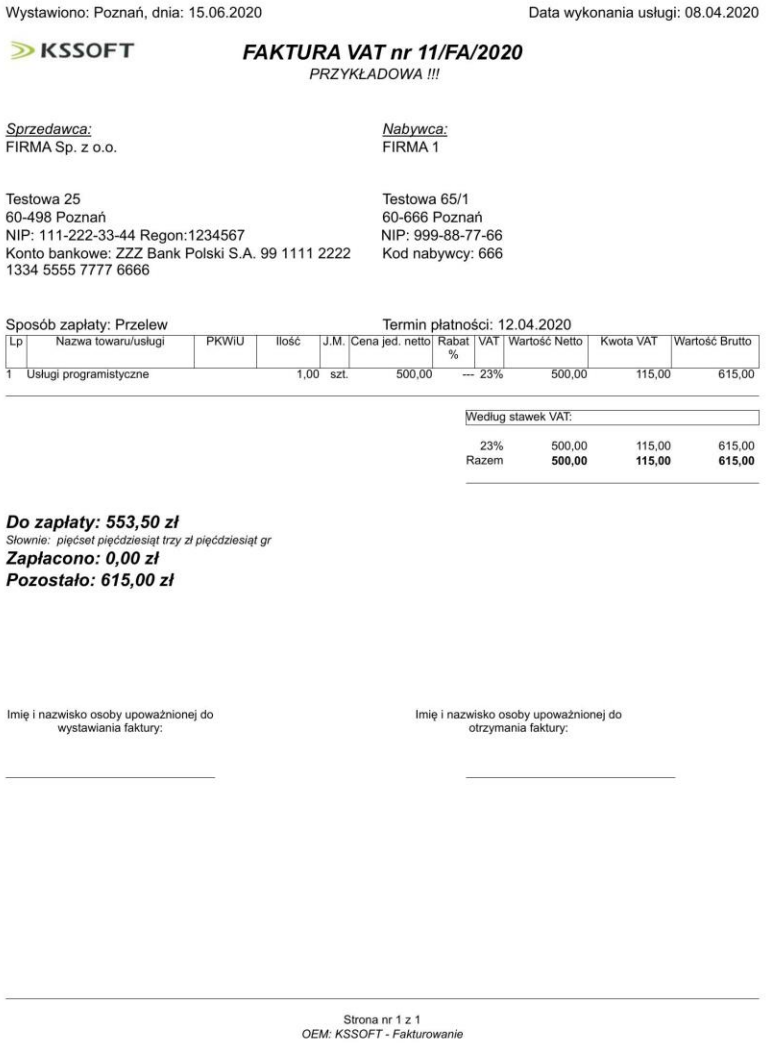
Referenz
12 34567 89012 34567 89012 34567

Zusätzliche Informationen
1. Steuerrate 2020
//S1/11/200627/30/115140892/31/200627/32/7.7/40/0:30

Zahlbar durch
Martina Muster
Bubenbergplatz 1
3011 Bern

Währung CHF **Betrag** 2 500.00

https://commons.wikimedia.org/wiki/File:Beispiel_QR-Rechnung.png



Wystawiono: Poznań, dnia: 15.06.2020 Data wykonania usługi: 08.04.2020

KSSOFT

FAKTURA VAT nr 11/FA/2020
PRZYKŁADOWA !!!

Sprzedawca:
FIRMA Sp. z o.o.

Testowa 25
60-498 Poznań
NIP: 111-222-33-44 Regon:1234567
Konto bankowe: ZZZ Bank Polski S.A. 99 1111 2222 1334 5555 7777 6666

Nabywca:
FIRMA 1

Testowa 65/1
60-666 Poznań
NIP: 999-88-77-66
Kod nabywcy: 666

Sposób zapłaty: Przelew Termin płatności: 12.04.2020

Lp	Nazwa towaru/usługi	PKWiU	Ilość	J.M.	Cena jed. netto	Rabat %	VAT	Wartość Netto	Kwota VAT	Wartość Brutto
1	Usługi programistyczne		1,00	szt.	500,00	---	23%	500,00	115,00	615,00

Według stawek VAT:

23%	500,00	115,00	615,00
Razem	**500,00**	**115,00**	**615,00**

Do zapłaty: 553,50 zł
Słownie: pięćset pięćdziesiąt trzy zł pięćdziesiąt gr
Zapłacono: 0,00 zł
Pozostało: 615,00 zł

Imię i nazwisko osoby upoważnionej do wystawiania faktury:

Imię i nazwisko osoby upoważnionej do otrzymania faktury:

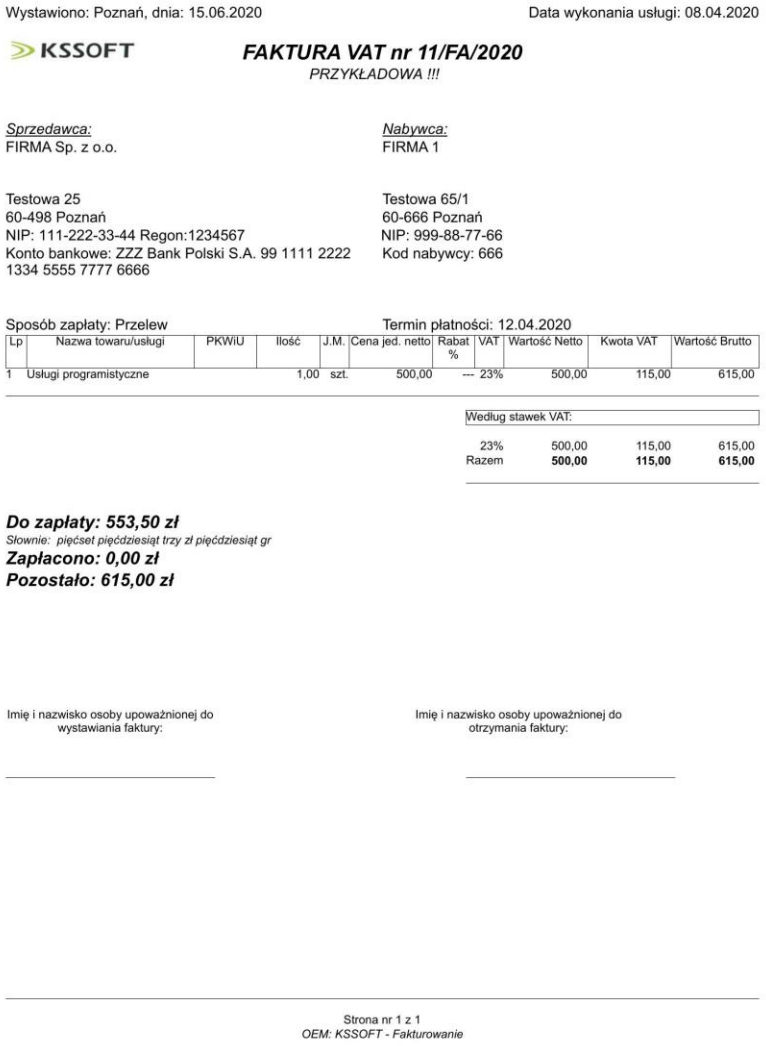
Strona nr 1 z 1
OEM: KSSOFT - Fakturowanie

https://github.com/vsymbol/CUTIE

LES FACTURES SONT-ELLES DES DONNÉES NON STRUCTURÉES ?

Donnée structurée : suit un format standardisé et une structure précise.

- Selon la définition donnée précédemment les factures sont bien des données non structurées
- La norme Facture-X va profondément changer les choses, puisqu'elle va associer l'image de la facture avec ses données structurées
- 1er juillet 2024 : passage à la facturation électronique se fera progressivement (jusqu'au 1er janvier 2026, en fonction des différentes tailles d'entreprises)
- Gros enjeux pour passer de documents d'un format non-structuré à un format structuré



Wystawiono: Poznań, dnia: 15.06.2020 — Data wykonania usługi: 08.04.2020

KSSOFT

FAKTURA VAT nr 11/FA/2020

PRZYKŁADOWA !!!

Sprzedawca:
FIRMA Sp. z o.o.

Testowa 25
60-498 Poznań
NIP: 111-222-33-44 Regon:1234567
Konto bankowe: ZZZ Bank Polski S.A. 99 1111 2222 1334 5555 7777 6666

Nabywca:
FIRMA 1

Testowa 65/1
60-666 Poznań
NIP: 999-88-77-66
Kod nabywcy: 666

Sposób zapłaty: Przelew — Termin płatności: 12.04.2020

Lp	Nazwa towaru/usługi	PKWiU	Ilość	J.M.	Cena jed. netto	Rabat %	VAT	Wartość Netto	Kwota VAT	Wartość Brutto
1	Usługi programistyczne		1,00	szt.	500,00	---	23%	500,00	115,00	615,00

Według stawek VAT:

23%	500,00	115,00	615,00
Razem	**500,00**	**115,00**	**615,00**

Do zapłaty: 553,50 zł
Słownie: pięćset pięćdziesiąt trzy zł pięćdziesiąt gr
Zapłacono: 0,00 zł
Pozostało: 615,00 zł

Imię i nazwisko osoby upoważnionej do wystawiania faktury:

Imię i nazwisko osoby upoważnionej do otrzymania faktury:

Strona nr 1 z 1
OEM: KSSOFT - Fakturowanie

QUELLES DIFFICULTÉS ?

Liées aux données

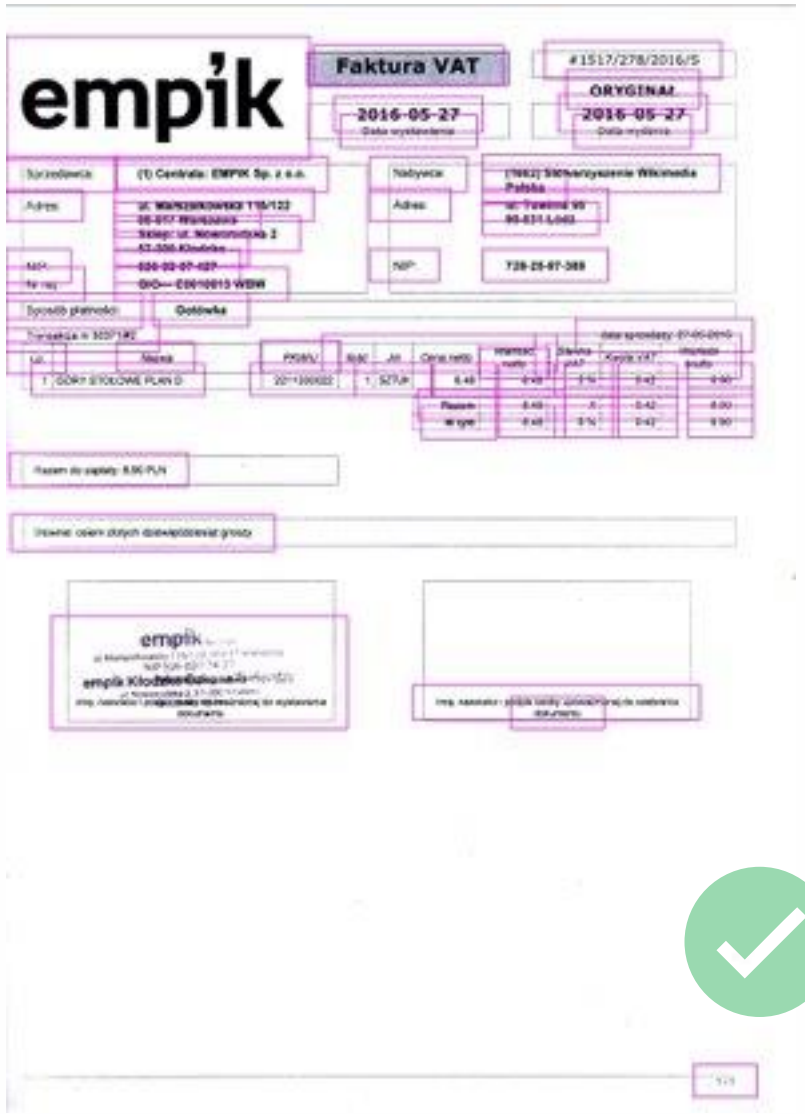
- ✓ L'organisation générale du document varie d'un fournisseur à un autre ou au fil du temps pour un même fournisseur ;
- ✓ Mais pour un même fournisseur plusieurs factures peuvent être extrêmement similaires, y compris pour des éléments sans pertinence.
- ✓ Le fait de scanner ou de photographier un document peut en modifier significativement l'organisation générale ;
- ✓ Identifier à quoi correspond un élément de la facture peut demander de prendre en compte du contenu éloignée spatialement ;
- ✓ Le vocabulaire spécifique est relativement restreint, mais le vocabulaire parasite peut lui être très varié (exemple facture fournisseur énergie VS facture libraire VS facture matériel agricole).
- ✓ Les factures sont des données personnelles dont l'utilisation soulève de nombreuses problématiques juridiques
- ✓ Obtenir de la vérité terrain est laborieux et coûteux

Liées aux modèles

- ✓ Initialement beaucoup de traitements réalisés sur la base de règles métiers ou de templates => difficilement généralisable, demande un effort conséquent pour être configuré de manière optimale
- ✓ Lorsque le texte n'est pas parfaitement aligné, les méthodes de type NER échouent dans la majorité des cas ;
- ✓ Entraîner une méthode avec des couches d'attention nécessitent de très nombreuses données.

PRÉ-TRAITEMENT

```
Empfangsschein

(CHI2 3456 7890 1234 5678 9
‘Steuerverwaitung der Stadt Bern
Bundesgasse 33

3011 Bein

Reteens
12 34567 89012 24567 89012 24567

‘Martina Muster

Bubenbergplatz 1
3011 Bern

CHE 2,500.00

Konto | Zahibar an
CH12 3456 7890 1234 5678 9
Steuerverwaltung der Stadt Bern
Bundesgasse 33

3011 Bern

Referenz
12 34567 89012 34567 89012 34567

Zusiitzliche Informationen
1. Steuerrate 2020
1S1/11/200627/30/115140892/31/200627/32/   7/40/
30

Zahlbar durch
Martina Muster
Bubenbergplatz 1
3011 Bern
```

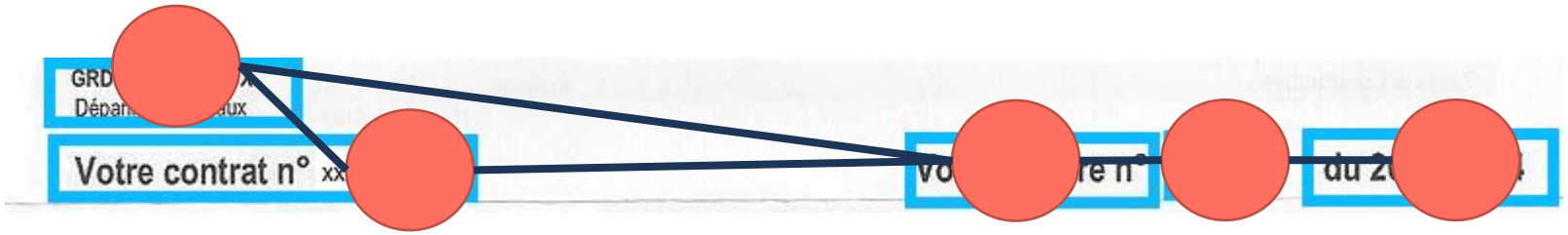



```
"text_boxes": [
    {
        "id": 0,
        "bbox": [
            0.0,
            0.0295,
            0.3854,
            0.1407
        ],
        "text": "empik\n",
        "bbox_type": "CONTENT"
    },
    {
        "id": 1,
        "bbox": [
            0.6821,
            0.0338,
            0.9059,
            0.0676
        ],
        "text": "#1517/278/2016/S\n\n",
        "bbox_type": "CONTENT"
    },
    {
        "id": 2,
        "bbox": [
            0.3902,
            0.0372,
            0.6291,
            0.0731
        ],
        "text": "Faktura VAT\n",
        "bbox_type": "CONTENT"
    },
```

On découpe le texte en **boîtes englobantes**

STRUCTURE DES DONNÉES

Deux boîtes englobantes sont considérées comme voisine si le segment connectant leurs barycentres n'intersecte aucune autre boîte englobante



Déterminer la catégorie sémantique de chaque boîte englobante

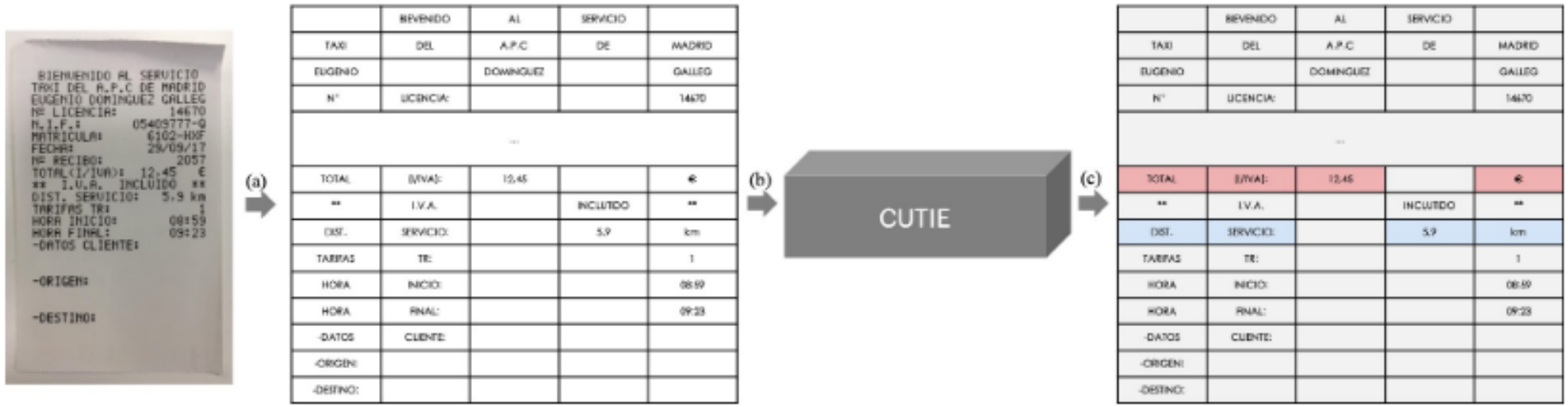
- X : contenu textuel de la boîte englobante
- Y : catégorie sémantique (fournisseur, client, TVA, date échéance, etc.)

(X,Y) peuvent être considéré comme un champs aléatoire conditionnel

Definition. *Let* $G = (V, E)$ *be a graph such that* $\mathbf{Y} = (\mathbf{Y}_v)_{v \in V}$, *so that* $\mathbf{Y}$ *is indexed by the vertices of* G. *Then* $(\mathbf{X}, \mathbf{Y})$ *is a conditional random field in case, when conditioned on* $\mathbf{X}$, *the random variables* $\mathbf{Y}_v$ *obey the Markov property with respect to the graph:* $p(\mathbf{Y}_v \,|\, \mathbf{X}, \mathbf{Y}_w, w \neq v) = p(\mathbf{Y}_v \,|\, \mathbf{X}, \mathbf{Y}_w, w \sim v)$, *where* $w \sim v$ *means that* w *and* v *are neighbors in* G.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

STRUCTURE DES DONNÉES



Transformation des boîtes englobantes en grille (graphe régulier, avec 4 ou 8 voisinage

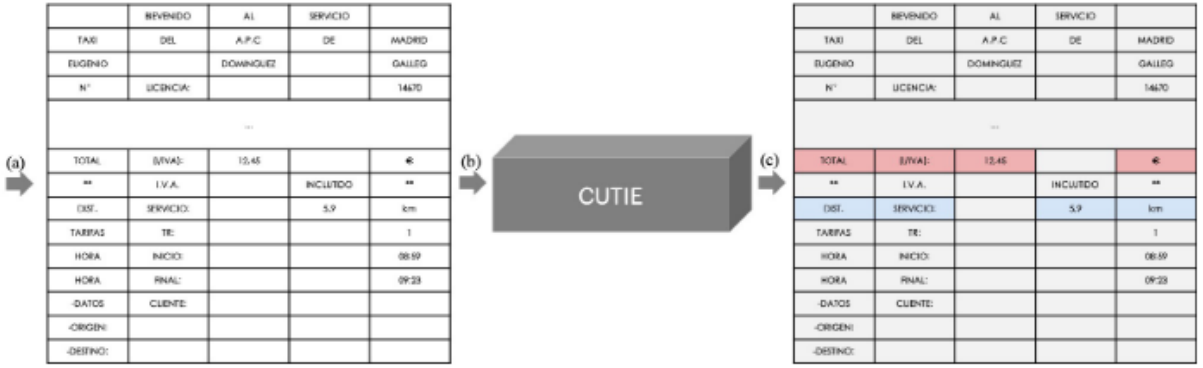
- Les nœuds de la grille correspondent à des tokens
- X : le contenu textuel du token
- Y : la catégorie sémantique **du token**

Cette modification du problème permet une attribution beaucoup plus précise du token (par exemple en autorisant l'identification des différents composants d'une adresse).

Il devient par contre plus délicat de partir sur l'hypothèse d'un champs aléatoire conditionnel.

CUTIE

Convolutional Universal Text Information Extractor



2022

Wang, W., Huang, Z., Luo, B., Chen, Q., Peng, Q., Pan, Y., ... & Zhang, Y. (2022). ERNIE-mmLayout: Multi-grained MultiModal Transformer for Document Understanding. *arXiv preprint arXiv:2209.08569*.

Lee, C. Y., Li, C. L., Dozat, T., Perot, V., Su, G., Hua, N., ... & Pfister, T. (2022). FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction. *arXiv preprint arXiv:2203.08411*.

2021

Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, *9*, 72894-72936.

Zhang, P., Xu, Y., Cheng, Z., Pu, S., Lu, J., Qiao, L., ... & Wu, F. (2020, October). TRIE: end-to-end text reading and information extraction for document understanding. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 1413-1422).

2020

Majumder, B. P., Potti, N., Tata, S., Wendt, J. B., Zhao, Q., & Najork, M. (2020, July). Representation learning for information extraction from form-like documents. In *proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 6495-6504).

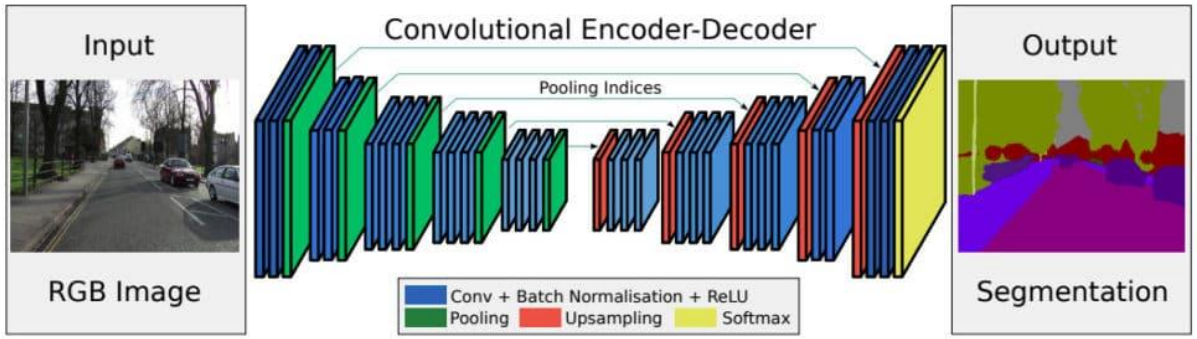
Denk, T. I., & Reisswig, C. (2019). Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

2019

Zhao, X., Niu, E., Wu, Z., & Wang, X. (2019). Cutie: Learning to understand documents with convolutional universal text information extractor. *arXiv preprint arXiv:1903.12363*.

Proposition très orientée technique, sans validation par les paires mais reprise dans en une quarantaine de travaux.

Inspiré par les travaux autour de la segmentation sémantique, notamment les approches encoder-decoder.



https://cnvrg.io/semantic-segmentation/

TRANSFORMATION DU TEXTE EN GRILLE

$$c_x^i = c_{gm} \frac{x_{left} + \frac{x_{right} - x_{left}}{2}}{w}$$

$$r_x^i = r_{gm} \frac{y_{top} + \frac{y_{bottom} - y_{top}}{2}}{h}$$

(c_x^i, r_y^i)	Coordonnées du texte n °1
(c_{gm}, r_{gm})	Taille de la grille
(w, h)	Taille du document
(x_{left}, y_{top}) (x_{right}, y_{bottom})	Coordonnées de la boîte englobante (coin haut gauche et coin bas droite)

Tokenisation :

1. Tokenisation classique (ponctuation et espaces) ;
2. Tokenisation Wordpiece
   - Cherche le token de plus grande taille à partir de la première lettre
   - Répète le processus avec ce qui reste du mot

Lexique :

- Conçu à partir des données d'entraînement
- Multi-alphabet
- Vocabulaire spécifique mais pas que

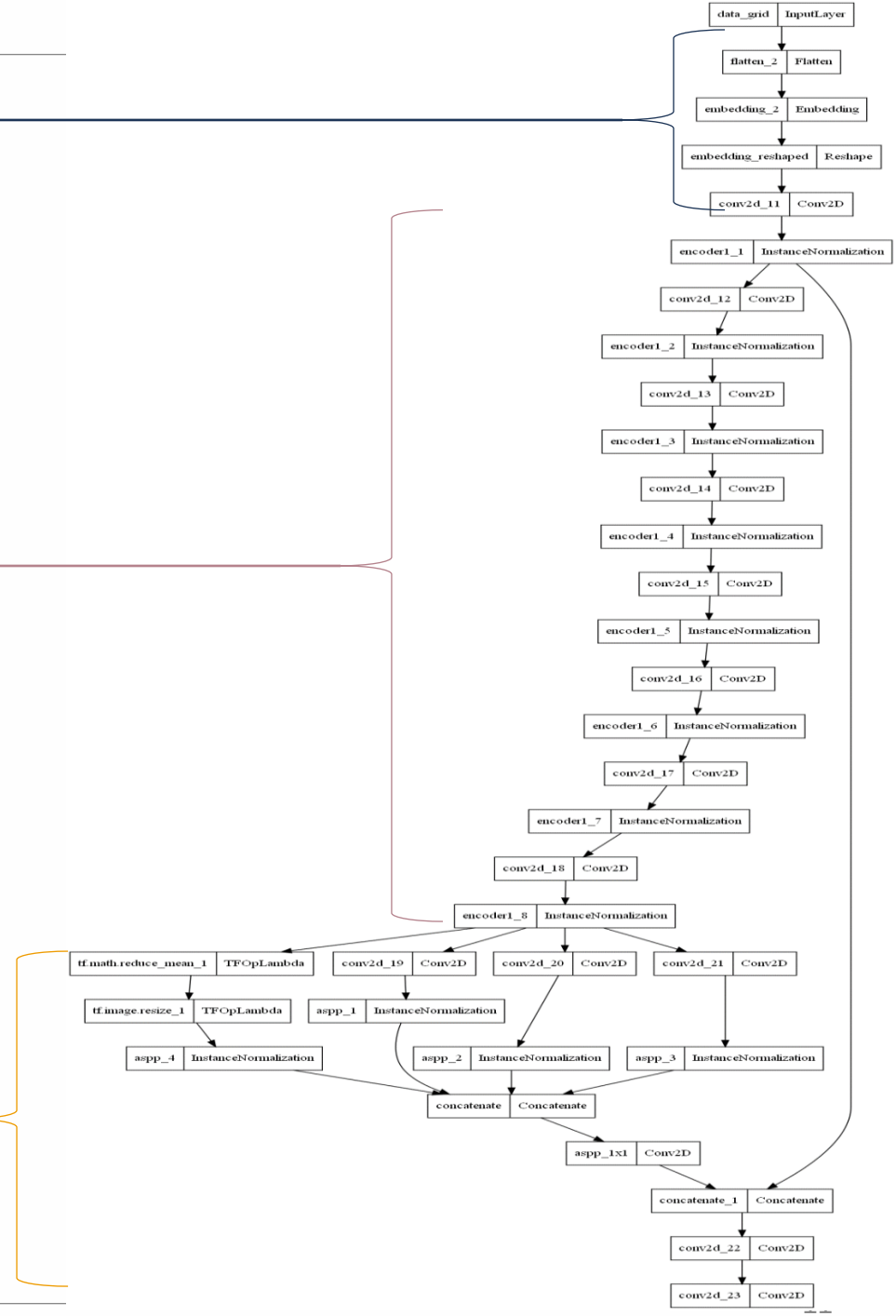
```
taxes
scientists
Sword
сезони
Bedford
##czyna
##Gш
Lakin
```

ARCHITECTURE CUTIE

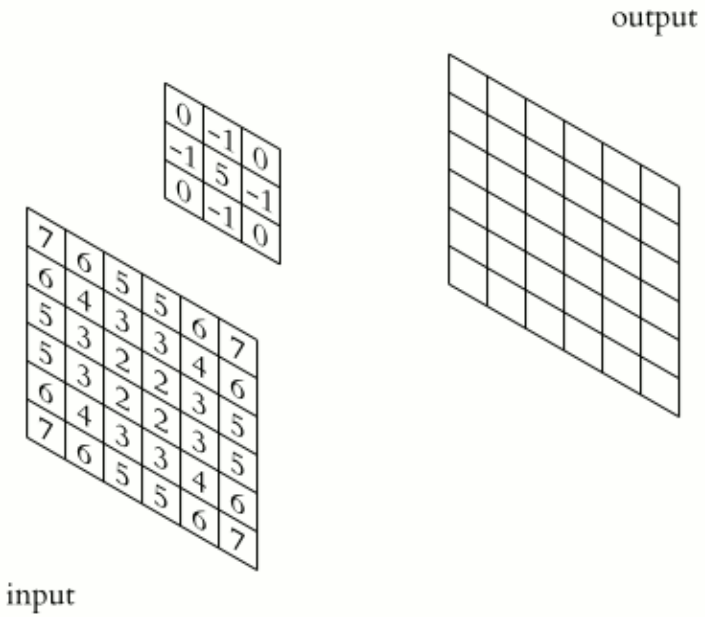
embedding

encoder
(convolutions et
atrous convolutions)

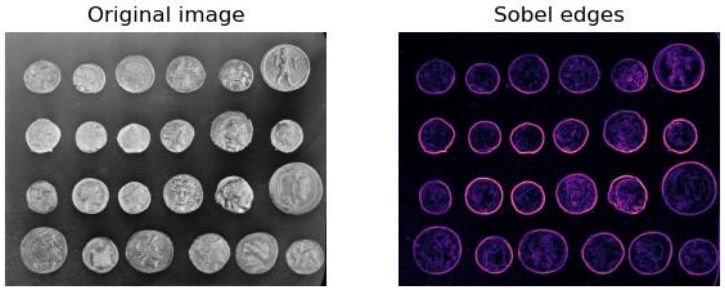
decoder
(concaténation
atrous convolutions)



ARCHITECTURE CUTIE



https://en.wikipedia.org/wiki/Kernel_(image_processing)

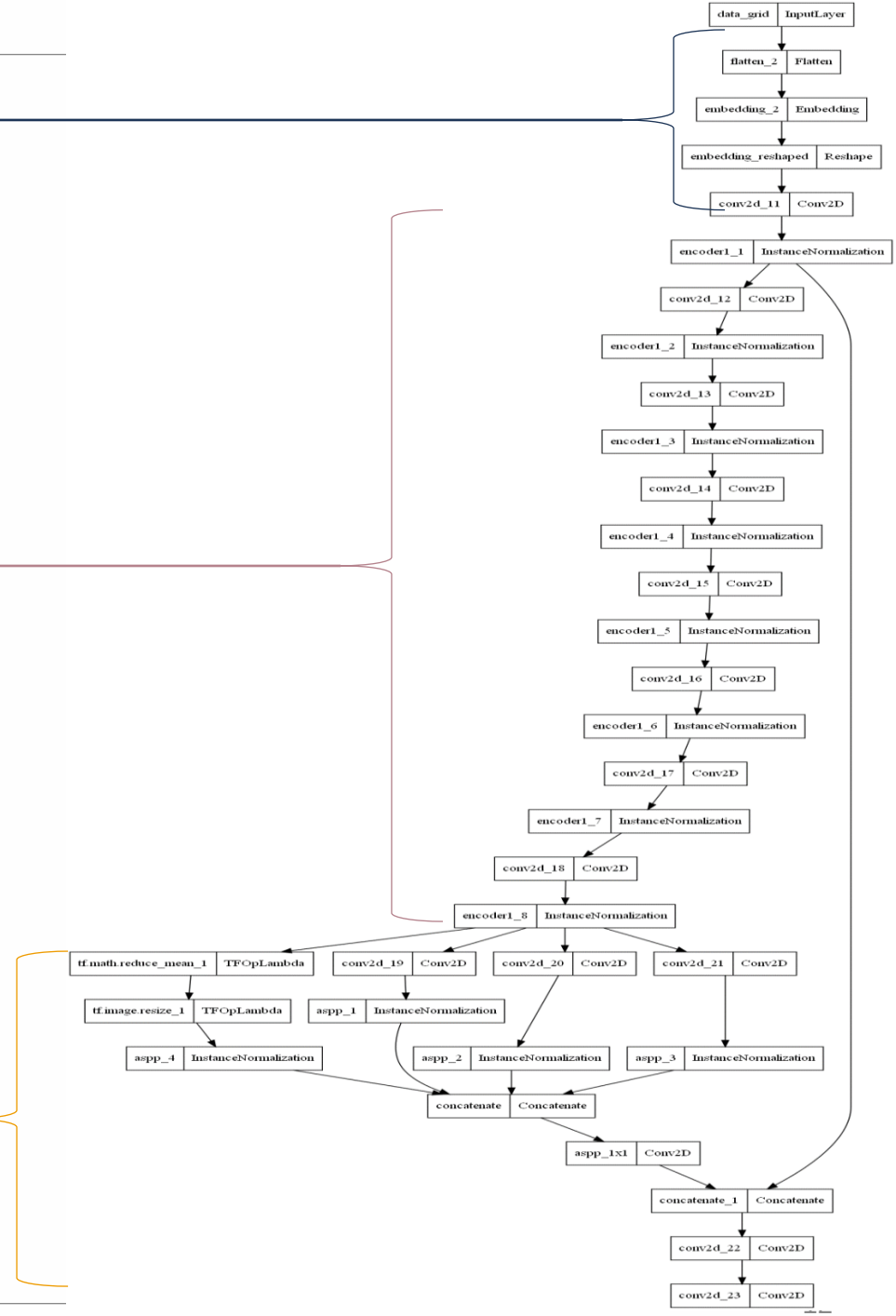


Source : scikit-image

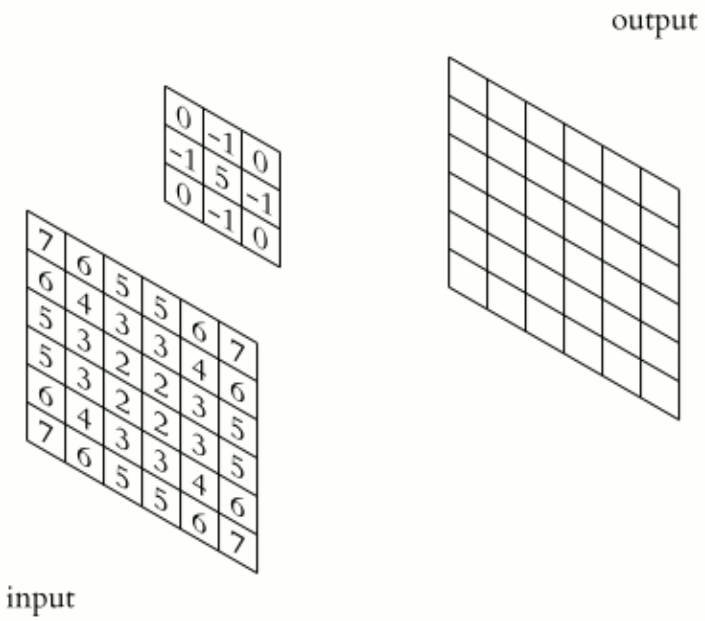
embedding

encoder
(convolutions et
atrous convolutions)

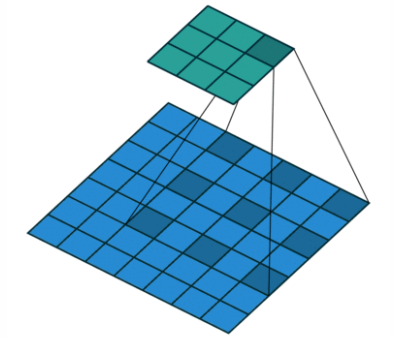
decoder
(concaténation
atrous convolutions)



ARCHITECTURE CUTIE



https://en.wikipedia.org/wiki/Kernel_(image_processing)

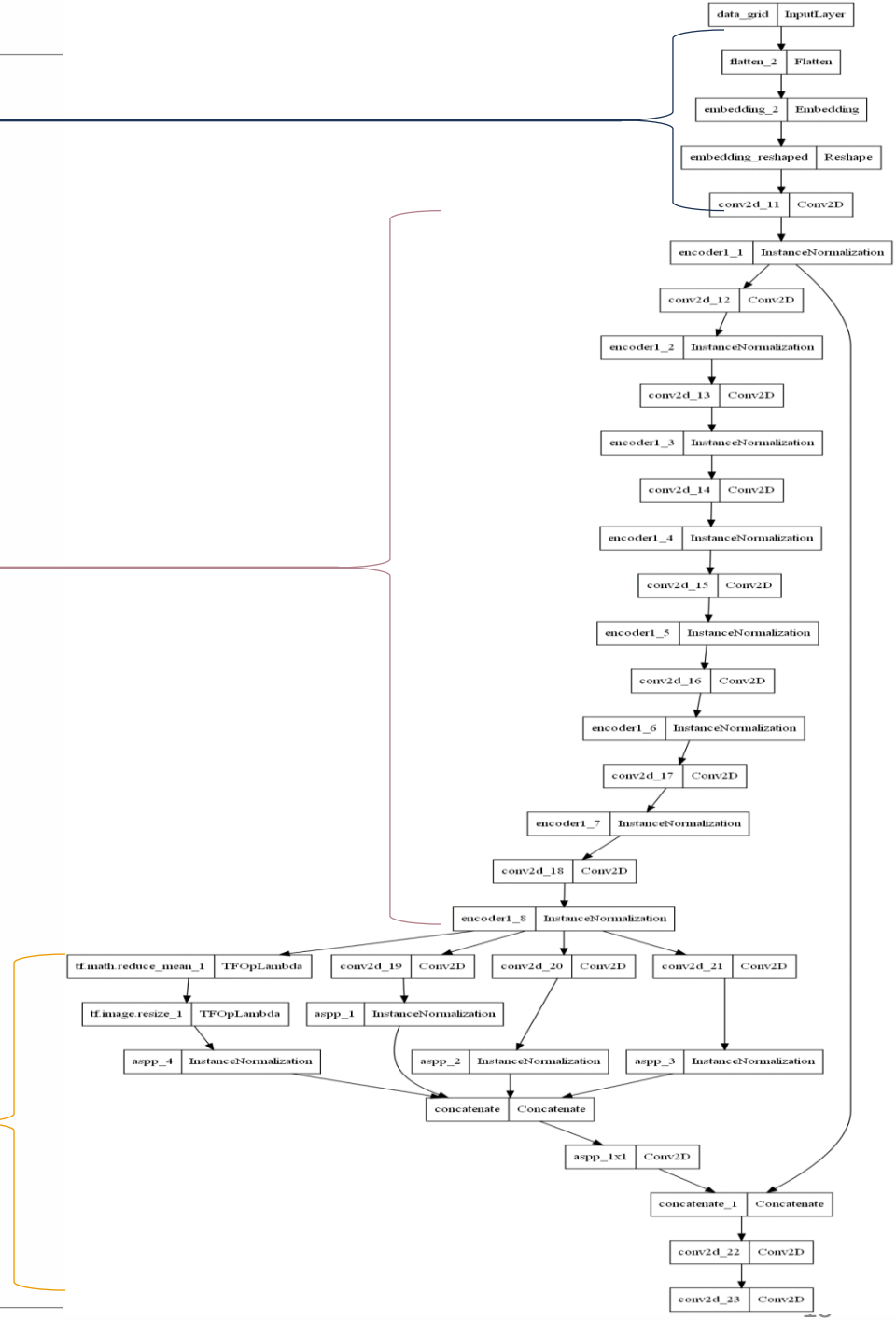


Towardsdatascience

embedding

encoder
(convolutions et
atrous convolutions)

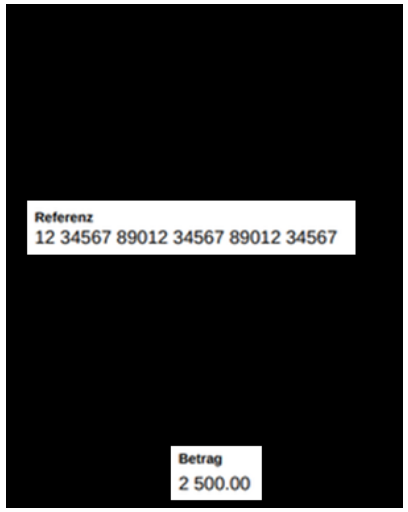
decoder
(concaténation
atrous convolutions)



ARCHITECTURE CUTIE

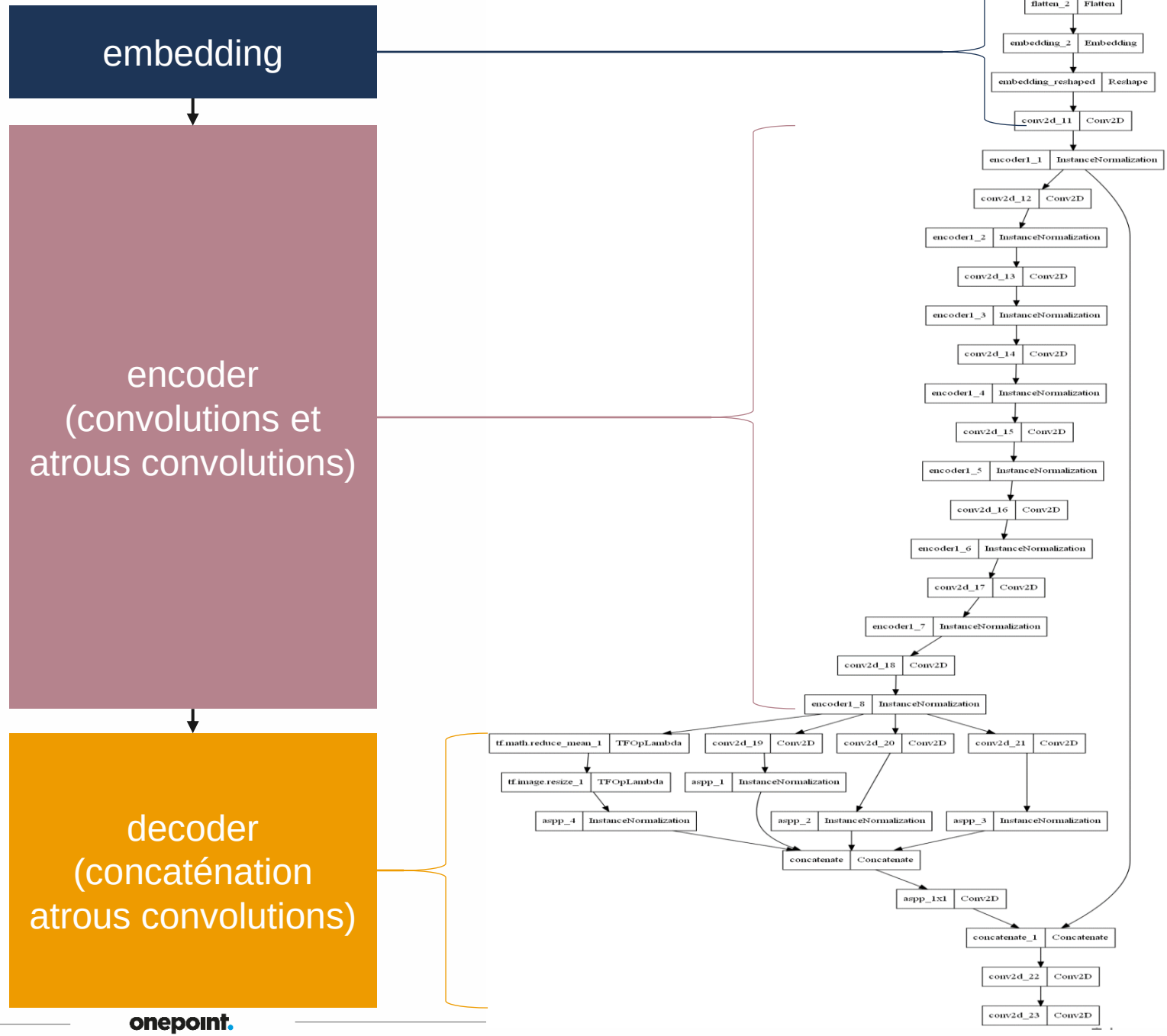
La couche d'embedding permet de transformer la grille de tokens (2D) en matrice 3D, ce qui permet un traitement plus fin au niveau des convolutions.

Durant l'entraînement, le réseau fixe les paramètres des filtres de convolutions permettant de mettre en évidence les éléments significatifs de la facture.



Les filtres de convolutions appris permettront de retrouver certains mots clés, certaines expressions régulières, etc.

La dernière couche produit un filtre de convolution par classe, à partir de laquelle la probabilité pour un token d'appartenir à cette classe est calculée.



EVALUATION

Métriques :

1. AP (Strict Average Precision) : tous les tokens détectés pour une classe doivent être corrects
2. SoftAP (Soft Average Precision) : autorise les faux positifs si et seulement si l'ensemble des vrais positifs sont trouves

Statut supposé réel de l'objet topique		Statut par rapport au critère d'analyse	
Qualité du constat	**Qualité du sujet**	**Jugement affirmatif (cas symptomatique)**	**Jugement réfutatif (cas asymptomatique)**
absence	exempt	**faux positif**	vrai négatif
présence	concerné	vrai positif	faux négatif

https://fr.wikipedia.org/wiki/Faux_positif

Table 2. Performance comparison on different types of documents. (AP/softAP)

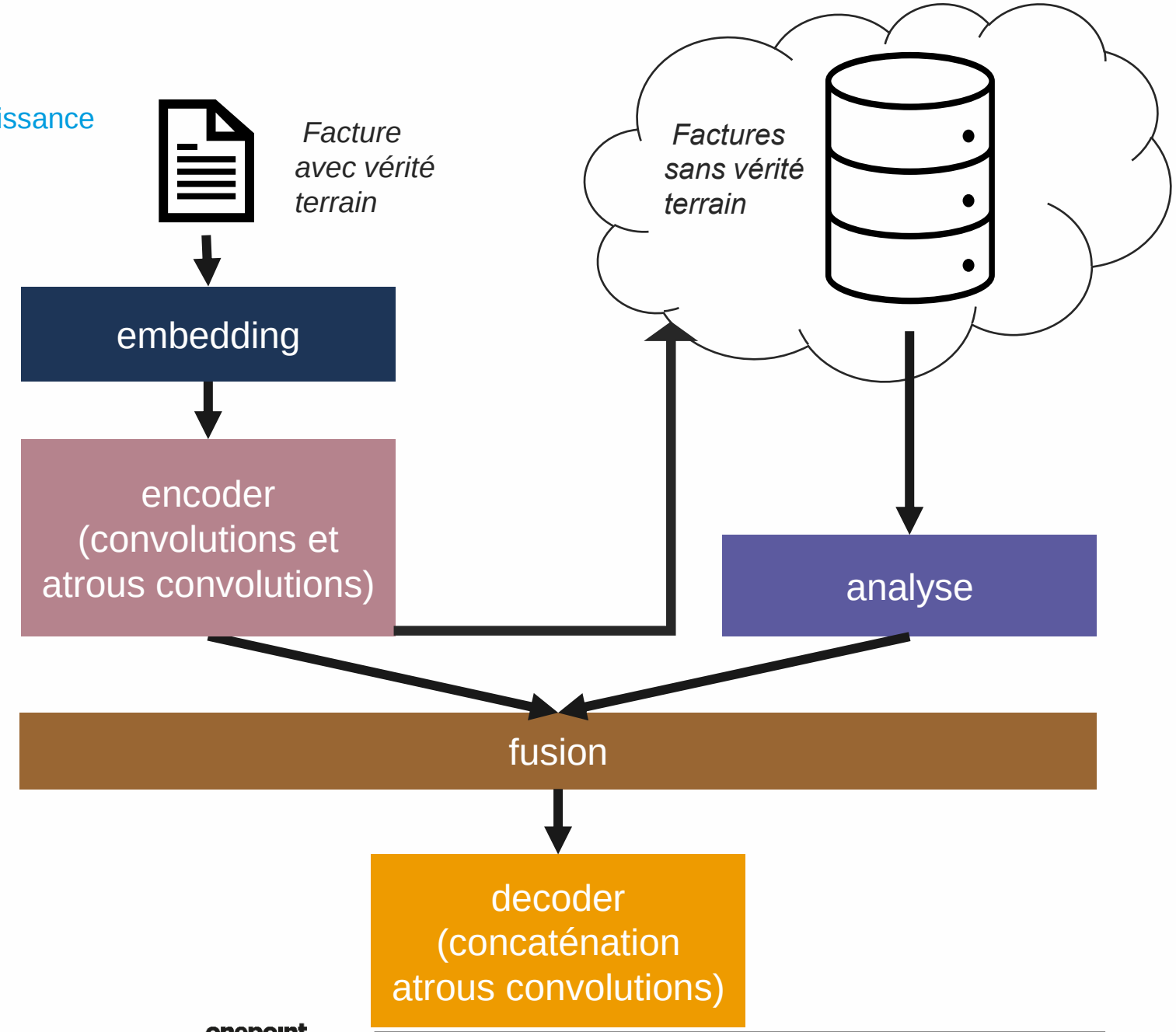
Method	#Params	Taxi	ME Meal	Hotel
CloudScan[9]	-	82 / -	64 / -	60 / -
BERT for NER[7]	110M	88.1 / -	80.1 / -	71.7 / -
CUTIE-A	67M	90.8 / 97.2	77.7 / **91.4**	69.5 / **87.8**
CUTIE-B	14M	**94.0 / 97.3**	**81.5** / 89.7	**74.6** / 87.0

BASE DE CONNAISSANCES

68 – Génération de code avec une base de connaissance

Motivations :

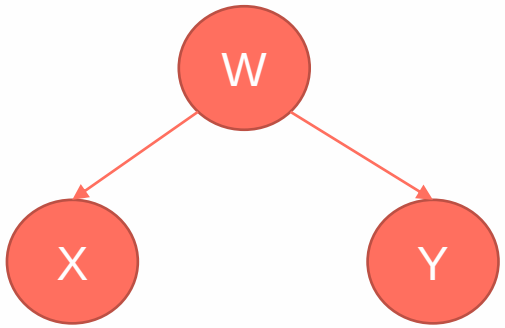
- Coût pour la création de la vérité terrain important => réduire la taille du dataset d'entraînement
- Similarités importantes entre factures d'un même fournisseur
- Similarités entre factures de différents fournisseurs



ETUDE D'IMPACT CAUSAL

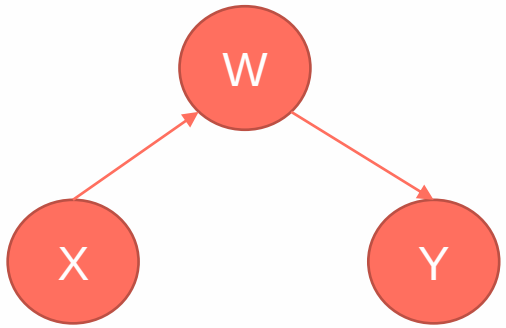
67 – Causalité 2 - Inférence causale avec des textes

Comprendre les données réellement utiles pour apprendre à identifier la catégorie sémantique d'un token => permettre d'anonymiser et de compléter les jeux de données.



Backdoor rule

Ex : pertinence du lexique
X : token = TVA (ou n'importe quel mot clé)
W : contenu textuel dans le voisinage de ce token
Y : montant TVA ou pas



Cause sans parent

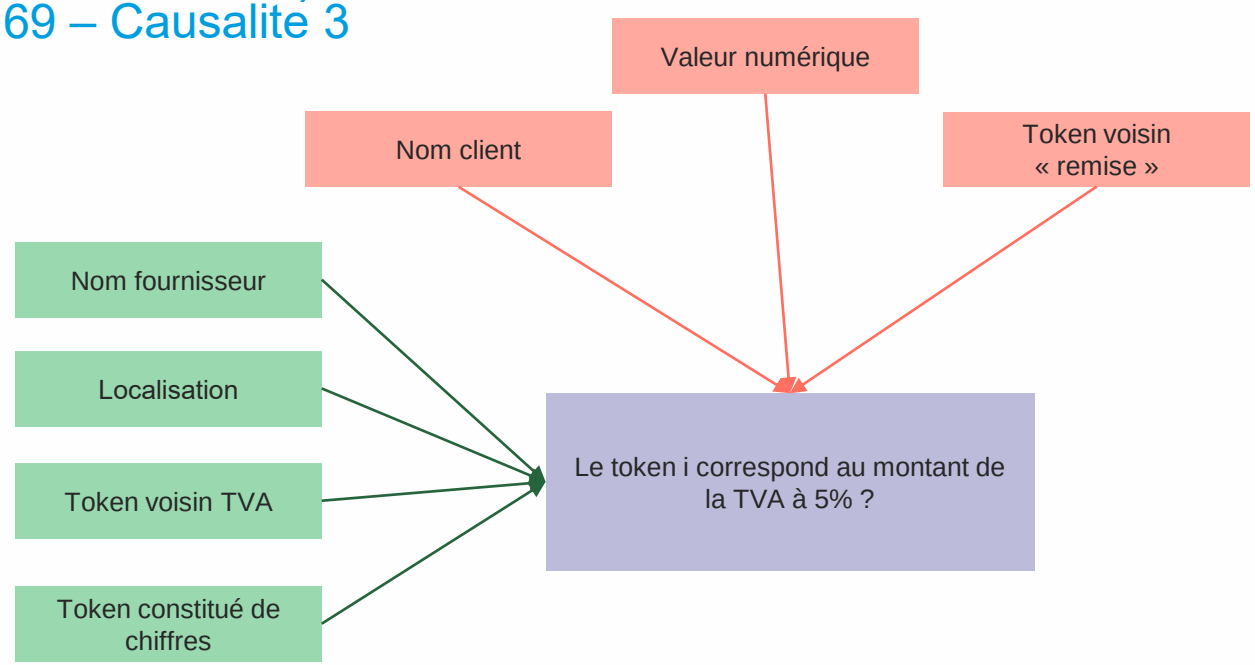
Ex : éléments pertinent pour déterminer le type de facture
X : est une adresse ou pas
W : contenu textuel au voisinage
Y : facture client ou fournisseur

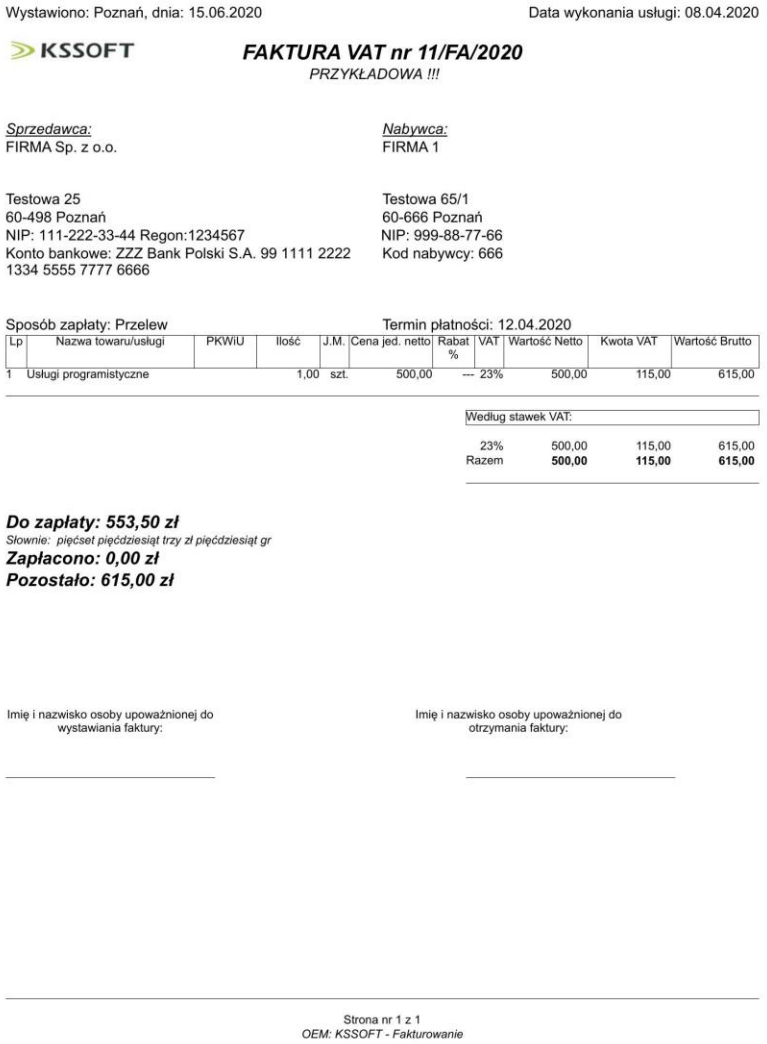
Génération d'un dataset semi-synthétique ?

Utilisation des propriétés sur les données

CORRÉLATIONS FALLACIEUSES (SPURIOUS)

69 – Causalité 3





Wystawiono: Poznań, dnia: 15.06.2020 — Data wykonania usługi: 08.04.2020

KSSOFT

FAKTURA VAT nr 11/FA/2020

PRZYKŁADOWA !!!

Sprzedawca:
FIRMA Sp. z o.o.

Testowa 25
60-498 Poznań
NIP: 111-222-33-44 Regon:1234567
Konto bankowe: ZZZ Bank Polski S.A. 99 1111 2222 1334 5555 7777 6666

Nabywca:
FIRMA 1

Testowa 65/1
60-666 Poznań
NIP: 999-88-77-66
Kod nabywcy: 666

Sposób zapłaty: Przelew — Termin płatności: 12.04.2020

Lp	Nazwa towaru/usługi	PKWiU	Ilość	J.M.	Cena jed. netto	Rabat %	VAT	Wartość Netto	Kwota VAT	Wartość Brutto
1	Usługi programistyczne		1,00	szt.	500,00	---	23%	500,00	115,00	615,00

Według stawek VAT:

23%	500,00	115,00	615,00
Razem	**500,00**	**115,00**	**615,00**

Do zapłaty: 553,50 zł
Słownie: pięćset pięćdziesiąt trzy zł pięćdziesiąt gr
Zapłacono: 0,00 zł
Pozostało: 615,00 zł

Imię i nazwisko osoby upoważnionej do wystawiania faktury:

Imię i nazwisko osoby upoważnionej do otrzymania faktury:

Strona nr 1 z 1
OEM: KSSOFT - Fakturowanie

- Afin de maximiser la capacité de généralisation du réseau de neurones nous voulons réduire l'influence des variables parasites (en rouge) sur l'attribution d'une classe i.
- La structure des données (boîtes englobantes + texte) rend extrêmement délicat le fait de faire varier les valeurs pour les variables parasites
- Une variable parasite pour attribuer une classe peut être une variable intéressante pour une autre classe