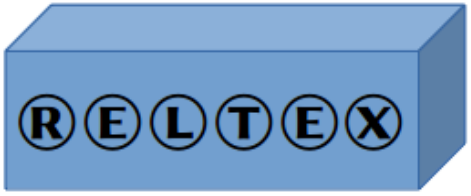
Apprentissage non supervisé pour l'extraction de relations d'hyperonymie à partir de textes scientifiques

Elena Manishina, Mouna Kamel, Nathalie Aussenac, Cassia Trojahn

l'IRIT, Toulouse

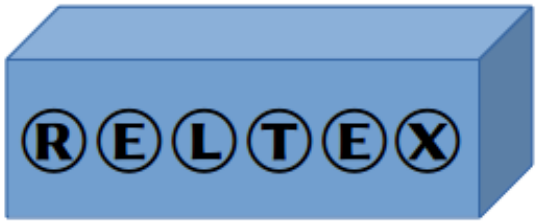
09 mai 2017



Equipe:

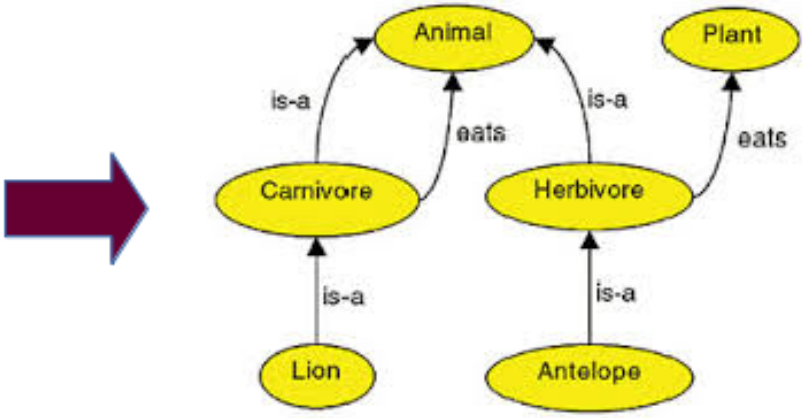
Nathalie Aussenac, Mouna Kamel, Elena Manishina, Cassia Trojahn

- **une bibliothèque scientifique numérique** - collections rétrospectives de la littérature scientifique dans toutes les disciplines: archives de revues, bases de données, corpus de textes, etc.
- **une platform** qui hébèrge plusieurs millions de documents numériques
  - un moteur de recherche
  - des services de traitement des données
  - une intégration à l'environnement numérique local
- différents domaines scientifiques
- 6 langues

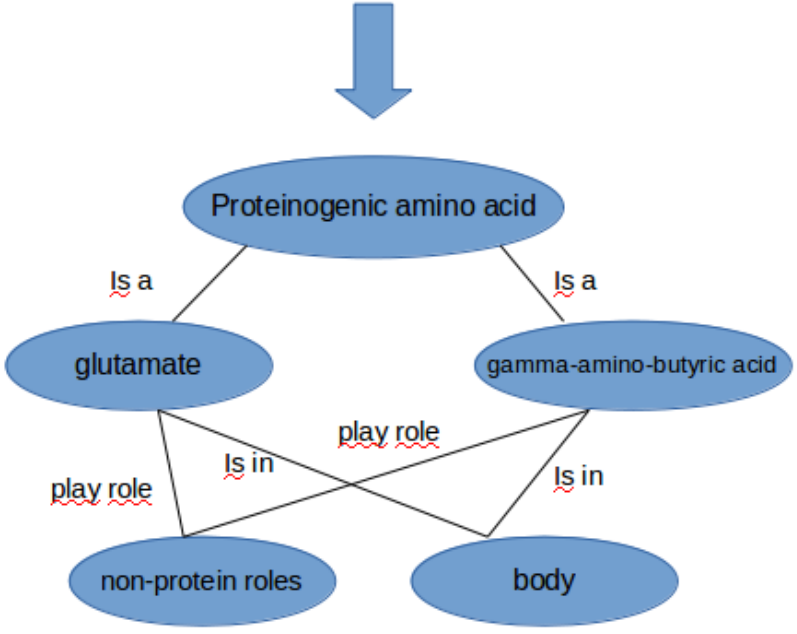


L'extraction de **connaissances** à partir de textes pour construire des **ressources sémantiques formelles** (ontologies)

Antelope is an animal of Herbivore family that eats plants. ....



Proteinogenic amino acids, such as **glutamate** (standard glutamic acid) and **gamma-amino-butyric acid** also play critical **non-protein roles** within the **body**.



Focus: nouveaux termes et relations qui ne sont pas présents dans des ontologies

Construction du corpus à partir de la collection ISTEX

La possibilité de créer des sous corpus (langue, domaine, periode, ...)

```
?q=language:eng AND corpusName:Nature  AND publicationDate:[2000 TO 2016] AND size:55000 ...
```

Choix de l'ontologie

NCIT - National Cancer Institute Thesaurus

Notre corpus final: plus de 50K docs, journal Nature, 2000-2010

National Cancer Institute Thesaurus[1]

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

118941 classes, 109 types de relations:

- Is_a
- Anatomic_Structure_Has_Location
- Anatomic_Structure_Is_Physical_Part_Of
- BioCarta_ID
- Biological_Process_Has_Associated_Location
- ...

Dans notre corpus:

- entités trouvées: **24K**
- relations présentes: **83**

[1] http://ncicb.nci.nih.gov/core/EVS

Structure des documents(articles)

- **texte (paragraphes)**

  Proteinogenic amino acids, such as **glutamate** (standard glutamic acid) and **gamma-amino-butyric acid** also play critical non-protein roles within the body.

- **tables**

Table 1: Alcaline reaction of test substances in the presence of a salt solutions

Salt	**test 1**	**test 2**	**alk. final**
$NaCl$	0.032	0.004	...
Na_2CrO_4	0.81	0.007	...
...	...	...	...

Structure des documents(articles)

- **structures enumeratives**
  We detected the presence of the following **metals**:
  1. **lithium** ($\approx 0.56 mgr/unit$)
  2. **sodium** ($\approx 0.04 mgr/unit$)
  3. **zink** ($\approx 0.014 mgr/unit$)
  4. **potassium** ($\approx 0.001 mgr/unit$)
  5. etc.

- **images et figures (légendes)**

Figure 1: **Oxides**, such as **iron(III) oxide** or rust, which consists of **hydrated iron(III)** oxides $Fe_2O_3 \Delta nH_2O$ and **iron(III) oxide-hydroxide** ($FeO(OH), Fe(OH)_3$), form when oxygen combines with other elements

Extraction des relations: le pipeline général

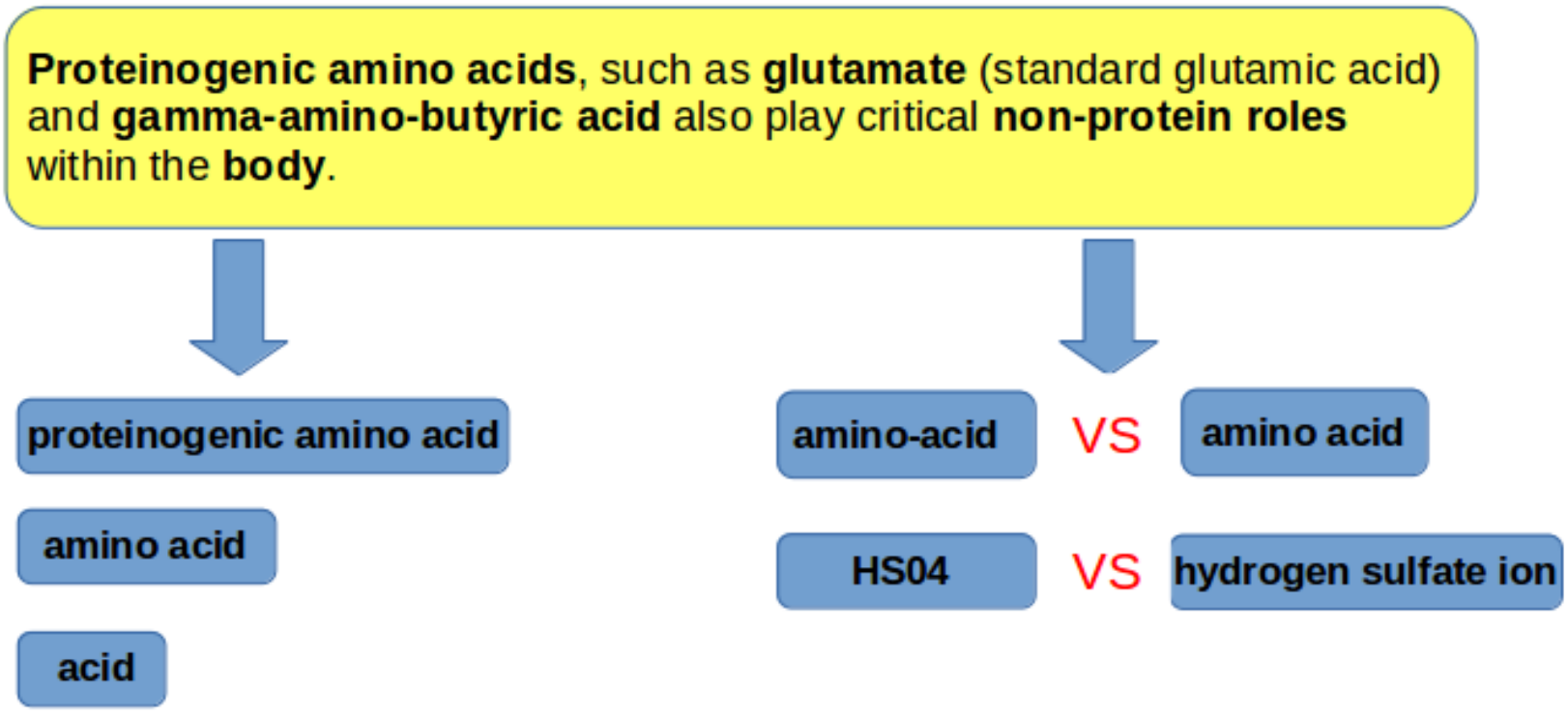
Etapes:

1. identifier les **structures** dans le document, découper le document en structures
2. construire un **corpus** pour chaque type de structure
3. identifier les **termes-candidats**
4. développer une liste **de paramètres/features** pour chaque structure
5. lancer un **algorithme d'extraction de relations** (avec les paramètres de l'étape 4)

NB!

- les corpus de taille differente => difference en performance

Etape 3: identifier les candidats



Focus: nouveaux termes, pas présents dans l'ontologie

Etape 3: identifier les candidats (II)

La procedure d'extraction à la base

- scoring distributionnel et compositionnel (NPs)
  - Termsuite [Cram et Daille, 2016]
  - Yatea [Hamon, 2006]

Mais:
- les termes ne sont pas dans l'ontologie
- pas d'experts pour l'evaluation manuelle

Afin d'evaluer notre approche:

- projeter les termes et les relations de l'ontologie sur le corpus
- lancer l'algo et voir les resultats

Construction des instances d'entrainement

Proteinogenic amino acids, such as **glutamate** (standard glutamic acid) and **gamma-amino-butyric acid** also play critical **non-protein roles** within the **body**.

glutamate + **gamma-amino-butyric acid**

glutamate + **proteinogenic amino acids**

glutamate + **non-protein roles**

etc...

Conditions

- candidats dans la phrase
- max 1 terme entre les candidats

Choix de l'algorithm

Nos constraintes:

- pas d'annotations (termes et relations) - coûteux de produire
- corpus de petite taille
- domaine specifique
- structures hétérogènes

Notre choix - apprentissage non-supervisé

- independent du domaine et de la langue
- independent du type des relations
- flexible and facile à ajuster à la tâche spécifique

Apprentissage non-supervisé: clustering

Objective: découvrir des patrons lexico-morpho-syntaxiques d'une façon non-supervisé en regroupant les patrons similaires

Algorithmes:

- **K-means** $\Leftarrow$ baseline
- **SOM (Self-organizing maps)** $\Leftarrow$ notre choix

KMeans VS SOM

- définir le nombre de clusters en avance (KMeans)
- distribution aléatoire des centroids initiaux $\Rightarrow$ influence sur les sorties/resultats
- **apprentissage competitif** (vector quantization) VS apprentissage par correction d'erreurs (backpropagation avec gradient descent)

Représentation vectorielle des paramètres

La base: répresenter le contexte avec word2vec[2]

Contexte:

- lemmas du contexte => **patrons lexicaux**:
  - 5 unités à gauche, droite et au milieu des candidats
- + Contexte POS => **patrons morphologiques**:
  - 5 unités à gauche, droite et au milieu des candidats
- + Arbres de dependence => **patrons syntaxiques**

La représentation vectorielle de chaque couple ⇒ beaucou de dimensions ⇒ utiliser SOM pour reduire la dimensionnalité

[2] Mikolov et al.,2013

Self-organizing maps

Le reseau de neurons artificiel:

- une façon de représenter les données multidimensionnelles dans l'espace 2D - **vector quantization**
- relations topologiques entre les instances sont préservées

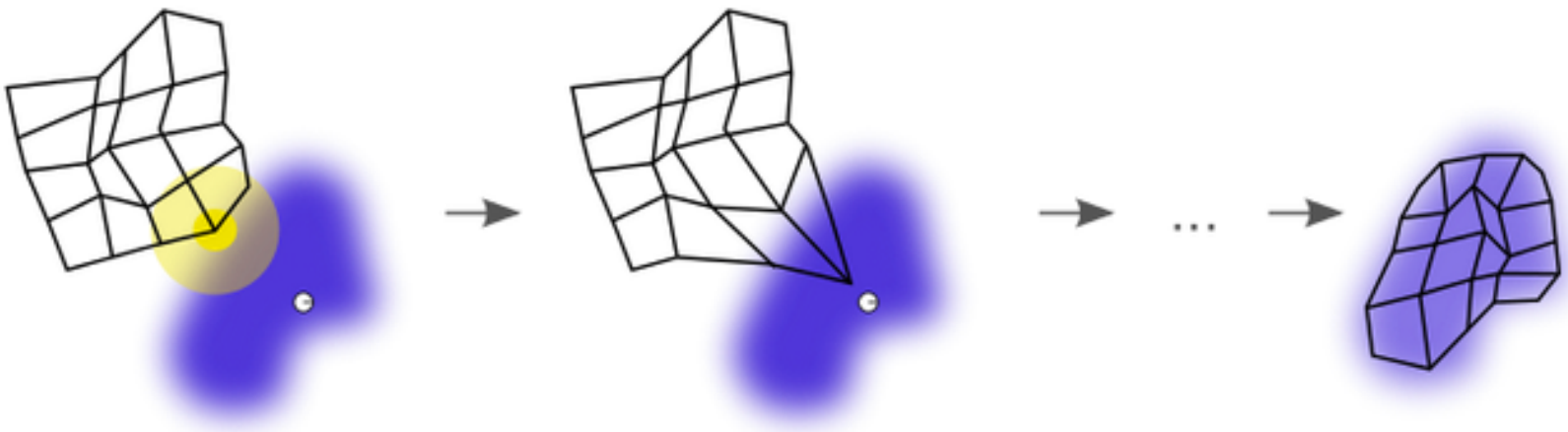
un neuron v avec le vector des poids $W_v(s)$:

$$W_v(S+1) = W_v(s) + \theta(u, v, s) \cdot \alpha(s) \cdot (D(t) - W_v(s))$$

L'algorithme:

- choisir les poids sur les noeuds
- prendre le vecteur $D(t)$
- traverser chaque noeud sur la carte
- mettre à jour les noeuds voisins
- augmenter s and répéter la procedure à partir de l'étape 2 si $s < \lambda$

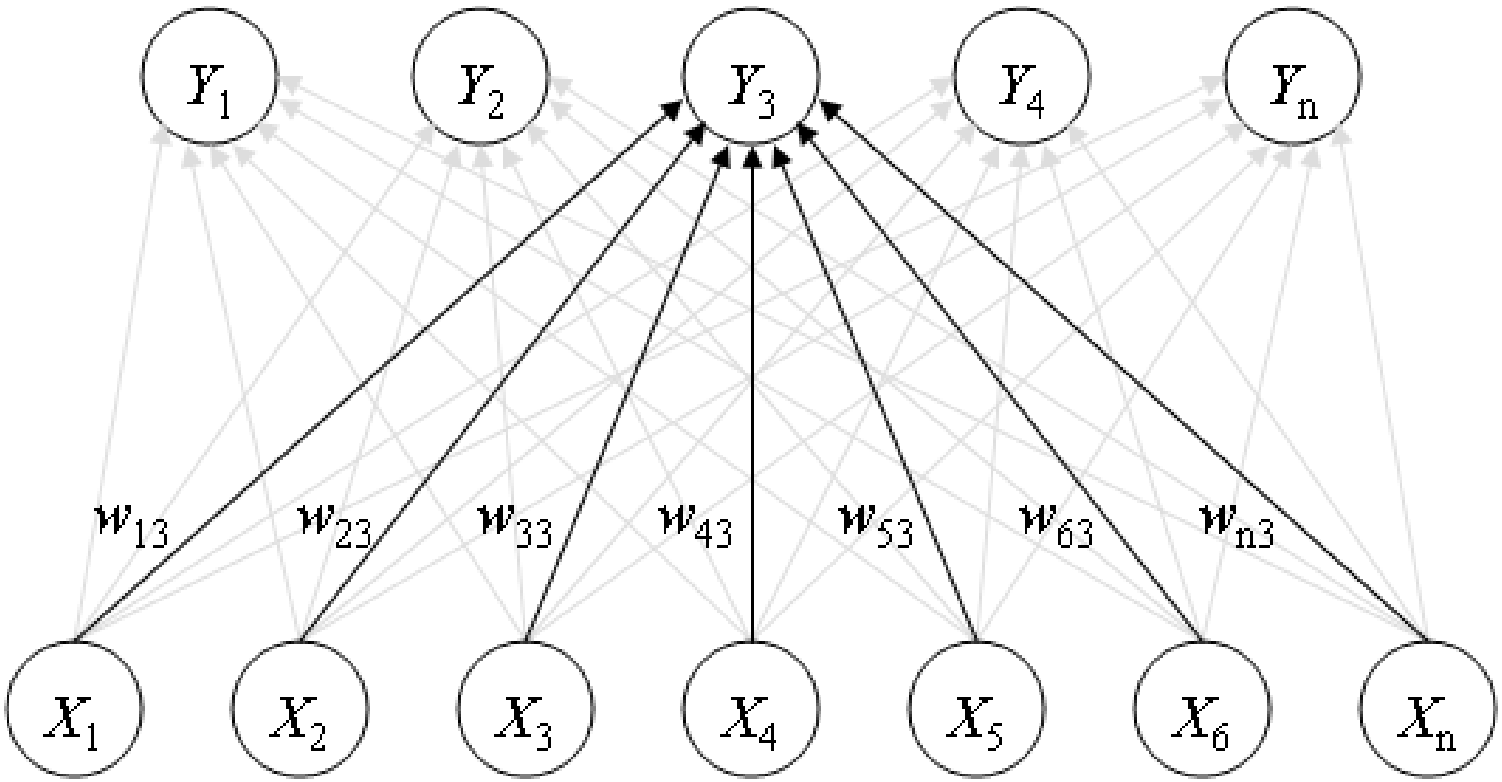
SOM training



SOM training[3]: apprendre la distribution des données dans le corpus

[3] https://en.wikipedia.org/wiki/Self-organizing_map

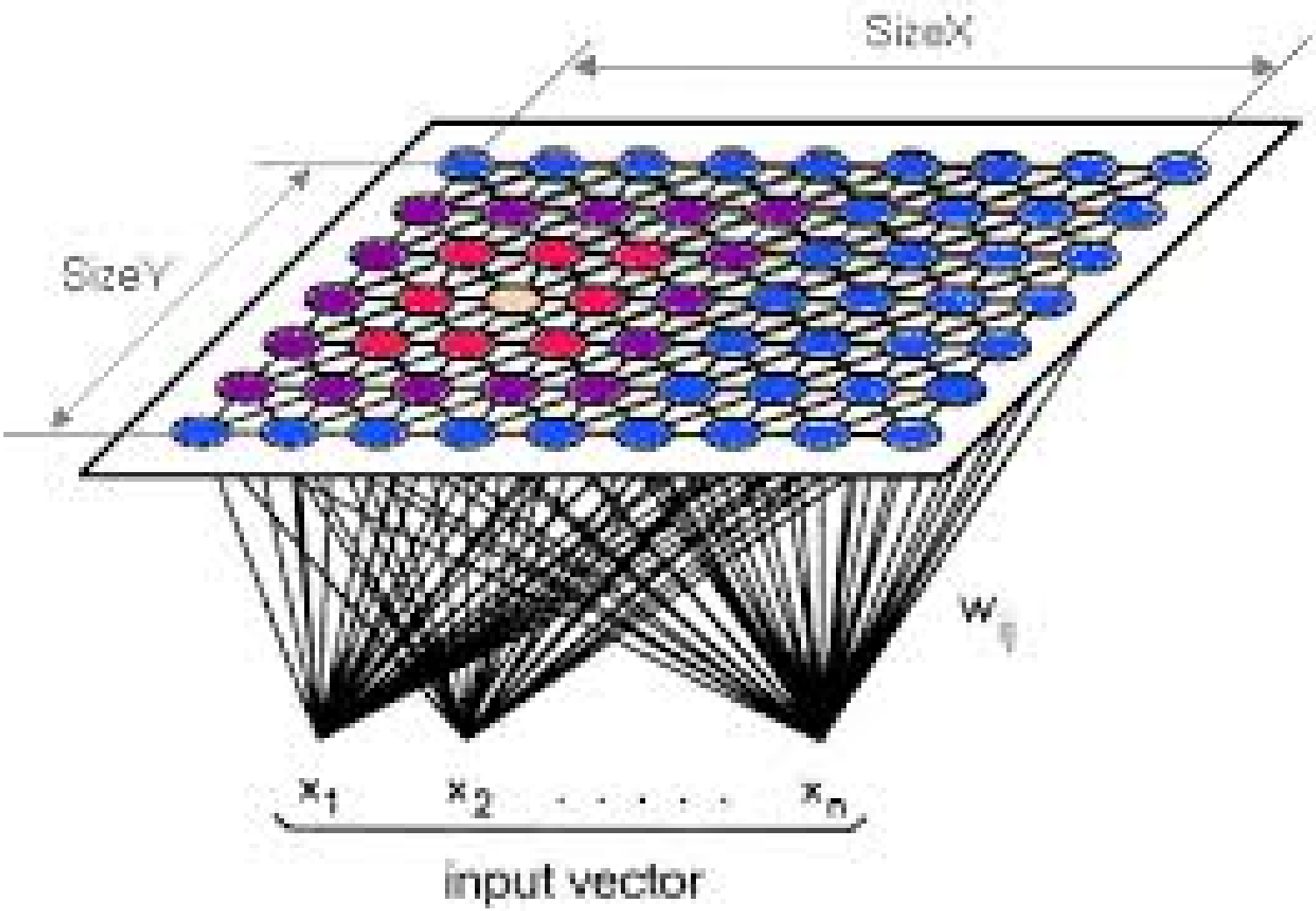
SOM training



SOM training[4]: les vecteurs des poids

[4] https://www.mnemstudio.org

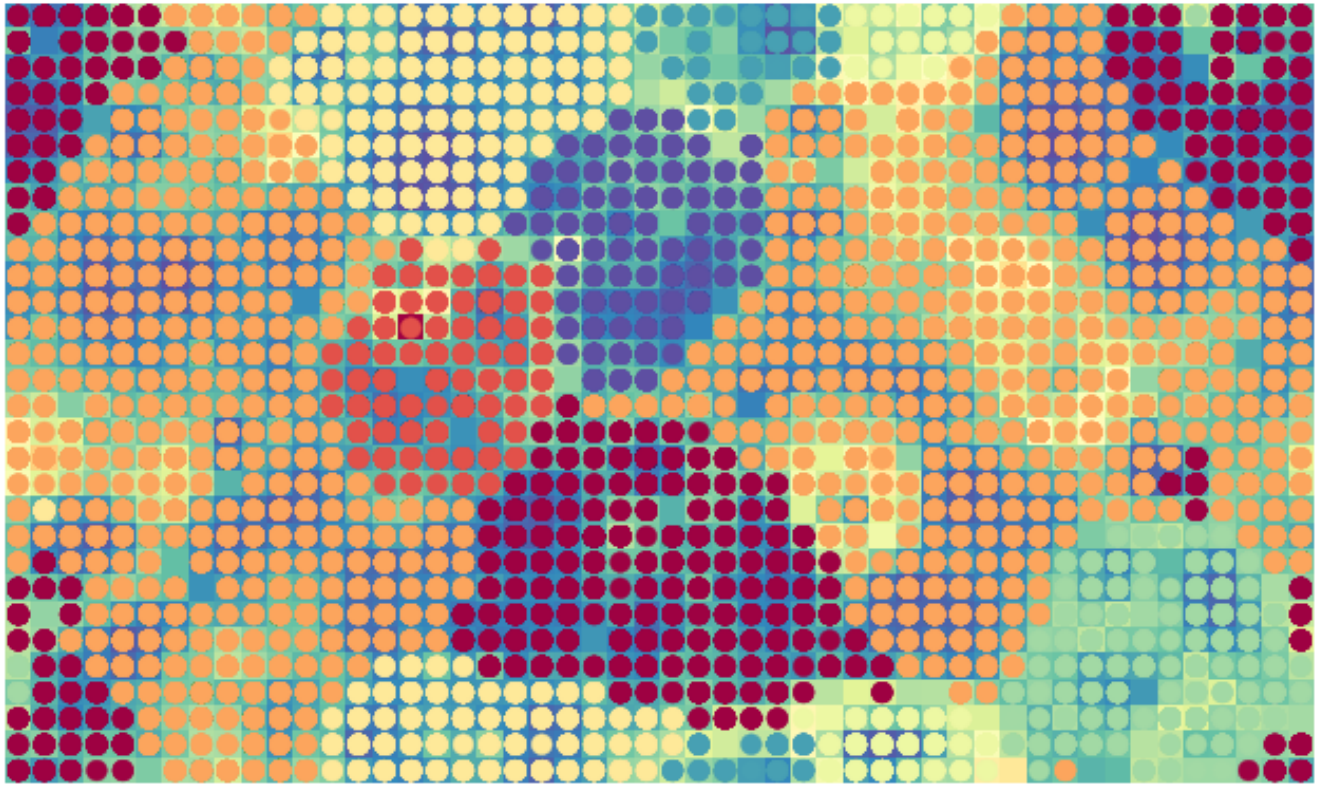
SOM training



SOM training.[5]: BMUs et voisins

[5] https://www.pitt.edu

Evaluation: 30x50 map



Hyperonyms

Table 2: F score: map size, features

map size	lem	lem+POS	lem+POS+Syn
30x50	63.6	78.2	?
70x90	65.1	80.3	?

- (30x50) TP et TN: **77.2%** et **81.1%** accuracy
- (70x90) TP et TN: **78.9%** et **83.5%** accuracy

Autres relations

- Biological_Process_Has_Location
- Gene_Has_Physical_Location
- Anatomic_Structure_Has_Location
- Chemical_Plays_Role_In_Bio_Process

Table 3: F score: rel, lem+POS

map size	BPL	GPL	ASL	CRBP
30x50	54.4	61.2	48.8	66.1
70x90	55.1	60.8	51.2	68.5

Conclusions

General:

- evaluation en cours
- clusteurs bien définis => capacité discriminative suffisante des paramèrtres
- certains clusteurs sont mieux separés que les autres

Futur:

- calculer P/R pour toutes les relations dans le corpus
- reproduire la procedure sur toutes les autres structures
- tester le pipeline sur le corpus Wikipedia

Merci!

Thank you!