Unsupervised relation extraction using self-organizing maps

Elena Manishina, Mouna Kamel, Nathalie Aussenac, Cassia Trojahn

l'IRIT, Toulouse

09 march 2017

RelTEX project

Team:

- Nathalie Aussenac
- Mouna Kamel
- Elena Manishina
- Cassia Trojahn

Funding:

Financement : ANR-10-IDEX-0004-02

ISTEX collection

What is it?

- **a digital library** - retrospective collections of scientific literature in all disciplines: journal archives, books, databases, texts corpora, etc.
- **a platform** that host several million digital documents:
  - a powerful search engine
  - data processing services: data extraction, text mining, etc.
  - integration into local digital environments (widgets)
- different domains
- 6 languages

Numbers:

- Springer e-books from 1995 to 2004 - **7500** titles
- Wiley journals - nearly **2200** titles from 1791 to 2011
- etc ...

Quering ISTEX and creating subcorpora

The possibility to create specific dataset (language, domain, time period, ...) to carry out specific experiments

$?q = language : eng\ AND\ corpusName :$
$Nature\ AND\ publicationDate : [2000 TO 2016]\ AND\ size :$
$10000\ AND\ ...$

- formats: plain text, pdf, xml, tei
- occasionally enriched with additional info (layout, metadata,...)

Our corpus: more than 50K docs, Nature, 2000-2016 (!)

Document structure (articles)

- **text (paragraphs)**

  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	...
...	...	...	...

Document structure (articles)

- **enumerative structures**
  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 and figures (captions)**

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

General pipeline

Steps:

1. identify structures in the document, split the document into structures
2. build a corpus for each structure type
3. extract features specific for this structure
4. run the algorithm on each corpus

NB!

- Select features for each structure
- training corpora of different sizes => difference in performance

Relation extraction

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

Relations:
glutamate is_a proteinogenic amino acid
gamma-amino-butyric acid is_a proteinogenic amino acid
glutamate is_in body
gamma-amino-butyric acid is_in body
...

Focus: new terms and relations, not present in ontologies

Step 1: identify the candidates

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

Term border identification

proteinogenic amino acid
amino acid
acid

Term variants

amino-acid VS amino acid (training)
HS04 VS hydrogen sulfate ion (evaluation)
...

Focus: new terms, not present in ontologies

Step 1: identify the candidates (II)

Custom extraction procedure

- distributional and compositional scoring (NPs)

To evaluate the approach:

- pick an ontology (NCIT, MESH)
- map the terms from the ontology onto the corpus
- look for relations present in the ontology
- calculate P and R for **that specific onology**

NCIT ontology

Statistics

- **118941 classes**
- **109 relation types**:
  - Is_a
  - Anatomic_Structure_Has_Location
  - Anatomic_Structure_Is_Physical_Part_Of
  - BioCarta_ID
  - Biological_Process_Has_Associated_Location
  - ...

Step 2: building training instances

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

Couples of terms

(proteinogenic amino acids, glutamate)
(proteinogenic amino acids, gamma-amino-butyric acid)
(glutamate, gamma-amino-butyric acid)

Constrains

- candidates within the phrase
- max 1 term between the candidates

Step 3: picking an algorithm (background)

Morpho-syntactic patterns

built manually for a specific domain and a given language

Supervised ML:

Lately:

- Zeng et al., 2014 (Convolutional Neural Networks)
- Liu et al, 2015 (Dependency-based Neural Networks)

But: require annotated corpus

Unsupervised ML

- Riedel et al, 2013 (augmented bootstrapping)
- Yan et al., 2009 (k-means) $\Leftarrow$ **state-of-the-art work well on large general domain corpora**

Step 3: picking an algorithm (background)

Our constraints:

- no annotations (terms and relations) - costly to produce
- smaller size
- specific domain
- heterogenous corpus structure

Our choice - unsupervised learning

- domain and language independent
- relation independent
- flexible and tunable

Unsupervised: clustering

Discover lexico-morpho-syntactic patterns in an unsupervised manner by grouping similar patterns

Algorithms:

- K-means
- BFR
- CURE
- **SOM (Self-organizing maps)** $\Leftarrow$ our choice

KMeans VS SOM

- define the number of clusters in advance (KMeans)
- random distribution of initial centroids $\Rightarrow$ influence the output
- **competitive learning** (vector quantization) VS error-correction learning (backpropagation with gradient descent)

Self-organizing maps

artificial neural network:

- a way of representing multidimensional data in much lower dimensional spaces (2 in our case) - **vector quantization**
- topological relationships within the training set are preserved

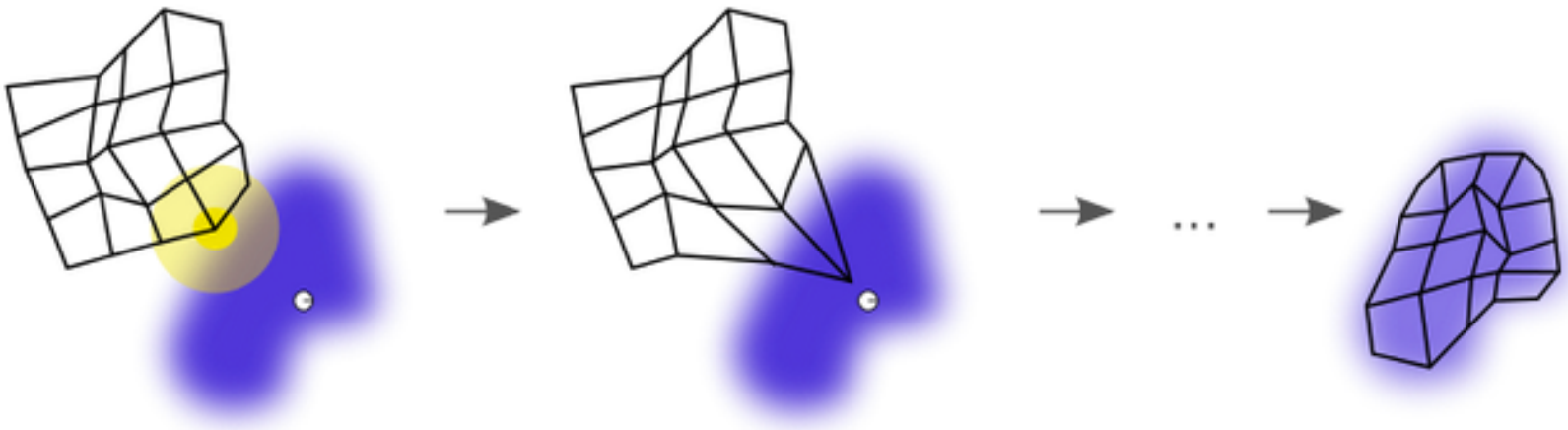
a neuron v with weight vector $W_v(s)$:

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

Algorithm:

- Randomize the map's nodes' weight vectors
- Grab an input vector $D(t)$
- Traverse each node in the map
- Update the nodes in the neighborhood of the best matching unit
- Increase s and repeat from step 2 while $s < \lambda$

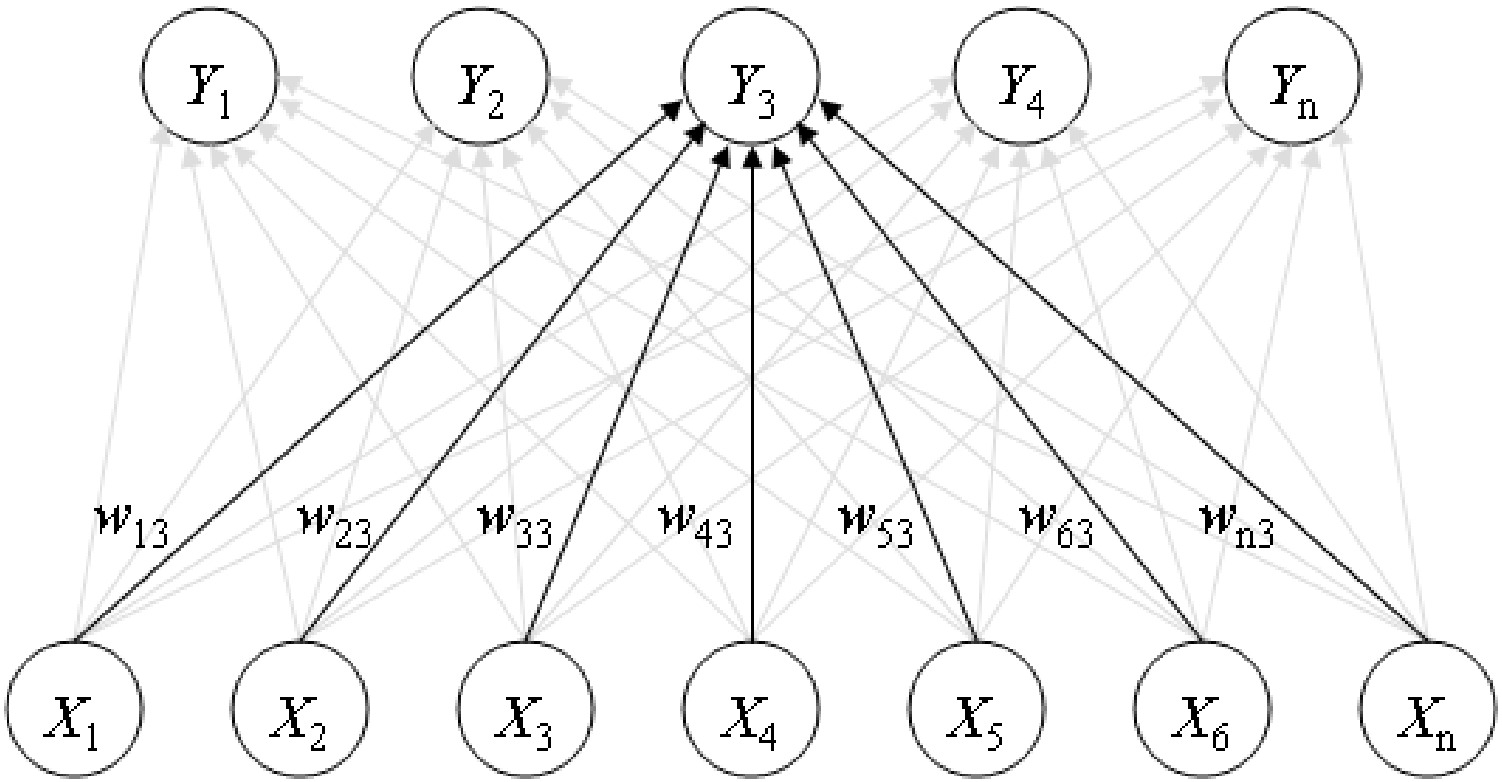
SOM training



SOM training[1]: approximating the data distribution

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

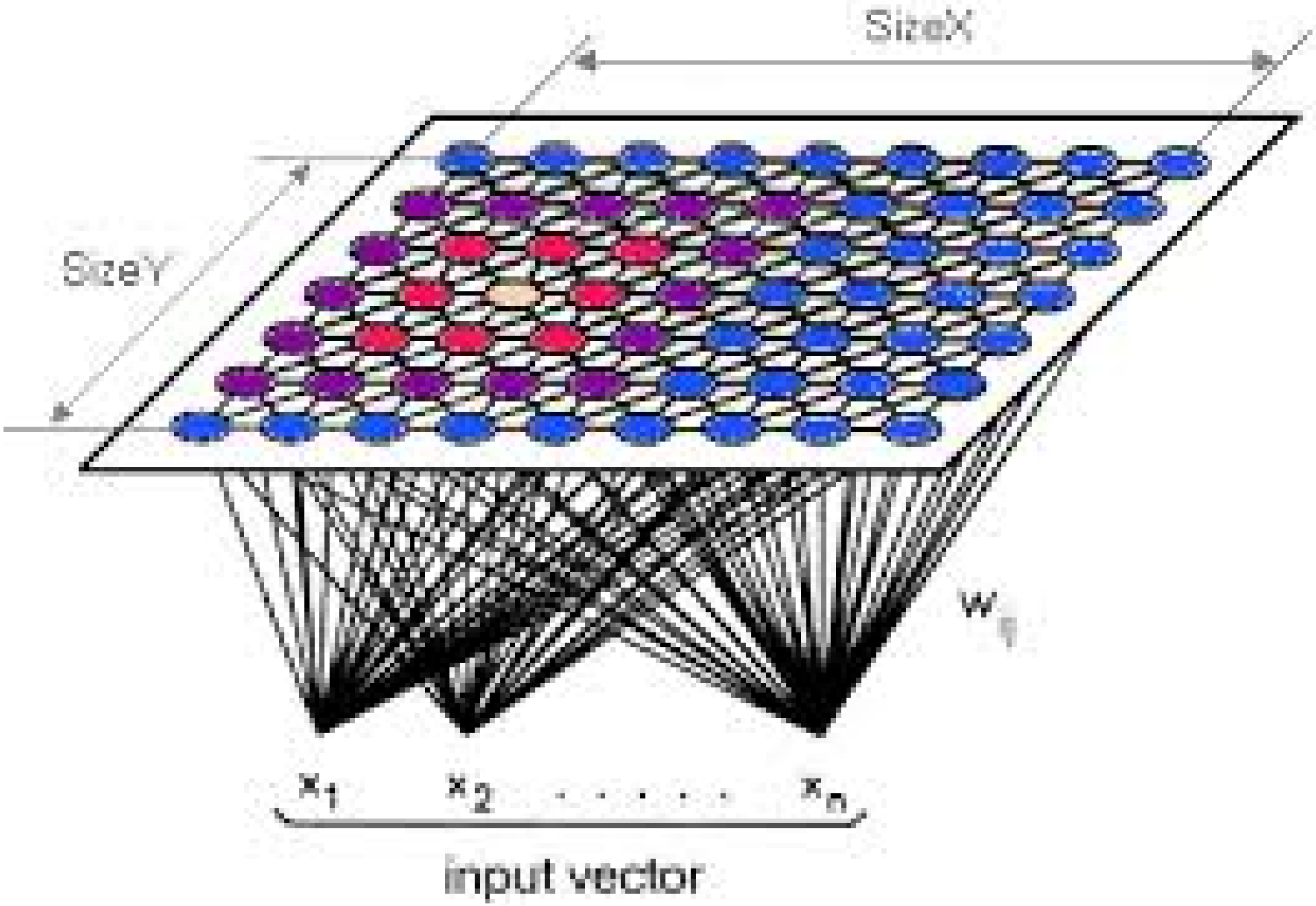
SOM training



SOM training[2]: weight vectors

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

SOM training



SOM training.[3]: BMUs and neighbors

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

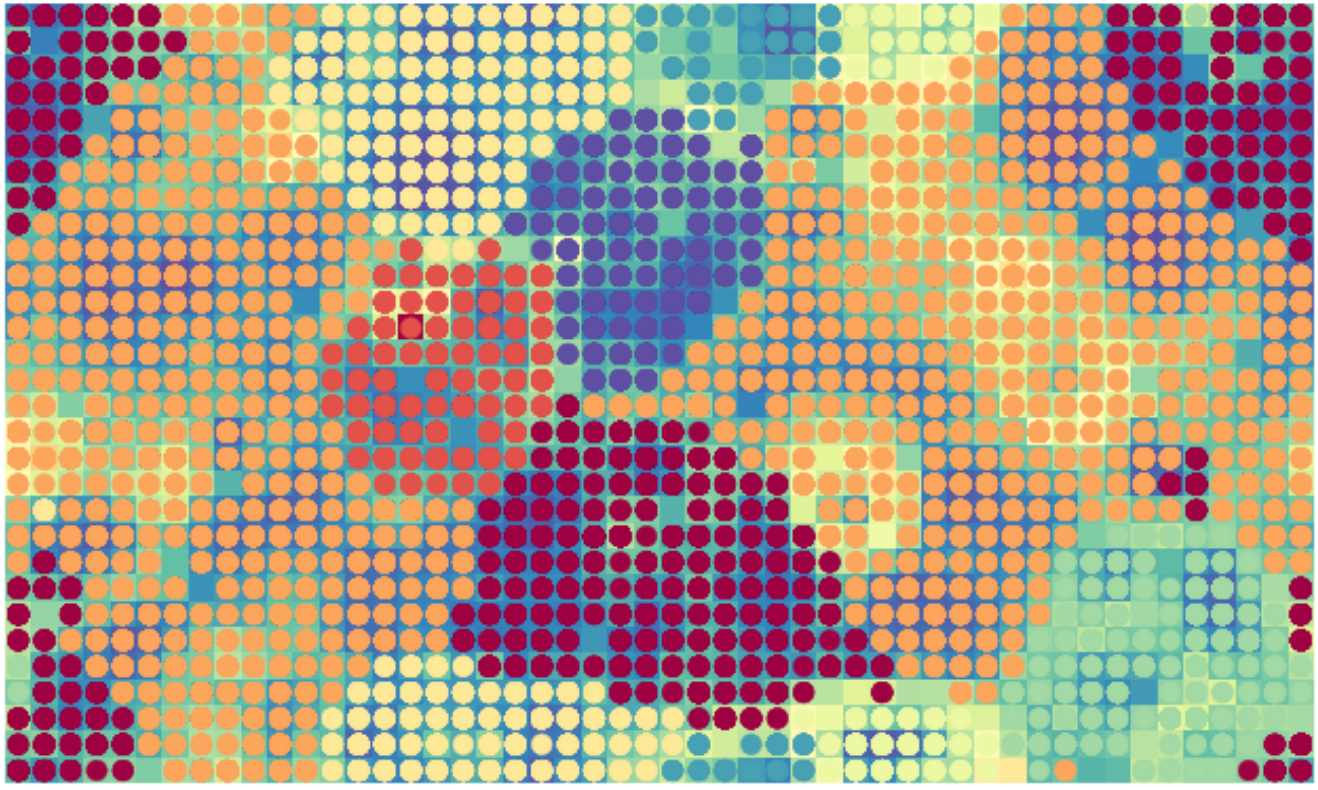
Features

Context:

- Context lemmas => **lexical patterns**:
  - 5 units left, right and between the candidates
- + Context POS => **morphological patterns**:
  - 5 units left, right and between the candidates
- + Context dependency trees => **syntactic patterns**

vector representation of each unit $\Rightarrow$ many dimensions $\Rightarrow$ use SOM dimension reduction capacity

Evaluation: our data



Hyperonyms

- TP and TN: **77%** and **81%** accuracy

Conclusion

General:

- on-going work
- preliminary results are encouraging :)

Future work:

- calculate P/R for other relations (define the relations first)
- reproduce the procedure on each structure corpus
- test the pipeline on Wikipedia corpus

Merci!

Thank you!

Danke!

Gracias!