RelTEX : Usage du corpus ISTEX pour l'exploitation de méthodes d'extraction de connaissances à partir de textes

Nathalie Aussenac-Gilles aussenac@irit.fr
Mouna Kamel - Cassia Trojahn
Elena Manishina (Post-doc)

Collaboration avec Cécile Fabre et
Anne Condamines (CLLE-ERSS)

RelTEX : projet chantier d'usage

- Domaines scientifiques concernés
  - TAL, extraction de connaissances
  - Web sémantique
- Problématique
  - Usage du corpus ISTEX pour l'exploitation de méthodes d'extraction de connaissances à partir de textes
- Problématique plus précise
  - Extraction de relations sémantiques entre entités ou entre classes à partir des textes
  - Évaluation d'approches faisant appel au langage naturel et à la structure des textes
- Projet complémentaire SemPédia (2015-2018)
  - financé par la région Midi-Pyrénées et la COMUE de Toulouse
  - Plateforme d'extracteurs de relations pour Wikipedia en français

Plan de la présentation

Extraction de relations

Etat de l'art

Méthodologie

Expérimentations et résultats

Conclusions et perspectives

Plan de la présentation

Extraction de relations

- Tâche appartenant à l'extraction d'information (IE) et qui vise à extraire des relations sémantiques entre des entités du monde.
- Tâche cruciale pour la construction de ressources ou applications TAL :
  - e.g. : WordNet, DBpedia, etc.
  - e.g. : Recherche d'Information, Systèmes Question-Réponse, etc.

- **Exemple** : relation R_1 d'hyperonymie entre E_1 et E_2 :

- Une fois extraite, cette information est formalisée (FOL, DL, etc.) :
  e.g. : $\forall x(\text{arbre}(x) \implies \text{plante-lignifiée-terrestre}(x))$
  e.g. : $\textit{arbre} \sqsubseteq \text{plante-lignifiée-terrestre}$

Analyser des collections ISTEX

- Caractériser le contenu d'articles scientifiques
- Articles scientifiques comme sources de connaissances
- Forme "prévisible" et variée
  - langage naturel rédigé (paragraphes) avec une structure discursive
  - structures de représentation de données (tableaux, listes) et de langage naturel (énumérations)
  - plan, sections avec titres et sous-titres
  - ces différentes structures nécessitent une analyse distincte

Objectifs généraux

- Utiliser la collection ISTEX au service des recherches en TAL
- Contribuer à nos recherches sur l'extraction de connaissances pour construire des ressources sémantiques formelles à partir de grandes masses de données textuelles, dans des domaines variés

Objectifs spécifiques

A court terme : définir des extracteurs de relations pour des corpus scientifiques

- Exploiter des indices linguistiques divers
  - Lexique, syntaxe
  - dépendances
  - **relations discursives et structure logique des documents**
- Focus sur peu de types de relations : hyperonymie, méronymie
- Proposer plusieurs techniques complémentaires : Patrons, composition de mots, co-occurrence, apprentissage supervisé

A plus long terme : Vers une plateforme modulaire d'extraction de relations répondant à différents besoins

Objectifs reformulés

Expérimenter plusieurs extracteurs de relations pour des corpus scientifiques

- tester des techniques déjà éprouvées sur des sous-corpus en exploitant des indices linguistiques divers
  - Lemmes, POS, dépendances
  - relations discursives et structure logique des documents
- Focus sur **tout** type de relations : relations du domaine du corpus
- Prospecter de nouvelles techniques d'extraction "ouverte" de relations : **apprentissage non supervisé** avec RNN

Plan de la présentation

Trois grandes familles de méthodes

- Approches symboliques à base de patrons lexico-syntaxiques
  - relations connues, formulations en corpus identifiées
  - adapter les patrons au corpus : coûteux
- Approches statistiques (apprentissage)
  - Intégrer au sein d'un processus de classification, des connaissances linguistiques au moyen de données annotées
  - supervisé : annotation manuelle d'exemples
  - semi-supervise (supervision "distante") : annotation automatique d'exemples à l'aide d'une ressource
  - non supervisé pas d'annotation
- Critères de choix
  - corpus : genre textuel, domaine, volume
  - disponibilité de ressources sémantiques ou d'expertise
  - disponibilité d'exemples annotés
  - application visée

Extraction de relations par apprentissage supervisé

- Deux sous-tâches : décider si 2 entités sont en relation ; identifier la nature de la relation
- Méthode :
  - classer des couples de termes à partir d'exemples
  - chaque couple est représenté par un ensemble de traits calculés à partir de son occurrence en corpus
- Modèles : SVM, MaxEnt, réseaux de neurones (RNN)
- Traits : contextes des entités (fenêtres de mots), traits syntaxiques, lexicaux et sémantiques, arbres de dépendances, fonctions noyaux
- Limites :
  - disposer d'exemples d'entraînement annotés
  - modèle à entraîner pour chaque domaine, corpus, langue ...
  - RNN : nécessitent un grand volume de données

Extraction de relations par apprentissage non supervisé

- K-Means adapté avec l'algorithme espérance-maximisation, classification sémantique basée sur des distances WordNet ou K-means avec analyse syntaxique
- Limites
  - performance moins bonne que les approches supervisées
  - un expert doit spécifier la sémantique des classes obtenues
  - ne permet pas de cibler des relations a priori
- Avantages
  - independant du domaine, de la langue
  - independant du type des relations
  - traitement de gros volumes de données
  - bas coût de mise en œuvre (pas d'annotation)

Réseaux de neurones pour le TAL

- Pour l'extraction de relations syntaxiques
  - réseaux de neurones profonds pour extraire des traits lexicaux des phrases [Zeng et al., 2014, Nguyen and Grishman, 2015]
  - apprendre les relations à partir du chemin de dépendances le plus court [Xu et al., 2015]
- Self-Organizing Map (SOM), traditionnellement utilisé pour la visualisation, appliqué en TAL à :
  - la classification de documents [Chifu and Cenan, 2004], [Henderson et al., 2002]
  - la résolution des co-références [Burkovski et al., 2011]
  - l'extraction de relations [Bloehdorn and Blohm, 2006].
- Intérêt : ne pas fixer a priori le nombre, la nature des relations

Plan de la présentation

Méthodologie adoptée

Une méthodologie commune aux expérimentations réalisées

- Extraction et préparation du corpus (traitement linguistique de base)
- Extraction de termes et balisage par les termes
- Préparation de l'extracteur
  - Entraînement de la méthode ou définition de patrons
  - Evaluation de l'extracteur sur un jeu de données
- Application de la méthode au corpus

Expérimenter 4 approches complémentaires

- Patrons lexico-sémantiques de la relation d'hyperonymie
- Apprentissage supervisé sur des structures énumératives
- Apprentissage non supervisé : K-means et SOM

Table – Nos expérimentations

Expé.	**Corpus**	**Langue**	**Algorithme**	**Relations**
Expé. 1	Agro., parag.	Fr	Patrons	Hyper.
Expé. 2	Biomed, SE	GB	Max Ent	Hyper.
Expé. 3	Biomed, parag	GB	K-means, SOM	Tout type

Plan de la présentation

Construction du corpus

Source : collection ISTEX
Anglais : biomedical, 2000-2010
Français : agroalimentaire, 100 documents
Sous-corpus : regroupement des éléments de structure identiques

Table – Statistiques des sous-corpus (anglais)

Corpus	**N unités**	**N phrases**	**N mots**
Corpus-Biomedical-Paragraph-EN	17125	81711	2383850
Corpus-Biomedical-SE-EN	81	509	18473
Corpus-Biomedical-Table-EN	53	-	491
Corpus-Biomedical-Figures-EN	31	49	323

Les structures dans les documents

En effet, la plupart des fichiers xml ou tei existants ne contiennent en fait que des méta-données (titre, auteurs, références, etc.). Aussi les corpus que nous avons sélectionnés se limitent aux quelques articles dont tout le texte est au format xml ou tei. La seconde difficulté concerne les textes en français. Ces textes, sans doute résultant d'une conversion pdf/txt ou pdf/xml à l'aide d'outils comme pdftoxml et pdftxt, ne comportent pas de caractères accentués.

Partant de ces observations, nous avons quand même pu exploiter 2 corpus :

- Un corpus en langue française : ce corpus appartient au domaine de l'agro-alimentaire et nous a été fourni par le laboratoire LINA impliqué également dans le projet ISTEX. Ce corpus est constitué de 159 articles au format tei qui permet de formaliser certains éléments de structure, dont les structures énumératives. Ce corpus sera référencé dans la suite de ce document par `Corpus-Agro-Alimentaire-FR`.

- Un corpus en langue anglaise : ce corpus appartient au domaine biomédical, et est composé de 988 articles issus de la revue *Nature* parus entre les années 2000 et 2012. Le choix de laisser de côté les articles dernièrement parus a été motivé par le fait que la connaissance contenue dans des articles récents est probablement nouvelle, et que cette connaissance n'a pas encore été formalisée au sein de ressources sémantiques, ressources qui dans ce contexte sont utiles pour l'évaluation. Ce corpus est en langue anglaise, car toujours en rapport avec l'évaluation, nous faisons le constat que très peu de ressources en français existent, notamment dans le domaine scientifique. Ce corpus sera référencé dans la suite de ce document par `Corpus-Biomedical-EN`.

Comme décrit dans la méthodologie, ces corpus ont fait l'objet de prétraitements classiques, et des sous-corpus spécifiques à certains types de structures ont été construits :

- le sous-corpus `Corpus-Agro-Alimentaire-SE-FR`, constitué structures énumératives verticales issues de `Corpus-Agro-Alimentaire-FR`
- le sous-corpus `Corpus-Biomedical-Paragraph-EN` constitué des paragraphes issus de `Corpus-Biomedical-EN`
- le sous-corpus `Corpus-Biomedical-SE-EN` constitué des structures énumératives issues de `Corpus-Biomedical-EN`
- le sous-corpus `Corpus-Biomedical-Table-EN` constitué des tableaux issus de `Corpus-Biomedical-EN`
- le sous-corpus `Corpus-Biomedical-Figures-EN` constitué des entêtes et légendes issus de `Corpus-Biomedical-EN`

Nous donnons ci-dessous des informations quantitatives sur ces sous-corpus :

Table 2: Statistiques des sous-corpus

Corpus	N unités	N phrases	N mots
Corpus-Biomedical-Paragraph-EN	17125	81711	2083850
Corpus-Biomedical-SE-EN	81	909	18473
Corpus-Biomedical-Table-EN	53	-	491
Corpus-Biomedical-Figures-EN	31	49	323

Ces différents corpus/sous-corpus ont été utilisés lors des différentes expérimentations que nous avons menées. Le corpus `Corpus-Agro-Alimentaire-FR` a été utilisé pour évaluer une approche par patrons (section 6), le sous-corpus `Corpus-Agro-Alimentaire-SE-FR` a été utilisé pour évaluer un système d'extraction de relations d'hyponymies sur un corpus scientifique (section 7), et le sous-corpus `Corpus-Biomedical-Paragraph-EN` a été utilisé pour expérimenter une approche par apprentissage non supervisé dans le cas de l'extraction de la relation d'hyperonymie, et dans le cas d'extraction de relations ouverte (section 8).

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 – 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)

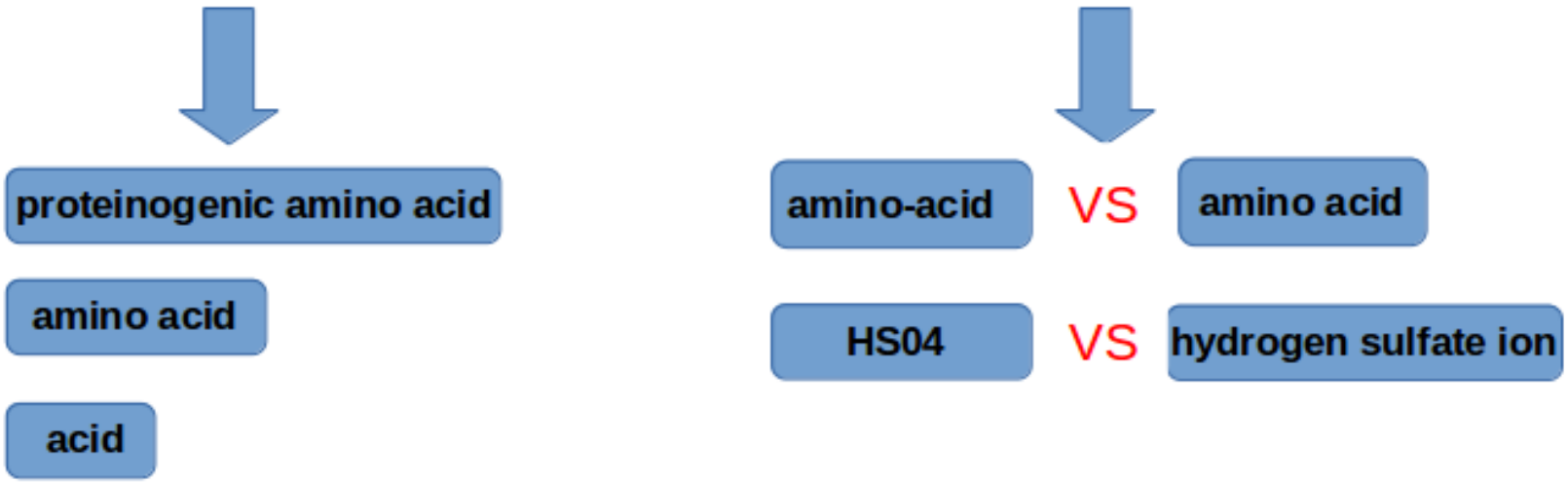
- Structures énumératives
  We detected the presence of the following **metals** :
  1. **lithium** ($\approx 0.56 mgr/unit$)
  2. **sodium** ($\approx 0.04 mgr/unit$)
  3. **zink** ($\approx 0.014 mgr/unit$)
  4. **potassium** ($\approx 0.001 mgr/unit$)
  5. etc.

- Images et figures (légendes)

Figure – **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

Identifier les candidats

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, pas présents dans les ontologies

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'algorithme et voir les resultats

Patrons Lexico-syntaxiques pour extraire les hyperonymies

Données

- Corpus : Agroalimentaire en Français
- Termes : termes extraits par TermSuite et Yatea

Patrons

- 34 patrons "génériques" de (Jacques et Aussenac, 2006)
- implémentés par exp. regulières en Python
- 41 relations trouvées

```
(
        ( { Token.category == "ADJ" } | { Token.category == "NOM" } )
        { Token.string == "," }
        ( { Token.category == "DET:ART" } | { Token.category == "DET:POS" } | { Token.category == "PRP:det" }
        )
        (   { Token.category == "PRP:det" } | { Token.category == "PRP" } | { Token.category == "NOM" } |
            { Token.category == "ADJ" } | { Token.category == "DET:ART" } | { Token.category == "DET:POS" } |
            { Token.category == "KON" } | { Token.category == "ADV" }
        )*
        (   ( { Token.lemma == "par" } { Token.lemma == "exemple" } ) |
            ( { Token.lemma == "en" } { Token.lemma == "particulier" } ) |
            ( { Token.lemma == "avant" } { Token.lemma == "tout" } )
        )
)
```

Patrons Lexico-syntaxiques pour extraire les hyperonymies

Evolutions - améliorations

- Définir "à la main" des patrons adaptés au corpus
- Projeter les couples de termes en relation trouvés par Expé 1 (ou d'autres méthodes) et observer des contextes pour définir de nouveaux patrons
- Expertise nécessaire pour la validation : utiliser une ontologie ou BC du domaine

Apprentissage supervisé d'hyperonymies à partir de structures énumératives

Données

- Corpus : Structures Enumératives dans le corpus Agroalimentaire en français
- Analyseur syntaxique : Talismane
- Termes : termes extraits par TermSuite et Yatea
- Données d'évaluation : 15 articles annotés à la main

Algorithme et traits

- Max-ent pour classer les paires + A* pour trouver les termes en relation
- Traits définis par J.P Fauconnier : POS, lemmes, présence de certains mots, distance entre mots, ...

Apprentissage supervisé d'hyperonymies à partir de structures énumératives

Table – Résultats en termes de Précision, Rappel et F-mesure

Précision	Rappel	F-mesure
0.38	0.61	0.47

Evolutions prévues

- Nouvel entraînement sur ce corpus
- Nouvelles heuristiques pour sélectionner les termes arguments des relations
- Autre type de relation hiérarchique

Expériences 3

Apprentissage non supervisé

- Deux algorithmes testés
  - SOM
  - K-means
- Les données
  - corpus de paragraphes, domaine de la biologie
  - transformer le corpus en un ensemble d'objets à classer (vecteurs)
- L'évaluation avec une ressource NCTI

Construction des instances d'entraînement

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
- max 5 tokens

Représentation vectorielle des paramètres

Objectif : Modeliser **le contexte semantique**

Hypothèse : **Le contexte semantique** characterise les differents **types** des relations

La base : répresenter le contexte avec word2vec [1]

Contexte :

- lemmes du contexte => **patrons lexicaux** :
  - 3 unités à gauche, droite et au milieu des candidats
- + Contexte POS => **patrons morphologiques** :
  - 3 unités à gauche, droite et au milieu des candidats

La représentation vectorielle de chaque couple ⇒ beaucoup de dimensions ⇒ utiliser SOM pour réduire la dimensionnalité

1. Mikolov et al.,2013

Apprentissage non-supervisé : clustering

Objectif : découvrir des patrons lexico-morpho-syntaxiques d'une façon non-supervisée 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 compétitif** (vector quantization) vs. apprentissage par correction d'erreurs (backpropagation avec gradient descent)

Self-organizing maps

Le réseau de neurones artificiels :

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

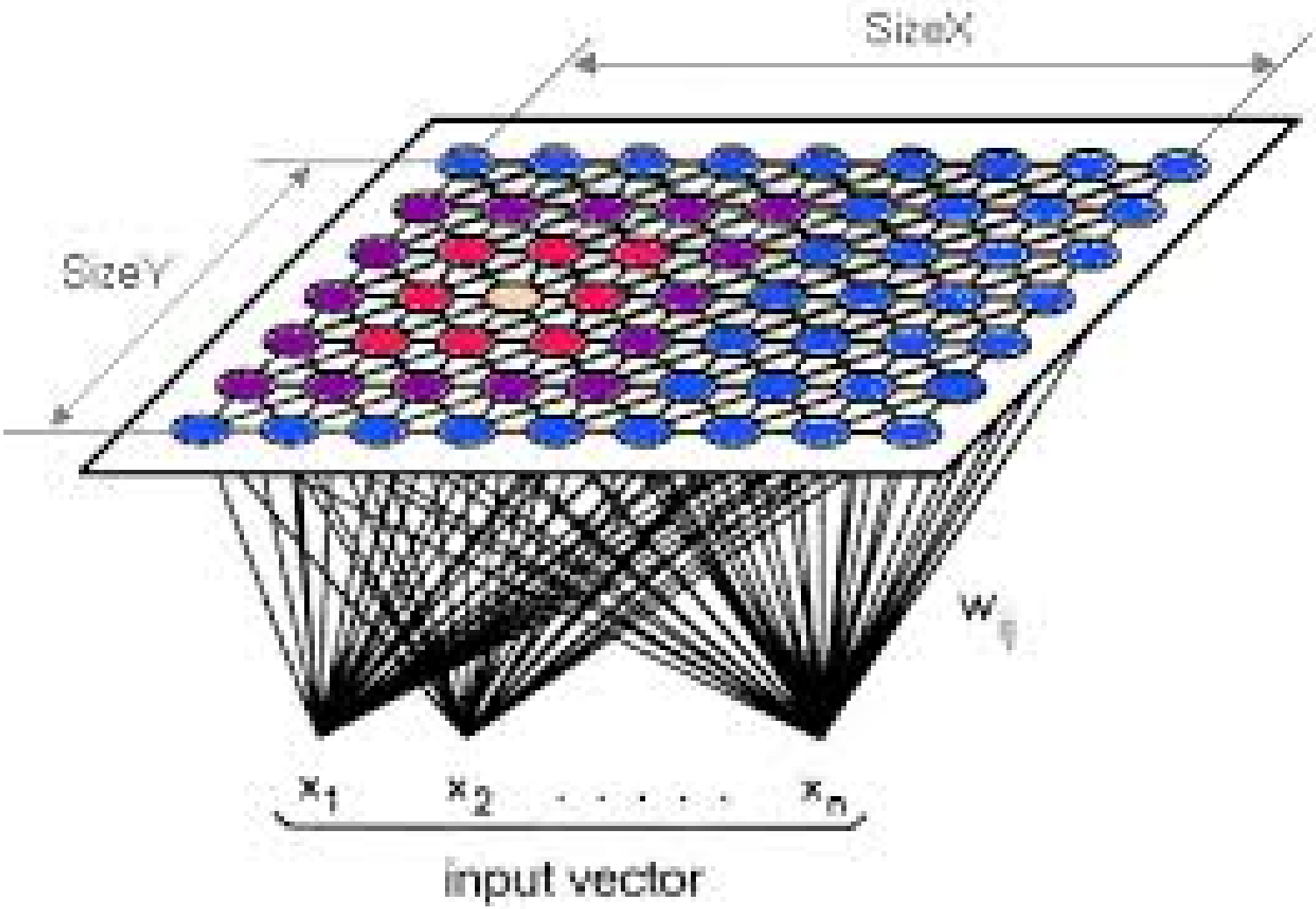
un neuron v avec le vector des poids $W_v(s)$:

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

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 et répéter la procedure à partir de l'étape 2 si $s < \lambda$

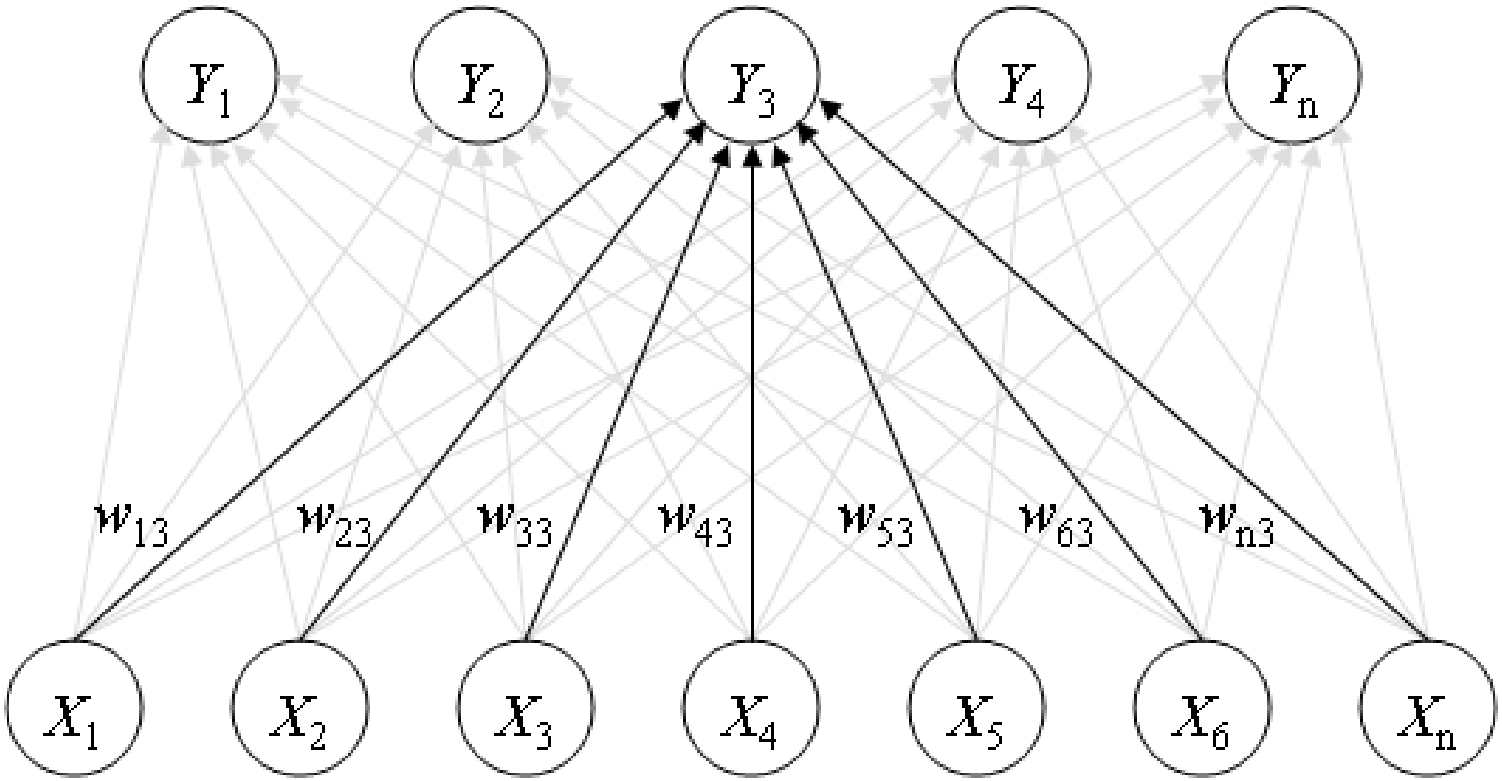
SOM training



SOM training. [2] : BMUs et voisins

2. https ://www.pitt.edu

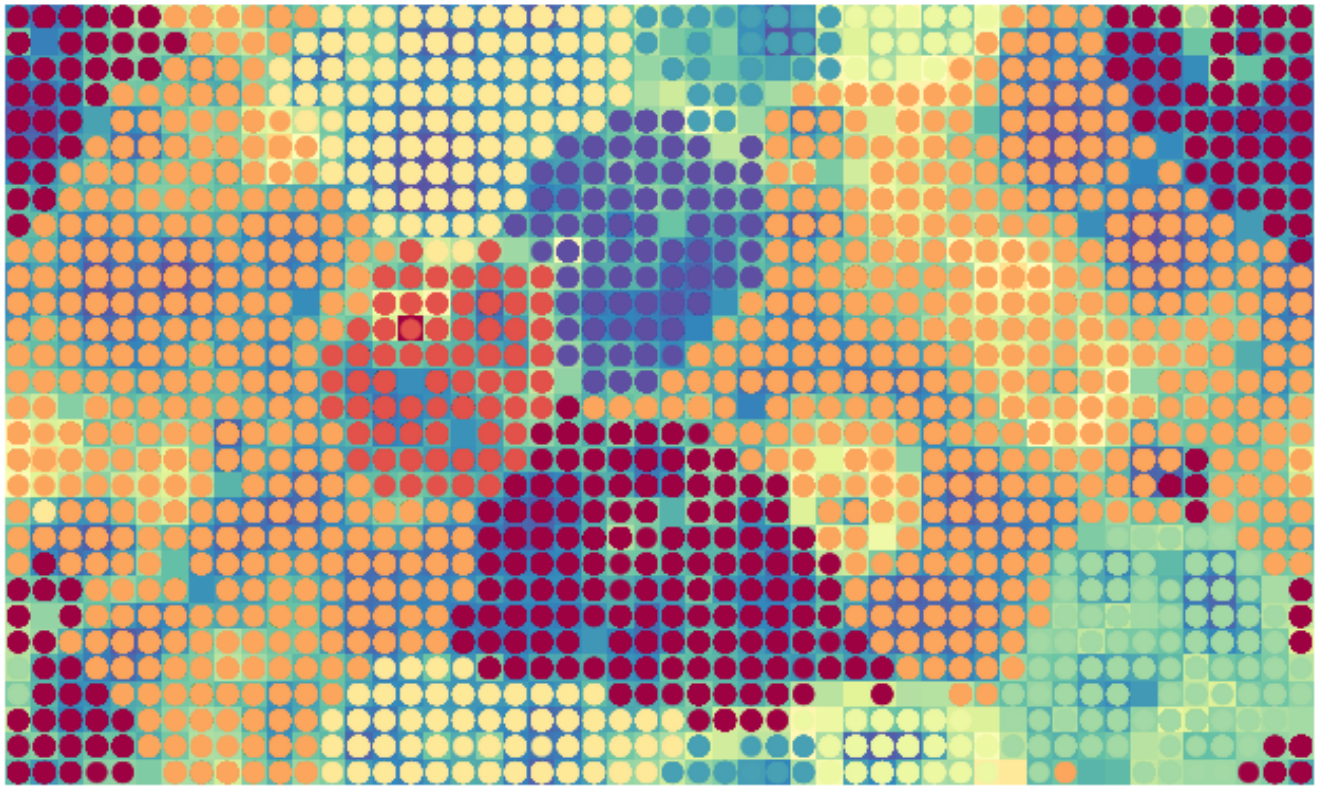
SOM training



SOM training [3] : les vecteurs des poids

3. https ://www.mnemstudio.org

Evaluation : 30x50 map



Paramètres testés

SOM

- taille du SOM (30x50 ; 70x90 ; 120x150)
- fenêtre contextuelle (3 mots VS 5 mots)
- w2v lemmes, POS, lemmes+POS
- vecteurs concatenés vs. vecteurs de valeurs moyennes
- decoupage en clusteurs forcé (8 clusteurs VS 30 clusteurs)

idem pour le **K-means**

Evaluation

National Cancer Institute Thesaurus [4]

118941 classes, 109 relations :

- Is_a
- Anatomic_Structure_Has_Location
- Anatomic_Structure_Is_Physical_Part_Of
  Gene_Prod_Plays_Role_In_Bio_Process
- Biological_Process_Has_Associated_Location
- ...

Dans notre corpus :

- entités trouvées : **24K**
- relations présentes : **83**
- couverture : **76%** (couverture MESH : **54%**)

4 : http ://ncicb.nci.nih.gov/core/EVS

Evaluation

Table – F-score : SOM, 30x50, lem

rel name	P	R	F
EO_Disease_Has_Property_Or_Attribute	0.23	0.45	0.3
Gene_Product_Has_Organism_Source	0.25	0.54	0.34
Gene_Plays_Role_In_Process	0.45	**0.85**	0.59
Conceptual_Part_Of	0.45	0.43	0.44
Procedure_Has_Excised_Anatomy	0.37	0.53	0.43
Gene_Prod_Plays_Role_In_Bio_Process	0.34	**0.92**	0.5
Procedure_Has_Target_Disease	0.42	0.18	0.25
Procedure_Uses_Manufactured_Object	0.38	0.50	0.43

Evaluation

Table – F-score : SOM vecteurs moyens, SOM concatenés et K-means concatenés (30x50, 3 tokens)

rel name	F1	F2	F3
EO_Disease_Has_Property_Or_Attribute	0.3	0.25	0.11
Gene_Product_Has_Organism_Source	0.34	0.31	0.24
Gene_Plays_Role_In_Process	**0.59**	**0.22**	0.11
Conceptual_Part_Of	0.44	0.43	0.18
Procedure_Has_Excised_Anatomy	**0.43**	**0.52**	0.25
Gene_Prod_Plays_Role_In_Bio_Process	0.5	0.33	**0.08**
Procedure_Has_Target_Disease	0.25	0.21	0.14
Procedure_Uses_Manufactured_Object	0.43	0.46	0.1

Evaluation

Table – F-score : lem+POS, vecteurs concatenés, par cluster

Méthode	PUM	PHA	GPRP	GPR	GP0S
Kmeans	0.1	0.25	0.11	0.08	0.24
30x50 SOM	0.46	0.52	0.22	0.33	0.31
70x90 SOM	**0.49**	0.52	**0.38**	0.31	**0.36**

Plan de la présentation

Conclusions

ISTEX est-il un "bon corpus" pour le TAL ?

- Beaucoup de fichiers txt bruités, à nettoyer
- Problèmes liés au balisage : HTML dans le txt, pas assez de balises dans le XML ou le TEI
- problème d'accentuation en français

L'apprentissage non supervisé est-il adapté à l'extraction de relation dans des corpus scientifiques ?

- Résultats moins bons que sur "très grand" corpus (Wikipedia)
- Résultats prometteurs car on trouve des classes de relations

Vers des outils modulaires et complémentaires

Poursuivre les recherches

- Mieux analyser les résultats pour tester une vraies complémentarité (cf travaux de R. Granada et C. Trojahn)
- Tester les apports des résultats d'une approche comme entrée ou traits pour une autre méthode
- Définir des enchaînements de techniques

Mieux intégrer dans ISTEX

- Vers des outils opérationnels et diffusables
- Complémentarité avec autres projets : TermSuite, TERRE-ISTEX, ...
- cf Latoe et Lara de JP Fauconnier sur Github

References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007).
Dbpedia : A nucleus for a web of open data.
In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

Bloehdorn, S. and Blohm, S. (2006).
A self organizing map for relation extraction from wikipedia using structured data representations.
In *Proc. Int. Workshop on Intelligent Information Access (IIIA-2006)*.

Burkovski, A., Kessler, W., Heidemann, G., Kobdani, H., and Schütze, H. (2011).
Self organizing maps in nlp : Exploration of coreference feature space.
In *International Workshop on Self-Organizing Maps*, pages 228–237. Springer.

Chifu, E. S. and Cenan, C. (2004).
Discovering web document clusters with self-organizing maps.
Sci. Ann. Cuza Univ., 15 :38–47.

Govindaraju, V., Zhang, C., and Ré, C. (2013).
Understanding tables in context using standard nlp toolkits.

References (cont.)


Henderson, J., Merlo, P., Petroff, I., and Schneider, G. (2002).
Using nlp to efficiently visualize text collections with soms.
In *Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on*, pages 210–214. IEEE.


Jannink, J. (1999).
Thesaurus entry extraction from an online dictionary.
Technical report, Standford University.


Kamel, M. and Aussenac-gilles, N. (2009).
can document structure improve ontology learning ? (regular paper.
In *Semantic Authoring, Annotation and Knowledge Markup Workshop - collocated with K-CAP 2009 (SAAKM 2009*, pages 1–8.


Nguyen, T. H. and Grishman, R. (2015).
Relation extraction : Perspective from convolutional neural networks.
In *Proceedings of NAACL-HLT*, pages 39–48.


O'Connor, M. J. and Das, A. (2011).
Acquiring owl ontologies from xml documents.
In *Proceedings of the Sixth International Conference on Knowledge Capture*, K-CAP '11, pages 17–24, New York, NY, USA. ACM.

References (cont.)


Sumida, A. and Torisawa, K. (2008).
Hacking wikipedia for hyponymy relation acquisition.
In *IJC-NLP*, number 8, pages 883–888.


Xu, K., Feng, Y., Huang, S., and Zhao, D. (2015).
Semantic relation classification via convolutional neural networks with simple negative sampling.
arXiv preprint arXiv :1506.07650.


Zeng, D., Liu, K., Lai, S., Zhou, G., Zhao, J., et al. (2014).
Relation classification via convolutional deep neural network.
In *COLING*, pages 2335–2344.

Exploitation de la structure de documents en TAL

- La structure de document exploitée conjointement au texte rédigé pour extraire et représenter des connaissances
- Structures exploitées :
  - dictionnaires ou thesaurus [Jannink, 1999]
  - tableaux [Govindaraju et al., 2013],
  - structures énumératives
  - infoboxes Wikipédia [Auer et al., 2007]
- Structure explicite de collections HTML ou XML
  - WikiText pour construire des taxonomies [Sumida and Torisawa, 2008],
  - pour construire des ontologies [Kamel and Aussenac-gilles, 2009, O'Connor and Das, 2011], des BC comme DBpedia