CÔNG CỤ X.ENT CHO TRÍCH XUẤT DỮ LIỆU THỰC THỂ, QUAN HỆ GIỮA THỰC THỂ VÀ HỖ TRỢ PHÂN TÍCH DỮ LIỆU TRONG CÁC TẠP CHÍ VỀ PHÒNG CHỐNG DỊCH BỆNH TRONG NÔNG NGHIỆP CỦA PHÁP

Phan Trọng Tiến*, Ngô Công Thắng

Khoa Công nghệ Thông tin, Học viện Nông nghiệp Việt Nam

Email: ptgtien@vnua.edu.vn*

Ngày gửi bài: 22.07.2015 Ngày chấp nhận: 03.09.2015

TÓM TẮT

Trích xuất thực thể là công việc trích xuất thông tin và phân loại thông tin trong văn bản theo những loại xác định trước như tên người, tổ chức, địa điểm, thời gian,... và một bước cao hơn là tìm mỗi quan hệ giữa các thực thể ví dụ như mỗi quan hệ giữa tên người với tên tổ chức. Công cụ x.ent được xây dựng để làm công việc như vậy, công cụ sử dụng các từ điển cho thực thể và các luật để trích xuất. Trong trích xuất quan hệ giữa các thực thể chúng tôi áp dụng hai phương pháp: phân tích cấu trúc của văn bản và sử dụng mô hình học không giám sát đó là phân tích tần suất xuất hiện của các thực thể. Công cụ x.ent có sẵn trên trang chủ R theo đường dẫn: http: //cran.r - project.org/web/packages/x.ent/index.html.

Từ khoá: Automat hữu hạn, nhận biết thực thể định danh, Perl, R, trích xuất thông tin, trích xuất thực thể, trích xuất quan hệ.

X.ent Package for Extraction of Entities, Relationships between Entities and Support Data Analysis in Epidemiological Journals in French Agriculture

ABSTRACT

Entity extraction is a task of information extraction and element classification in text such as the names of persons, organizations, locations, times, etc. and to find relationship between entities such as the relationship between the names of persons with the organizations. The X.ent tool was built solve this task. It uses dictionaries matching and hand - crafted rules to extract. In extracting the relationship between the entities, we applied two methods: analysis of text structures and unsupervised learning approach called coo – ccurrence analysis. This tool is available on the site of R at the links: http: //cran.r - project.org/web/packages/x.ent/index.html.

Keywords: Entity Extraction, Information Extraction (IE), Named entity recognition (NER), Perl, Relation Extraction, R.

1. ĐẶT VẤN ĐỀ

Chúng ta đang sống trong thời đại bùng nổ về công nghệ thông tin, theo thống kê, mỗi ngày có 540 triệu tin nhắn văn bản được gửi đi trên toàn thế giới, 143 tỷ email được trao đổi, 40.000 gigabyte dữ liệu được tạo ra bởi Máy gia tốc hạt lớn (LHC - Large Hadron Collider), 400 triệu cập nhật trạng thái trên trang mạng xã hội Twitter được đăng, 104.000 giờ video được thêm vào YouTube, v.v. (theo NASATI) và nó còn tiếp tục tăng lên trong thời gian tới.

Việc xử lý và phân tích dữ liệu lớn dựa trên những nghiên cứu trong nhiều lĩnh vực bao gồm khoa học máy tính, thống kê, toán học, kỹ thuật dữ liệu, nhận dạng mẫu, trực quan hóa, trí tuệ nhân tạo, máy học và tính toán hiệu năng cao.

Với lượng dữ liệu rất lớn, nó có thể chứa cả những thông tin dư thừa, vì vậy việc trích xuất

thông tin (IE) là một bước rất quan trọng để lấy được ra những thông tin cần thiết cho việc phân tích dữ liệu. Hiện nay trích xuất thông tin được sử dụng trong rất nhiều lĩnh vực ứng dụng như để tìm hiểu về xu hướng kinh doanh chủ yếu của người dùng, ngăn ngừa bệnh tật, phòng chống tội phạm, lĩnh vực tin sinh học, phân tích chứng khoán, v.v.

X.ent là một công cụ được chúng tôi xây dựng cho việc trích xuất dữ liệu văn bản (trích xuất thực thể và quan hệ giữa các thực thể), ngoài ra chúng tôi còn xây dựng một số tính năng bằng đồ hoạ được viết trên R để cung cấp cho người sử dụng các tính năng phân tích dữ liệu sau khi trích xuất. Công cụ này là sự kết hợp các ngôn ngữ lập trình khác nhau: Perl cho phần trích xuất dữ liệu, R cho việc hỗ trợ phân tích kết quả. Sau khi hoàn thành chúng tôi đã gửi công cụ của chúng tôi lên trang chủ của CRAN (là một trang chứa các gói ứng dụng của R) và được các chuyên gia thống kê học ở đây chấp nhận, hiện tại người sử dụng có thể tải về và cài đặt trực tiếp từ máy chủ CRAN. Đây là sản phẩm được tôi hoàn thành trong quá trình học cao học tại Pháp năm 2012 - 2014.

2. VẬT LIỆU VÀ PHƯƠNG PHÁP

2.1. Vật liệu

Dữ liệu được chúng tôi trích xuất là các báo cáo về phòng chống dịch bệnh cho cây trồng của Pháp, có 12 thực thể chúng tôi quan tâm là cây trồng (crops), bệnh (diseases), sinh vật phá hoại (pests), các sinh vật có lợi khác (auxiliaries), vị trí địa lý (regions, towns), ngày tháng của báo cáo (date), số của báo cáo (issues), hoá chất sử dụng (chemicals), các giai đoạn phát triển cây trồng (developmental stage), sự gây hại với cây trồng (crop damage), khí hậu (climate), mức độ tiêu cực (negative). Các quan hệ giữa các thực thể mà chúng tôi quan tâm: cây trồng với bệnh và cây trồng với sinh vật phá hoại.

Ở Pháp, hàng tuần các nhà nông học sẽ tạo các báo cáo để thông tin cho người nông dân về các tấn công của dịch bệnh và côn trùng đối với cây trồng. Mục tiêu của các báo cáo này là khuyến khích người nông dân sử dụng các phương pháp điều trị để chống lại các sinh vật gây hại. Ấn bản đầu tiên được ra đời vào năm 1946 và đều là các bản đánh máy (bản in), từ năm 2001 tất cả các ấn bản được xuất bản theo định dạng PDF. Pháp được chia làm 22 vùng và các vùng nước ngoài, mỗi vùng sẽ xuất bản các báo cáo riêng.

Nguồn dữ liệu của dự án có 50.000 bản báo cáo, trong đó có khoảng 20.000 là dạng các trang in. Chúng tôi cần scan các bản giấy này và nó được chia sẻ tại thư viện BNF (Bibliothèque François - Mitterrand) và sau đó được chuyển đổi sang dạng text nhờ kỹ thuật OCR (Optical Character Recoginition) bởi Jouve Corp.

Đây là dự án được tài trợ bởi Bộ Nông nghiệp và Nghiên cứu Pháp, dự án bao gồm các chuyên gia sinh vật học và sinh thái học nghiên cứu các tác nhân gây bệnh: dịch tễ học và khoa học môi trường (các dự báo về sâu bệnh) với một mạng lưới gọi là PIC (Intergrated Crop Protection). Có 4 chuyên gia về khoai tây và lúa mì từ PIC đồng hành cùng chúng tôi trong dự án này, dự án có tên VESPA (Valeur et optimisation des dispositifs d'épidémiosurveillance dans une stratégie durable de protection des cultures - Ước lượng và tối ưu hoá các thiết bị giám sát dịch tễ học trong chiến lược bảo vệ sự bền vững cho cây trồng).

2.2. Phương pháp

Trích xuất thông tin (IE) là một tác vụ tự động trích xuất để có được thông tin có cấu trúc từ các tài liệu không cấu trúc hoặc bán cấu trúc mà máy tính có thể đọc được. Trong hầu hết các trường hợp, hoạt động này liên quan đến xử lý các văn bản ngôn ngữ con người hay nói cách khác là xử lý ngôn ngữ tự nhiên (Natural Language Processing)

Mục tiêu chính của chúng tôi là trích xuất quan hệ giữa thực thể cây trồng với các tác nhân gây hại cho cây trồng cùng với mức độ gây hại của chúng. Trích xuất thông tin là một công cụ tốt trong xử lý ngôn ngữ tự nhiên. Các bước thực hiện trong xử lý dữ liệu trích xuất thông tin:

Toute reproduction même partielle est soumise à notre autorisation

Grandes Cultures

Service Régional de la Protection des Végétaux Z.I. de Thise - BP 989 25022 - BESANCON Cedex Tél : 03 81.80.10.71 - Fax : 03 81.88.27.77

Imprimerie du SRPV Bourgogne - Le Directeur-Gérant : J.C. RICHARD - CPPAP n° 1700 AD - © SRPV 1986 - ISSN : 0758-2374

Service Régional de la Protection des Végétaux Z.I. Nord - BP 177 21205 - BEAUNE Cedex Tél : 03 80.26.35.45 - Fax : 03 80.22.63.85

TARIF COURRIER 320 F - FAX 370 F. - Page 1

Bourgogne et Franche-Comté

Bulletin n° 27/97 - 28 octobre 1997

D LP 3 1 -10- 9 7 0 3 1 8 4 1

COLZA

Stades : 6 à 10 feuilles.

Ravageurs

Les **pucerons** sont encore très présents. L'effectif est soit en stabilisation pour des populations importantes soit en progression en situations d'attaques moyennes ou de recolonisation après traitement. L'arrivée de conditions plus froides devrait enfin faire régresser les populations.

Essais pucerons d'automne sur colza : voir premiers résultats page suivante.

Quelques captures de **charançon du bourgeon terminal** sont encore observées mais le vol s'est essentiellement regroupé sur la période du 16 au 22 octobre. Les larves d'**altises** sont rares sur pétioles. Des comptages larvaires faisant un bilan altises-charançon du bourgeon terminal seront réalisés ultérieurement.

Au cours des dernières semaines, des larves de **mouche du chou** (asticots blancs aux collets et sur les pivots des colzas) ont été signalées en quelques endroits : Saponcourt, Delain (70), Mirebeau, Seurre (21)... Ces dégâts sont liés essentiellement à la précocité des dates de semis et/ou de levée. Aucune lutte n'est possible en post-levée de la culture.

Préconisations : Les interventions contre le charançon et/ou les pucerons ont dû être réalisées. Compte-tenu de la baisse des températures, peu favorable à l'activité des insectes, ne plus envisager d'intervention insecticide sur colza.

Désherbage de rattrapage sur sanves et ravenelles

Après les premières gelées hivernales qui provoquent la mise en place de la cuticule cireuse du colza, l'utilisation de Cent 7 est possible en rattrapage contre les sanves et les ravenelles (les autres crucifères sont peu sensibles au produit). L'efficacité, optimale lorsque le traitement est effectué au stade rosette des crucifères adventices, décroit à partir du stade montaison de celles-ci (souvent atteint cet automne) mais le produit sensibilise néanmoins les plantes au gel hivernal.

. Dose d'emploi du Cent 7 : 0,4 l/ha

Précautions : proscrire toute association avec un autre produit phytosanitaire ; traiter lors d'un radoucissement.

CEREALES

Stades : Selon les dates de semis, les cultures peuvent atteindre le début tallage, voire 2-3 talles pour les situations les plus avancées. Les semis de début octobre sont à leur 3ème feuille.

Ravageurs

Le vol de **pucerons** se ralentit sensiblement avec l'abaissement des températures (voir ci-après le graphe des captures à la tour d'Auxerre).

Les comptages réalisés en cultures donnent des résultats variables selon les secteurs et le stade des cultures. Le grand Val de Saône présente les populations les plus importantes (Saône-et-Loire, Jura, secteur Val de Saône et début plaine Dijonnaise en Côte-d'Or). Des niveaux atteignant 10 à 15 % de pieds porteurs sur cultures à 2 feuilles et au-delà peuvent être observés. Les autres secteurs présentent des niveaux variables mais généralement inférieurs, ne dépassant pas 3 à 5 % de pieds porteurs pour les mêmes stades.

Les **cicadelles** restent présentes dans les secteurs favorables de l'Yonne et de la Nièvre où les captures dans le réseau de cuvettes n'ont pas diminué.

Préconisations :

➢**Pour les parcelles atteignant 3 feuilles-début tallage, en tous secteurs,** les niveaux de population observés et la durée de présence des pucerons justifient une intervention rapide.

➢**Pour les parcelles de Saône-et-Loire, Jura et Sud Est Côte-d'Or actuellement à 2 feuilles-3ème feuille pointante,** l'intervention est aussi à réaliser très prochainement.

➢**Pour les parcelles des autres secteurs au même stade,** les populations observées dépassent rarement 3 à 5 % de pieds porteurs. L'intervention peut être encore différée.

➢**Pour les parcelles n'ayant pas atteint 2 feuilles,** toute intervention est actuellement prématurée. La surveillance doit cependant se maintenir dans le grand Val de Saône.

Pour les situations de l'Yonne et de la Nièvre, exposées au risque cicadelles, il convient de rester attentif aux fortes populations sur les jeunes levées et intervenir le cas échéant, même sur des stades jeunes.

COLZA

Plus d'intervention contre les ravageurs.

CEREALES

Protection anti-pucerons selon les secteurs et les stades.

Hình 1. Báo cáo về dịch bênh cây trồng vùng Bourgogne và Franche - Comté

Bước 1: Nhận biết các thực thể định danh (Named Entity Recognition - NER)

Bước 2: Trích xuất quan hệ

Bước 3: Trích xuất thông tin ngữ cảnh như mức độ gây hại, giai đoạn phát triển của cây trồng, khí hậu, địa lý...

Có rất nhiều giải thuật và phương pháp thực hiện trích xuất thực thể định danh (NER) như: các thuật toán về phân loại theo partern - based (dựa theo các quy luật trích xuất của các chuyên gia), các thuật toán về thống kê như HMM (Hidden Markov Model), MaXent (Maximum Entropy Modeling) hay CRF (Conditional Random Fields).

2.2.1. Trích xuất thực thể định danh

a. Sử dụng từ điển cơ sở

Khi trích xuất dữ liệu, có những thực thể chúng ta có thể xây dựng các từ điển của thực thể để thực hiện cho việc trích xuất, ví dụ từ điển về cây trồng (crops), bệnh (diseases), sinh vật phá hoại (pets), các sinh vật có lợi khác (auxiliaries), vị trí địa lý (regions, towns), hoá chất điều trị (chemicals). Các từ điển được chúng tôi xây dựng theo nguyên tắc sau: từ đầu là từ khoá gốc, sau đó phân loại của từ đó, N là gốc (node) của các loại khác, L là lá của từ loại đó (leaf), với một từ khoá gốc có thể có các dạng biến đổi của nó như dạng số ít, số nhiều, không dấu, từ đồng nghĩa, từ viết tắt, v.v.

b. Sử dụng các luật trích xuất

Có những loại thực thể mà chúng ta không thể xây dựng được từ điển cho thực thể đó, ví dụ như các giai đoạn phát triển của cây trồng, hay đánh giá mức độ gây hại với cây trồng hay là dữ liệu kiểu ngày tháng, v.v. Vì vậy chúng tôi phải xây dựng các luật trích xuất sử dụng công cụ Unitex, có thể xem tại địa chỉ http: //www – igm.univ – mlv.fr/~unitex/ (Paumier et al.), được phát triển bởi Đại học Paris – Est. Các luật trích

blé:N:blé:BLE:blés:Triticum:blé dur:blé tendre:
blé dur:L:BLE DUR:T. durum:Triticum durum:bles durs:blés durs:blé dur:
blé noir:L:BLE NOIR:f. esculentum:fagopyrum esculentum:sarrasin:bles noirs:blés noirs:blé noir:sarrasins:
blé tendre:L:BLE TENDRE:T. aestivum:Triticum aestivum:blé froment:blés froments:ble froments:blé tendre:blés tendres:bles tendres:

wheat *(species)*
durum wheat *(variety)*

buckwheat *(variety)*

soft wheat *(variety)*

	entities						
	auxiliaries	crops	pests	diseases	chemicals	regions	towns
#entries	28	114	373	275	4968	26	33161
#leafs	28	103	334	241	4968	26	33161
#concepts	0	18	53	40	0	0	0
#lexems	107	727	2673	1846	4968	869	89603

Hình 2. Cấu trúc từ điển và thống kê từ điển mà chúng tôi đã xây dựng

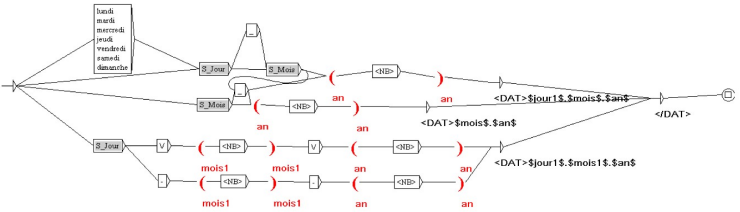
xuất này chính là các automat hữu hạn, được xây dựng bằng giao diện đồ hoạ. Ví dụ để trích xuất dữ liệu ngày tháng năm trong báo cáo, chúng tôi dựa theo cấu trúc dữ liệu ngày tháng trong các văn bản mẫu ví dụ chúng có định dạng "xx {January|February…} xxxx" thì chúng ta có thể xây dựng quy luật như hình 3.

Trong dự án này, với sự hỗ trợ của các chuyên gia về nông nghiệp chúng tôi xây dựng các luật trích xuất hay chính là ngữ pháp khác nhau cho việc luật trích xuất, có một số quy tắc để lấy được dữ liệu như sau:

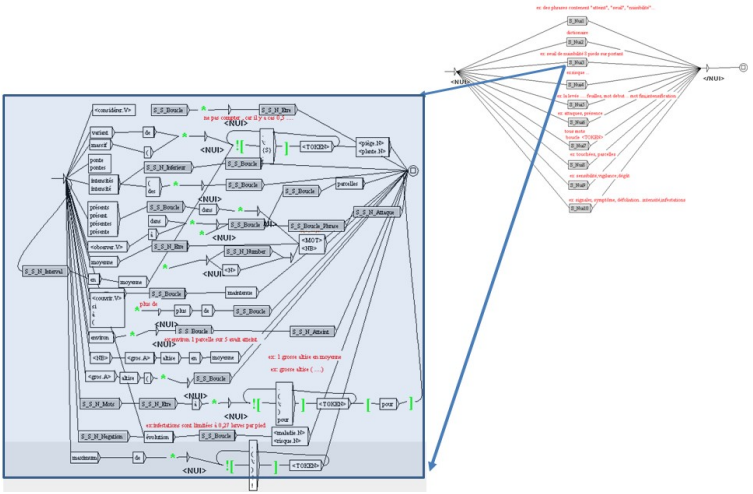
- < các từ trong từ điển>
- < từ khoá đánh dấu bắt đầu>…. < kết thúc câu>
- < từ khoá đánh dấu bắt đầu>…. < từ khoá đánh dấu kết thúc>
- < từ trong từ điển>… < từ khoá đánh dấu kết thúc>
- < từ khoá đánh dấu bắt đầu>… < từ trong từ điển>

2.2.2. Trích xuất quan hệ

Trích xuất quan hệ giữa các thực thể vẫn là bài toán tương đối phức tạp, có nhiều phương pháp trích xuất khác nhau đã được đề xuất như xây dựng luật trích xuất quan hệ, các phương pháp Bootstraping, Supervised, Distant Suppervision hay các phương pháp Unsupervised (Zettlemoyer, 2013). Chúng tôi đề xuất hai phương pháp trích xuất quan hệ: phương pháp phân tích cấu trúc tài liệu và phương pháp mô hình học không giám sát sử dụng tần suất xuất hiện dữ liệu của các thực thể (co – occurrence).



Hình 3. Luật trích xuất ngày tháng được xây dựng bằng công cụ Unitex



Hình 4. Ngữ pháp trích xuất đánh giá mức độ gây hại với cây trồng

a. Phân tích cấu trúc tài liệu

Tổ chức của một tài liệu (tiêu đề, tiêu đề con, phần tham chiếu, các phân đoạn, các bảng, các ảnh, phần giới thiệu, phần tổng kết, phần thảo luận) có thể ảnh hưởng tới việc trích xuất. Chúng tôi gọi đây là kiến trúc của một tài liệu. Tuy nhiên nhiều kiến trúc là có sẵn và tập các heuristics là không giới hạn.

Heuristics 1: Thực thể chính

Thực thể chính xảy ra ở vị trí tiêu đề hoặc tiêu đề con của đoạn hoặc của một phần của đoạn.

Trong hình 5 chúng ta nhìn thấy rằng thực thể chính xảy ra ở đầu của mỗi đoạn, trong ví dụ này là thực thể cây trồng (blé, betterave)

Heuristics 2: Lấy giá trị đầu tiên

Với các thực thể khác nhau, có thể trong dữ liệu chúng ta tìm thấy nhiều giá trị của thực thể đó, nhưng chúng ta chỉ lấy giá trị đầu tiên trong báo cáo đó.

Trong hình 5 chúng ta nhìn thấy các thực thể như vị trí địa lý, ngày xuất bản của báo cáo, số của báo cáo.

Heuristics 3: Vùng không tìm kiếm

Một vài đoạn trong văn bản có thể chứa các tiêu đề mà trong đoạn đó có thể có chứa các thực thể nhưng nó không có liên kết với thực thể chính hoặc thông tin của ngữ cảnh. Ví dụ như thông tin phụ trợ, hoặc chú thích hoặc thông tin được chích từ một nguồn dữ liệu khác.

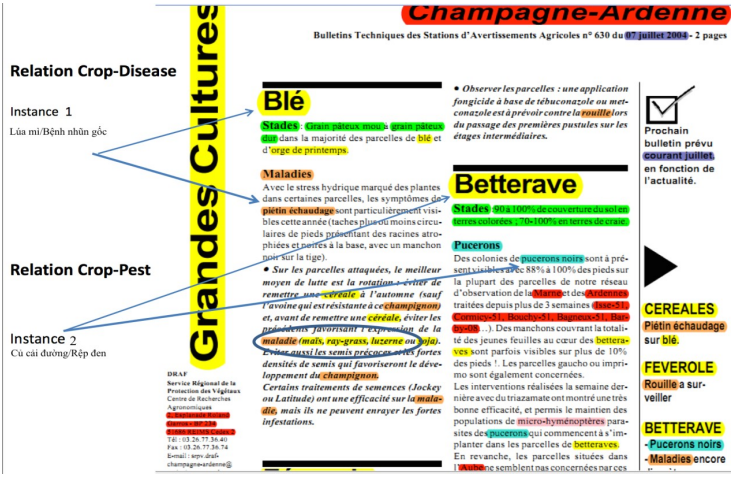
b. Mô hình học không giám sát sử dụng tần suất xuất hiện

Định nghĩa 1: Đơn vị văn bản và thực thể

Một đơn vị văn bản (TU) là một danh sách liên kết mà chứa các từ W và các thực thể E. Một thực thể có thể là một từ hoặc một tập các từ liên tiếp nhau.

Định nghĩa 2: Vị trí thực thể

Đặt E_i là một thực thể gốc. Một tài liệu được chia thành các đơn vị văn bản (TU). Một đơn vị văn bản có thể là một phần của một đoạn, một câu hoặc một đoạn văn. Gọi P_w^i là vị trí của các từ khoá và P_{TU}^i là tiêu đề của thực thể E_i trong tài liệu. Chúng ta định nghĩa một cửa sổ mà W_L là số từ tại vị trí bên trái từ P_w^i và W_R là số từ ở bên phải của P_w^i. W_R có giá trị là ∞ nghĩa là cửa sổ sẽ bắt đầu tại đầu của văn bản, tương tự như vậy W_L có giá trị là ∞, cửa sổ sẽ tới cuối của văn bản.



Hình 5. Chú thích bằng tay trong một tài liệu của dự án

Ghi chú: Màu vàng: cây trồng, màu xanh lá cây: các giai đoạn phát triển cây trồng, màu nâu: bệnh cây trồng, màu đỏ: vị trí địa lý, màu xanh da trời: sinh vật gây hại, màu tía: các sinh vật có lợi, màu xanh đen: thời gian

Kiểu 1: Tần suất xuất hiện của đơn vị văn bản. Đặt E_i là thực thể gốc và E_j là một thực thể khác. Chúng ta định nghĩa tần xuất xuất hiện bởi một hàm nhị phân cooc(E_i,E_j) như sau:

$$cooc(E_i,E_j) = \begin{cases} 1 \text{ nếu } P_w^i \in P_{TU}^i \text{ và } P_w^j \in P_{TU}^j \text{ và } P_w^i \text{ thoả mãn } heuristics\ 1,2 \text{ và } 3 \\ 0 \quad \text{trường hợp còn lại} \end{cases}$$

Kiểu 2: Tần suất xuất hiện của cửa sổ, giống như kiểu 1, nhưng thoả mãn:

$cooc(E_i,E_j) = 1$ nếu $(P_w^i - W_L) \le P_w^i \le (P_w^i + W_R)$

Kiểu 3: Các ràng buộc tần suất xuất hiện, giống như kiểu 1 hoặc kiểu 2. Nhưng đặt một danh sách các điểm đánh dấu m_k, ít nhất một điểm đánh dấu m_k cần nằm giữa E_i và E_j, vì vậy ta có:

$cooc(E_i, E_j) = 1$ nếu $\left|P_w^i - P_w^k\right| \le \left|P_w^i - P_w^j\right|$

2.2.3. Định dạng dữ liệu đầu vào và đầu ra

Kết quả trích xuất được lưu trữ theo định dạng giống định dạng CSV (hình 6 bên phải), đầu tiên là tên của tệp báo cáo, tiếp theo là ký hiệu của thực thể ("r" cho vùng, "p" cho cây trồng...) hoặc quan hệ (p: m là quan hệ giữa cây trồng và bệnh...), tiếp theo đó là dữ liệu trích xuất gắn với thực thể hoặc quan hệ mà chúng ta trích xuất được theo loại nào đó.

Ngoài ra để đánh giá độ hiệu quả của công cụ x.ent, chúng tôi so sánh kết quả trích xuất với các công cụ khác (http8, http9, 2014), chúng tôi phải biến đổi dữ liệu theo chuẩn của CoNLL (Conference on Natural Language Learning) cho các mô hình máy học sử dụng phương pháp thống kê. Chúng tôi phải thực hiện số hoá bằng tay 37 tệp để đánh giá kết quả. Định dạng dữ liệu (hình 6 bên trái) gồm hai cột: cột đầu tiên là các từ được cắt ra theo đúng thứ tự của các câu, cột thứ 2 là phân loại của từ đó, "O" là từ không thuộc phân loại nào, "PLA" là từ thuộc phân loại tên cây trồng, v.v.

Champagne-Ardenne		REG
.	O	
BSV	O	
du	O	
09/06/2011	O	
--	O	
semaine	O	
23	O	
A	O	
RETENIR	O	
CETTE	O	
SEMAINE	O	
.	O	
TOURNESOL	PLA	
:	O	
Pucerons	BIO	
:	O	
fin	O	
du	O	
risque	O	
.	O	
Absence	O	
de	O	
maladies	O	
.	O	
MAïS	PLA	
:	O	
Pyrale	BIO	

5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:r:$:CHAMPAGNE-ARDENNE:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:$:colza:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:b:$:charançon de la tige du colza:charançon de la tige du chou:Méligèthes:mouche du chou:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:d:$:25.03.2010:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:v:$:St Dizier:Reims:Charleville:Langres:TROYES:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:n:$:peu nuisible:à risque:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:s:$:c2:c1:d1:f1:
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:$$:colza:charançon de la tige du colza:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:$$:colza:mouche du chou:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:$$:colza:Méligèthes:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:s:$$:colza:c2:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:s:$$:colza:d1:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:s:$$:colza:c1:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:n:$$:colza:charançon de la tige du chou:peu nuisible:1
5_-_BSV_CHAMPAGNE-ARDENNE_COLZA_2010_03_25_cle09c576:p:b:n:$$:colza:mouche du chou:à risque:1

Hình 6. Định dạng đầu vào và đầu ra theo chuẩn CONLL và định dạng đầu ra của x.ent

3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Định giá kết quả trích xuất

Để đánh giá hiệu quả công cụ x.ent, chúng tôi so sánh kết quả trích xuất với các công cụ trích xuất khác.

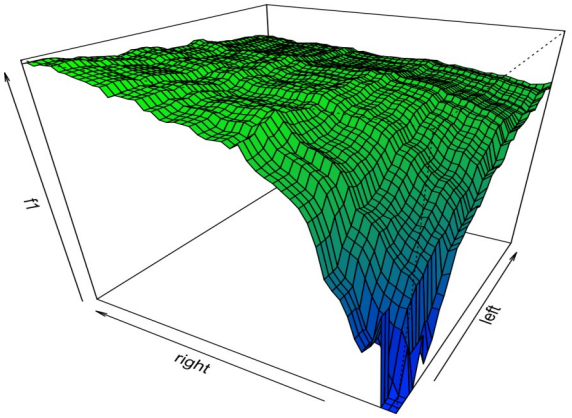
Trước hết, về trích xuất thực thể định danh, chúng tôi so sánh với công cụ LingPipe (http9, 2014) sử dụng trích xuất bằng so khớp với dữ liệu trong từ điển và công cụ SNER (http8, 2014) sử dụng mô hình học máy có giám sát CRF.

Các tham số cho việc định giá kết quả đó là F – score hay F1 (công thức 3), Recall (công thức 2) và Precision (công thức 1).

Kết quả trích xuất của x.ent cho kết quả tốt như công cụ Lingpipe. Lingpipe cũng có cách các cách tiếp cận trên cơ sở mô hình Hidden – markov nhưng nó cho kết quả ít tốt hơn.

Tiếp theo, chúng tôi so sánh kết quả trích xuất của x.ent sử dụng phân tích cấu trúc với cách tiếp cận Coo – currence với các tham số cửa sổ khác nhau, tức là độ rộng của cửa sổ của một đơn vị văn bản sẽ thay đổi về bên trái và bên phải so với thực thể gốc. Hình 7 hiển thị kết quả mà chúng tôi thay đổi sổ của đơn vị văn bản từ thực thể gốc, chúng tôi thử nghiệm cửa sổ bên trái và bên phải thay đổi từ 0 đến 500 từ. Chúng tôi nhận thấy kết quả tốt nhất khi số từ bên trái tiến dần tới 0 (gần tới thực thể gốc) và số từ bên phải tiến dần tới 500.

Bảng 2 cho chúng ta biết kết quả trích xuất quan hệ trong tập dữ liệu này thì phương pháp phân tích cú pháp sẽ hiệu quả hơn F – score khoảng 55%, trong khi phương pháp Coo – ccurrence khoảng 42%. Với dạng tập dữ liệu có cấu trúc, việc sử dụng phương pháp phân tích cấu trúc để tìm ra mối quan hệ sẽ hiệu quả hơn. Ngược lại phương pháp Coo – currence sẽ hiệu quả hơn với tập dữ liệu không có cấu trúc. Trong các bảng dưới, PET là từ viết tắt của thực thể sinh vật gây hại cây trồng, MAL là bệnh của cây trồng, PLA là thực thể tên của cây trồng, REG là thực thể về vị trí địa lý, TOT là kết quả trung bình của các thực thể. PLA – MAL và PLA – PET là mối quan hệ của các thực thể được nêu ở trên.



Hình 7. So sánh kết quả trích xuất quan hệ sử dụng Coo - currence bằng việc thay thế tham số các cửa sổ khác nhau

$$0 \le P \le 1,\ P = \frac{\#\ Tổng\ số\ kết\ quả\ trả\ lời\ đúng}{\#\ Tổng\ số\ kết\ quả\ mà\ công\ cụ\ tìm\ được} \quad (1)$$

$$0 \le R \le 1,\ R = \frac{\#\ Tổng\ số\ kết\ quả\ trả\ lời\ đúng}{\#\ Tổng\ số\ kết\ quả\ đúng\ nhất\ có\ thể} \quad (2)$$

$$0 \le F1 \le 1,\ F1 = \frac{(\beta^2+1)*P*R}{(\beta^2*R+P)} \quad (3)$$

Bảng 1. Định giá kết quả trích xuất thực thể định danh

	X.ENT			SNER			LINGPIPE		
	P	R	F1	P	R	F1	P	R	F1
PET	96.46	95.52	95.98	92.66	71.41	80.52	96.45	95.53	95.99
MAL	96.97	95.53	96.24	95.46	77.38	85.38	96.97	95.52	96.24
PLA	88.80	98.67	93.47	93.99	82.68	87.94	88.80	98.67	93.47
REG	100	100	100	93.20	73.73	81.92	100	100	100
TOT	94.33	96.67	95.48	93.68	76.85	84.41	94.34	96.65	95.48

Bảng 2. Định giá kết quả trích xuất quan hệ giữa các thực thể

	X.ENT			COOCCURRENCE		
	P	R	F1	P	R	F1
PLA - PET	53.4	75.8	52.7	36.4	50.5	42.3
PLA - MAL	58.1	69.5	63.3	41.3	38.7	40.0
TOT	55.3	73.1	62.9	38.1	45.4	41.4

3.2. Phân tích và thống kê dữ liệu sau trích xuất

Công cụ x.ent được phát triển bằng ngôn ngữ Perl cho phần chức năng trích xuất dữ liệu và quan hệ và được đóng gói thành một gói R và có sẵn trên R platform (R Development Core Team). Gói công cụ này cũng cung cấp các hàm trên R hỗ trợ cho người sử dụng phân tích và thăm dò kết quả sau khi trích xuất như: các đồ thị hiển thị sự xuất hiện đồng thời, biểu đồ tần xuất, biểu đồ Venn, biểu đồ chồng xếp lên nhau và sử dụng các giả thuyết thống kê để kiểm tra mối liên hệ giữa các quan hệ.

Trên hình 8 chúng ta nhìn thấy một ví dụ hiển thị song song đồng thời giữa hai thực thể (e1 và e2), e1 là thực thể gốc mà chúng ta tìm kiếm quan hệ với chúng, e2 là một thực thể khác loại ví dụ "mouche du chou" là một trường hợp của thực thể sinh vật gây hại cho cây trồng, "mildiou" là một trường hợp của thực thể bệnh.

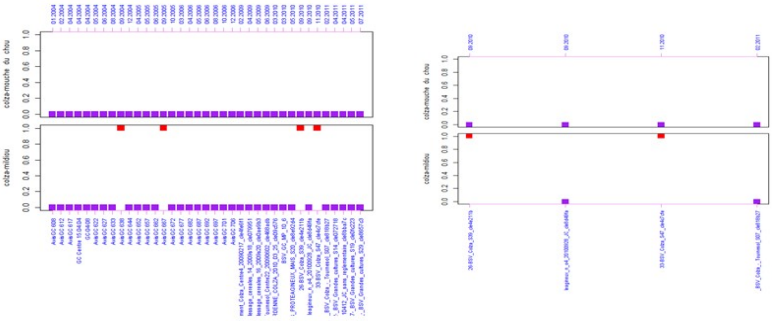
Trong R, bạn có thể đánh như sau:

xplot(e1 = "colza",e2 = c("mouche du chou", "mildiou"))

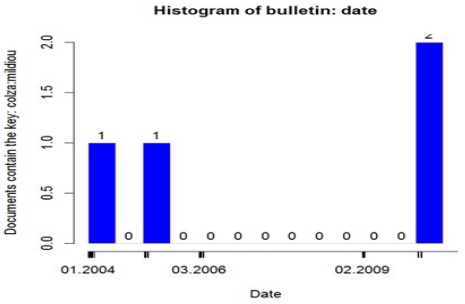
Chúng ta có thể thêm các ràng buộc về thời gian như:

xplot(e1 = "colza",e2 = c("mouche du chou", "mildiou"),t = c("09.2010","02.2011"))

Nhìn vào biểu đồ, người sử dụng có thể biết được tồn tại quan hệ ở trong báo cáo nào và ngược lại. Biểu tượng màu đỏ chỉ tồn tại, màu tím là không tồn tại trong báo cáo.



Hình 8. Biểu đồ so sánh sự xuất hiện đồng thời hay không của các thực thể trong tài liệu



Hình 9. Biểu đồ hiển thị tần xuất theo thời gian của các báo cáo

Biểu đồ tần xuất (histogram) thực hiện thống kê có bao nhiêu báo cáo chứa thực thể, hoặc chứa một quan hệ nào đó theo thời gian. Trong hình 9 là câu lệnh:

xhist("colza: mildiou"), nhìn vào đồ thị, người sử dụng có thể biết được trong giai đoạn nào xuất hiện nhiều bệnh "mildiou" với cây "colza".

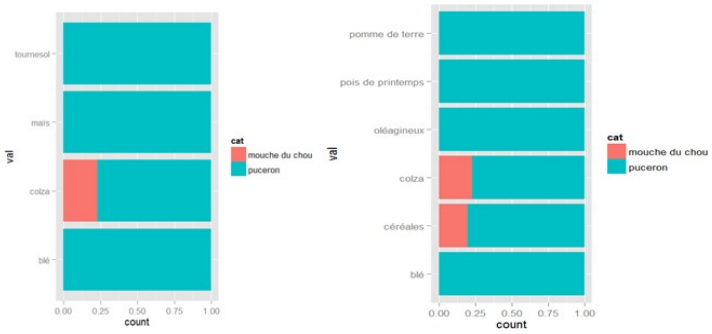
Đồ thị dạng chồng xếp là một trường hợp khác để người sử dụng có thể phân tích được quan hệ giữa các thực thể, ví dụ như quan hệ với cây trồng, dựa vào dữ liệu trích xuất, người sử dụng có thể biết được cây trồng nào thường bị tấn công bởi sinh vật phá hoại nào, còn loại khác thì không. Trong hình 10 là câu lệnh:

xprop(c("blé","maïs","tournesol","colza"),c(" mouche du chou", "puceron"))

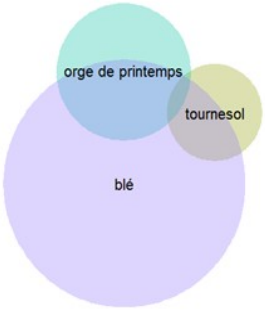
Nhìn vào đồ thị kết quả, chúng ta biết rằng cây "colza" là cây củ cải đường có thể bị tấn công bởi "mouche du chou" là ruồi dấm và "puceron" là rệp. Trong khi các loại cây khác như "tournesol" là cây hướng dương, "maïs" là cây ngô, "blé" là cây lúa mì chỉ bị tấn công bởi "puceron".

Một bài toán khác đặt ra sau khi trích xuất đó là phân tích sự xuất hiện đồng thời của các thực thể hoặc các quan hệ trong các báo cáo. Trong hình 11 là ví dụ so sánh sự xuất hiện đồng thời của các cây "blé", "orge de printepmps" và cây "tournesol", chúng ta có thể thực hiện trong R như sau:

xvenn(c("blé","orge de printemps","tournesol")



Hình 10. Biểu đồ dạng chồng xếp



Hình 11. Biểu đồ dạng Venn

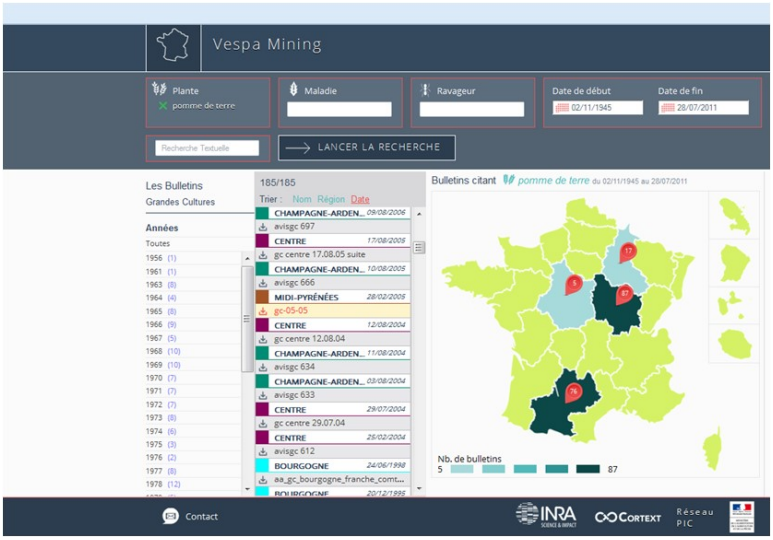
Bảng 3. So sánh các cặp quan hệ

	Relation	KOLMOGOROV	WILCOXON	STUDENT	GrowthCurves
700	blé: méligèthe/blé: thrips	1.00	0.13	0.13	0.02
543	blé: cicadelle/blé: pyrale	1.00	0.00	0.00	0.02
613	blé: criocère/blé: thrips	1.00	0.00	0.00	0.02
689	blé: méligèthe/blé: puceron des épis de céréales	0.91	0.00	0.00	0.02

Để đánh giá khả năng xuất hiện đồng thời của các quan hệ của các thực thể khác nhau, chúng tôi cũng đề xuất sử dụng các phân bố xác suất để đánh giá độ tương đồng của các quan hệ hay trong bài toán đánh giá về cây trồng với dịch bệnh, dùng các phân bố xác suất để đánh giá xem các bệnh nào có thể xảy ra ở cùng thời điểm. Chúng tôi đề xuất sử dụng các phân bố xác suất: Kolmogorov, Wilcoxon, Student, GrowthCurves để tính độ tương đồng của các quan hệ với nhau. Các giá trị p - value này sẽ giúp người sử dụng đánh giá các cặp quan hệ này có xảy ra tại cùng một thời điểm hay không.

3.3. Tích hợp kết quả trích xuất

Công cụ x.ent thực hiện trích xuất thông tin, kết quả là một định dạng theo kiểu CSV, vì vậy thường sẽ gây khó khăn cho người sử dụng thông thường. Chúng tôi đã xây dựng một ứng dụng Web có tên PESTOBSERVER, tại địa chỉ http: //www.pestobserver.eu, tích hợp kết quả trích xuất dữ liệu và có liên kết với tài liệu gốc của báo cáo cây trồng đó. Trên giao diện này cho phép tìm cây trồng, quan hệ cây trồng với bệnh và sinh vật gây hại với cây trồng trong một khoảng thời gian nào đó. Sau đó nó sẽ tìm kiếm đưa ra tất cả các bài báo cáo liên quan đến chủ để mà người sử dụng cung cấp.



Hình 12. Giao diện người dùng cuối tích hợp kết quả x.ent

4. KẾT LUẬN

Chúng tôi đã xây dựng thành công một công cụ có tên là x.ent và đã áp dụng công cụ này cho trích xuất thông tin vào trong các dữ liệu là các báo cáo về phòng chống dịch bệnh cho cây trồng của Pháp. Công cụ này trích xuất quan hệ crops/diseases và crops/pests có độ chính xác F - score 62%.

Ngoài ra, chúng tôi còn xây dựng được một platform giao diện thân thiện với người sử dụng mà tích hợp kết quả trích xuất kết hợp cùng với vị trí địa lý nơi xảy ra dịch bệnh và liên kết với báo cáo gốc.

Chúng tôi cũng quan tâm tới việc trợ giúp người sử dụng khám phá các mối quan hệ tiềm năng giữa các thực thể. Hai hướng mà chúng tôi đã và đang tiếp tục thực hiện: Thứ nhất, cung cấp giao diện trực quan dưới dạng đồ hoạ (các đồ thị, bảng biểu) để cho người sử dụng dễ dàng so sánh được kết quả và đưa ra các đánh giá như đồ thị so sánh đồng thời, đồ thị tần xuất, biểu đồ Venn, biểu đồ chồng xếp và áp dụng các phân bố thống kê để đánh giá kết quả. Thứ hai là tích hợp kết quả trích xuất vào trong một platform thân thiện với người dùng kết hợp với các thông tin thực tế. Ở đó, người sử dụng có thể duyệt qua tập liệu thông qua quan hệ các thông tin phụ trợ (vị trí địa lý, mức độ thiệt hại) sử dụng bản đồ địa lý và có thể phản hồi lại với các tài liệu gốc.

Ngôn ngữ tiếng việt khá là phức tạp so với ngôn ngữ tiếng anh như cấu trúc từ, ngữ pháp... Chúng tôi đang tiếp tục nghiên cứu nhằm cải tiến công cụ này để có thể xử lý với ngôn ngữ tiếng việt.

LỜI CẢM ƠN

Tôi xin gửi lời cám ơn đặc biệt tới người đã hướng dẫn tôi Dr. Nicolas Turenne (Paris - Est University), người đã cùng sát cánh với tôi trong thời gian thực hiện dự án; Prof. Kurt Hornik (Vienna University), người đưa ra những phản biện về khía cạnh kỹ thuật; Roselyne Corbière (INRA - Rennes center) và Vincent Cellier (INRA - Dijon center) về những góp ý cho ý tưởng giao diện, chức năng người dùng cuối, và tới Jean - Noel Aubertot (INRA - Toulouse center) về ý tưởng cho việc xây dựng bộ dữ liệu về phòng chống dịch bệnh cho cây trồng. Cảm ơn những đồng nghiệp làm việc tại labo INRA - LIGM đã trợ giúp về công nghệ, kỹ thuật trong thời gian tôi thực hiện dự án của tôi ở đây.

TÀI LIỆU THAM KHẢO

Abacha A.B., Zweigenbaum P. et Max A. (2012). Extraction d'information automatique en domaine médical par projection inter - langue: vers un passage à l'échelle (Automatic Information Extraction in the Medical Domain by Cross - Lingual Projection) [in French]. *La conférence JEP - TALN - RECITAL 2012, volume 2: TALN*, p. 15 - 28.

Carpenter B. (2007). LingPipe for 99.99% Recall of Gene Mentions. *Proceedings of the 2nd BioCreative workshop*, Valencia, Spain.

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A. et Billot S. (2011). Intégrer des connaissances linguistiques dans un CRF: application à l'apprentissage d'un segmenteur - étiqueteur du français. *TALN*. Montpellier, p. 1 - 12.

Faure C., Delprat S., Mille A. et Boulicaut J. - F. (2006). Utilisation des réseaux bayésiens dans le cadre de l'extraction de règles d'association. *Actes 6ème Journées Francophones Extraction et Gestion de Connaissances EGC'06*, p. 569 - 580.

Finkel J.R., Grenager T. and Manning C. (2005). Incorporating Non - local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (Stroudsburg, PA, USA, 2005), p. 363 - 370.

http1 Stackoverflow (2014). http: //stackoverflow.com.

http2 Manuel d'Utilisateur « Writing R Extentions » (2014). http://cran.r-project.org/doc/manuals/R-exts.html.

http3 O beautiful code, « How R Searches and Finds Stuff » (2014). http://obeautifulcode.com/R/How-R-Searches-And-Finds-Stuff/.

http4 Précision et rappel (2007). http://benhur.teluq.ca/SPIP/inf6104/article.php3?id_article = 98&id_rubrique = 10&sem = Semaine%208.

http5 Wilkipedia (2014). http://fr.wikipedia.org.

http6 Les Résaux Bayésiens (2014). http://w3.jouy.inra.fr/unites/miaj/public/matrisq/Contacts/abari.07_ 03_12. expo2.pdf

http7 Traitement Automatique du Langage Naturel (2014). http://lipn.univ-paris13.fr/~audibert/pages/enseignement /TAL_ITCN.pdf.

http8 Stanford Named Entity Recognizer (2014).http://nlp.stanford.edu/software/CRF-NER.shtml.

http9 LingPipe (2014)http://alias-i.com/lingpipe/.

http10 Information Extraction And Named Entity Recognition (2014). https://web.stanford.edu/class/cs124/lec/ Information_Extraction_and_Named_Entity_Recognition.pdf.

http11 Les Réseaux Bayésienes. http://www.bayesia.com/fr/technologie/reseaux-bayesiens.php.

Lafferty J., McCallum A. et Pereira F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Dep. Pap. CIS*.

Moncla L. (2013). Automatic Annotation of Motion Expressions and Place Named Entities. *2nd Unitex/GramLab*.

Paumier S. et Martineau C. (2006). Manuel d'Utilisateur Unitex 3.1 Beta. Université Paris - Est Marne - la - Vallée. version 1.2.

Sutton C. et McCallum A. (2010). An Introduction to Conditional Random Fields for Relational Learning. *1011.4088 [stat]*, p. 5 - 32.

R Development Core Team, R (2015). A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3 - 900051 - 07 - 0 (2015). URL http: //www.R - project.org/

Tannier X. (2012). Traitement Automatique des Langue. Université Paris - Sud.

Turenne N. (2013). *Knowledge Needs and Information Extraction*. Wiley - ISTE.

Zettlemoyer L. (2012). Relation Extraction. University of Washington.