

آليات استخلاص المصطلح

Term extractor tools

✉ ماجدة الرامي

Rami.majda@gmail.com

جامعة ابن طفيل / المغرب

تاريخ النشر: 2022/01/23

تاريخ القبول: 2022/01/06

تاريخ الاستلام: 2021/10/31



ABSTRACT:

ملخص البحث

Due to the fast development of technology and the evolution of knowledge, the manual traditional term extracting techniques became almost useless. For that, it was necessary to consider using new digital extracting tools that are more sophisticated and easily updated to automatically analyze specialized texts and extract the words that are most likely terms. In this paper, we evaluate several term extraction tools presenting their adopted strategies; as a result, we are going to draw conclusions considering the most common problems related to these term exactors.

Keyword: terms extraction, simple term, compound term, term candidates.

إن التطور السريع للمعارف وتزايد كمية المعلومات على الإنترنت في بعد كوني جعل من عملية الاستخلاص اليدوي للمصطلح غير ممكنة. ولتفحص الكم الهائل من المتون الرقمية أصبحت عملية استخلاص المصطلحات تتم بطريقة آلية بواسطة أدوات معلوماتية مخصصة لذلك والتي تمكن من تسهيل جمع المعطيات المصطلحية وتوظيفها. نقدم من خلال هذه الورقة عرض لمجموعة من البرمجيات المخصصة للاستخلاص المصطلح في بعد لُغوي كوني وطريقة اشتغالها والمشاكل التي تطرحها.

الكلمات المفتاحية: استخلاص المصطلحات، مصطلح مركب، مصطلح بسيط، المصطلح المرشح.

مجلة لغة – كلام / مختبر اللغة والتواصل / جامعة غليزان (الجزائر)

1. مقدمة:

عرف الإنتاج المصطلحي في بعد لغوي كوني تطوراً كبيراً بفعل التقدم الحاصل والمتسارع في ميادين مختلفة. كما ساعد النشر باللغات الحية في المجالات العلمية المحكمة والمفهرسة على ظهور عدد كبير من المفاهيم الجديدة في مختلف الحقول المعرفية المستجدة. وتُشكّل المنشورات المذكورة متونا رقمية متاحة ومتعلقة بمجالات معرفية محددة متضمنة للعديد من المصطلحات التي يستوجب تحديثها توازياً مع تطور هذه المتون. تتوفر اللغات الحية مثل الفرنسية والإنجليزية على العديد من الأدوات المعلوماتية لتفحص المتون الرقمية بغية استخلاص المصطلحات آلياً، بينما نجد أن اللغة العربية للأسف لا تحظى بهذه التقنيات لأن البرمجيات التي تدعم عدداً من اللغات الأوروبية، لا تدعم اللغات السامية بما فيها اللغة العربية.

تُعتبر عملية استخلاص المصطلحات من التقنيات الآلية التي تدخل في إطار المعالجة الآلية للغات الطبيعية، وتُمكن من البحث عن المصطلحات انطلاقاً من متون لغوية متخصصة بواسطة برمجيات تسمى مستخلصات المصطلحات. تعمل هذه الأخيرة على إحصاء واستخراج جميع المصطلحات المرشحة الممكنة، معتمدة في ذلك على مجموعة من المعايير التي تختلف فيما بينها في طريقة تحديدها للمصطلحات. تساعد عملية استخلاص المصطلحات على تسهيل عمل المصطلحي وعلى تسهيل عمل المترجم المتخصص في مجال معين، وذلك بتيسير عملية جمع المعطيات المصطلحية بهدف توظيفها في تطبيقات مختلفة نذكر منها: الترجمة بمساعدة الحاسوب والفهرسة الآلية،¹ ويمكن أن تستخدم أيضاً كأداة مساعدة لاختيار مفردات المعاجم المتخصصة لبناء الأنطولوجيا.²

سنقدم من خلال هذه الورقة معاينة لبعض الأدوات المخصصة لاستخلاص المصطلحات بالنسبة للغتين الفرنسية والإنجليزية والاستراتيجيات التي تتبناها ومجموع من الخلاصات التي توصلنا إليها.

قسمنا هذه الدراسة إلى ثلاثة أقسام. القسم الأول، سنتطرق من خلاله إلى تعريف المصطلح آلياً والقسم الثاني يشمل الاستراتيجيات المتبناة في عملية استخلاص المصطلحات، أما القسم الثالث فخصصناه لعرض بعض برمجيات استخلاص المصطلحات المتوافرة ثم القسم الخامس خصصناه إلى مجموع الخلاصات التي توصلنا إليها من خلال تجربتنا لهذه البرمجيات.

2. تعريف المصطلح المستخلص آلياً

يعتبر المصطلح بمثابة لفظ ينتمي معناه إلى مجال تخصص معين. وينقسم إلى قسمين: مصطلح بسيط ومصطلح مركب.

— مصطلح بسيط: يتكون من كلمة خطية واحدة، على سبيل المثال: تضخم، استثمار.

– مصطلح مركب: يتألف من كلمتين أو عدة كلمات، على سبيل المثال: قاعدة معطيات، عقد عمل.

يمكن أن يأتي المصطلح على شكل متلازمة لفظية لما يكون مؤلفا من مصطلح ووحدة معجمية أخرى ويكون بينهما تلازم دلالي، مثل: رأس مال. ويمكن أيضا للمصطلحات أن تنتمي إلى مقولات نحوية أخرى كالأفعال والصفات، إلا أن وورودها في المتون قليلا ولا تتم فهرستها في المعاجم المتخصصة وبنوك المصطلحات الآلية، إلا إذا كانت: تستعمل حصريا في مجال متخصص دون سواه، على سبيل المثال: إعداد برنامج في مجال المعلومات وكذلك تصفح الأنترنت في مجال الأنترنت. أو إذ كانت تحمل معنى إنسانيا، وتكون مرفقة بفواعل دلالية ذات معنى متخصص، على سبيل المثال: برمج يمكن ترتبط بفاعل أول: برمجية وفاعل ثاني: خبير في المعلومات وكلاهما مرتبط بمجال المعلومات.³

3. استراتيجيات تحديد المصطلحات

تعتمد مستخلصات المصطلحات على ثلاث استراتيجيات أساسية لاستخلاص المصطلح: استراتيجية لسانية واستراتيجية إحصائية واستراتيجية هجينة تجمع بين الاستراتيجيتين اللسانية والإحصائية.

1.3. الاستراتيجية اللسانية

تنبني هذه الاستراتيجية على افتراض مفاده أن المركبات الاسمية التي يمكن أن تكون مصطلحات تتألف من عدد محدود من المقولات النحوية التي تكون محددة مسبقا. فإذا طلبنا من مستخلص المصطلحات التعرف على المقولات النحوية التالية: اسم متبوع بصفة واسم متبوع باسم واسم متبوع بحرف جر متبوع باسم الخ، سيقوم حينئذ مستخلص المصطلحات باستخراج جميع المركبات الاسمية المتوافقة مع هذه المقولات النحوية والتي تتخذها البرمجية بمثابة قوالب نموذجية يُعتمد عليها لاستخراج مصطلحات من نص متخصص. وتتطلب هذه الاستراتيجية معرفة جيدة باللغة التي يتم الاشتغال عليها، لأن تقسيم المقولات النحوية يختلف من لغة إلى أخرى.

كما تستعين بعض برمجيات استخلاص المصطلحات التي تتبنى هذه المقاربة بواسم آلي صرف- تكميلي لتحليل النصوص وتحديد المقولات النحوية التي تتألف منها، مثل: النسخة التجريبية لبرمجية جينيا ديموتاجر Genia Tagger Demo الخاص باللغة الإنجليزية.

وقد عمل بوريكو (1992) Bourigault على تطوير برمجية تسمى ليكستر LEXTER تعتمد على استراتيجية لسانية لمعالجة النصوص باللغة الفرنسية وتمكن من تحديد سلاسل الألفاظ الأكثر طولاً وعزلها بواسطة حدود، عن طريق إجراء تقطيعات مختلفة في النص، مرتكزة في ذلك على المقولات

النحوية التي لا تساهم في تكوين المصطلحات، مثلاً: الضمائر والأفعال وحروف الجر وعلامات الترقيم، وهي بالتالي تقوم بعزل الوحدات المصطلحية التي يمكنها أن تكون بمثابة مصطلحات:

L'/opérateur /passe en/ recirculation directe sur puisard /grâce la/vanne manuelle d'isolement d'enceinte /qui est équipée du/ clapet de sécurité rapide.⁴

يجري العامل مروراً مباشراً إلى الحوض بواسطة صمام يدوي حابسي. علماً أن هذا الأخير مجهز بصمام أمان سريع.

بعد رسم الحدود تم الاحتفاظ بالمقولات الاسمية التالية:

Opérateur

recirculation directe sur puisard

vanne manuelle d'isolement d'enceinte

clapet de sécurité rapide

العامل

مروراً مباشراً إلى الحوض

صمام يدوي حابسي

صمام أمان سريع.

2.3. الاستراتيجية الإحصائية

تعتمد هذه الاستراتيجية على وتيرة تردد الوحدات المعجمية وعدد مرات تكرارها. فالوحدات التي عادة ما تتكرر أكثر في النصوص المتخصصة عادة ما تُشكّل مصطلحات. وكذلك الألفاظ التي تظهر مع بعضها مرات عديدة داخل نص متخصص عادة ما تكون مصطلحات، ويتم تصنيفها ضمن المصطلحات المركبة.

توجد العديد من الطرق الإحصائية لاستخراج المصطلحات، وتستند معظمها على ما يسمى بالمعلومة المتبادلة (information mutuelle).

استخدم هذه الاستراتيجية شيرش وهانكس (1990) Church and Hanks. وتُمكن من استخراج المصطلحات المركبة والمتلازمات اللفظية وترتكز على افتراض مفاده أن كل لفظين متجاورين في وثيقة نصية، يمكن أن نُكوّنَ منهما سلسلة مؤلفة من ألفاظ (س وص) عبر تطبيق بعض الحسابات الإحصائية المبنية على مؤشر الاحتمال. ويمكن حساب المعلومة المتبادلة من خلال احتساب احتمال

ورود الزوج (س،ص) معا من جهة، على احتمال ورود كل من س وص على حدة من جهة أخرى، ويُقدَّر الاحتمال بالأخذ بعين الاعتبار مجموع ورود الزوج (س،ص) وورود س وص على مجموع كلمات نص متخصص، ثم يتم تطبيق القاعدة التالية:

$$IM(ص،س) = \log_2 \frac{P(ص،س)}{P(ص)P(س)}$$

IM يدل على المعلومة

س يدل على الكلمة الأولى للزوج (س،ص)

ص يدل على الكلمة الثانية للزوج (س،ص)

P يدل على الاحتمال

وحسب ما ورد في أعمال شيرش وهانكس،⁵ تكون المعلومة المتبادلة قريبة من 0، إذا كان احتمال تجاوز اللفظين س وص ضعيف جدا. وتكون المعلومة المتبادلة أكبر من 0، إذا كان تجاوز س وص مثير للاهتمام من حيث تأليفهما لمصطلح ما. وكلما ارتفعت المعلومة المتبادلة كلما ارتفعت مصداقية تأليف س وص لمصطلح مركب ما.

3.3. الاستراتيجية الهجينة

تعتبر الأكثر شيوعا، حيث تجمع بين التقنيات اللسانية والإحصائية في آن واحد. تم اقتراح هذه الاستراتيجية من طرف داي (1994) Daille التي وصفها بالمقاربة الهجينة، وعملت على تَقْعِيدِهَا في برمجة تسمى أكابيت Acabit، وتعمل هذه الأخيرة على:

- استخراج المصطلحات المركبة من نص موسوم مسبقا،
- تحديد القوالب النموذجية التي تتألف منها سلاسل الكلمات التي تم استخراجها أي استراتيجية لسانية،
- تطبيق مختلف الحسابات الإحصائية عبر مراحل، وذلك لتصفية اللوائح المرشحة التي تم استخراجها مسبقا، وتحديد وضعها المصطلحي وانتقاء المصطلحات الممكنة منها.

4. برمجيات استخلاص المصطلح

هي عبارة عن برمجيات وظيفتها العثور في نص أو مجموعة من النصوص المتخصصة على الألفاظ التي من الممكن أن تشكل مصطلحات. توجد العديد من البرمجيات المخصصة لاستخلاص

المصطلح أليا من متن متخصص وتختلف فيما بينها من حيث الاستراتيجيات المعتمدة في استخلاص المصطلح المرشح.

1.4. برمجة فايف فالتر Fivefibre

يتعلق الأمر ببرمجة مفتوحة المصدر، تدعم اللغة الإنجليزية وتمكّن من استخلاص المصطلحات الأكثر ملاءمة وذلك بالاعتماد على المادة اللغوية المتوفرة على الأنترنت. تأخذ المصطلحات التي يتم استخلاصها أشكالاً مختلفة من الصيغ النصية، مثل: (JSON.XML .HTML .TEXTE. PHP)، بالإضافة إلى أن هذه الخدمة مدمجة مجاناً ضمن موقع ياهو Yahoo انطلاقاً من دليل هذه البرمجة. لم يتم ذكر المقاربة المعتمدة، ولقد لاحظنا من خلال تجربتنا للبرمجة أنها تمكّن من تحديد المصطلح مقترناً بعدد مكوناته (ألفاظه) وعدد تردده أي مقارنة إحصائية⁶ كما هو مبين في الجدول (1).

قمنا بتجريب برمجة فايف فالتر على مقطع من نص ينتمي إلى الطب باللغة الإنجليزية:

نص (1):

Aspirin is often used as an analgesic to relieve pains. Aspirin is also used long-term, at low doses, to help prevent heart attacks, strokes, and lower the risk of forming a blood clot in the coronary arteries. People at high risk of Heart attack should take a daily low dose of aspirin to reduce the risk of another heart attack.

يستخدم الأسبرين في كثير من الأحيان بمثابة مُسكّن لتخفيف الآلام. ويستخدم أيضاً على المدى الطويل، في جرعات منخفضة، للمساعدة على منع النوبات القلبية والسكتات الدماغية، والتقليل من خطر تجلط الدم في الشرايين التاجية. لذلك يجب على الناس الأكثر عرضة لخطر الإصابة بالنوبات القلبية أن يأخذوا جرعة يومية منخفضة من الأسبرين، للحد من خطر الإصابة بنوبة قلبية ثانية:

الجدول (1): ناتج تجريب برمجة فايف فالتر وفق ثلاث حقول: المصطلح المرشح وتردده وعدد ألفاظه.

Term	Occurrence	Word count
heart attack	2	2
blood clot	1	2
aspirin	3	1
risk	3	1
heart attacks	1	2

pains	1	1
dose	1	1
People	1	1
arteries	1	1
doses	1	1
strokes	1	1

Powered by FiveFilters.org

نلاحظ من خلال الجدول رقم (1)، أن المرشحات التي تم استخراجها على أساس مصطلحات تنتمي إلى المقولات النحوية الاسمية. كما نلاحظ أيضا أن هذه البرمجية تركز على الأجزاء الأساسية للمصطلح، وتستثني المقولات النحوية الأخرى التي تدخل في تركيب المصطلح كالصفات، على سبيل المثال: بالنسبة للمصطلح المرشح coronary arteries (الشرايين التاجية) تم الاحتفاظ بلفظ arteries (الشرايين) واستبعاد لفظ coronary (التاجية) بما أنه صفة. ونفس الملاحظة تنطبق على لفظ analgesic (مُسَكِّن) الذي تم استبعاده من لائحة المصطلحات المرشحة لأنه أيضا يُعتبر صفة. لا تعالج هذه البرمجية الخصائص الصرفية للمصطلحات المرشحة التي تم إدراجها ضمن اللائحة، بحيث تم تخصيص مدخلين في الجدول للفظ heart attack (نوبة قلبية) في صيغة المفرد وفي صيغة الجمع. وقد أدرجت هذه البرمجية ضمن لائحة المرشحات بعض الألفاظ التي لا يمكن اعتبارها مصطلحات مثل: people (الناس) الذي يعد لفظا منتسبا إلى اللغة العامة.

2.4. برمجية ترنسلات لابز Translated labs لاستخلاص المصطلحات.

تدعم برمجية ترنسلاتد لابز⁷ ثلاث لغات: الفرنسية، والإنجليزية، والإيطالية. تم تصميم هذه التكنولوجيا لمساعدة المترجمين من أجل التغلب على الصعوبات التي يواجهونها أثناء الترجمة، وكذلك لتبسيط عملية إنشاء المعاجم. تعمل هذه البرمجية على عرض اللفظ في سياقه، لتسهيل عمليات التحقق من مقبولية المصطلحات.

وتعتمد ترنس لابز في طريقة تحديدها للمصطلحات على مقارنة إحصائية تنبني على إحصائيات قانون بواسون (les statistiques de loi poisson) وعلى طريقة الإمكانية القصوى (Maximum de vraisemblances)، وهما طريقتان حسابيتان تُستعملان في مجال الإحصاء وتعتمدان على مؤشر الاحتمال لاستخراج المصطلحات المركبة. كما تتوفر هذه البرمجية على محلل صرفي للاحتمال يأخذ بعين الاعتبار جميع الاحتمالات لكي تكون سلسلة الألفاظ المستخرجة مصطلحا وتتوفر أيضا منهج إحصائي ينبني على أساس وتيرة ورود المصطلحات.

قمنا بتجريب برمجية ترنسلات لابز على نص (1) باللغة الإنجليزية فحصلنا على النتيجة التالية :

جدول (2):ناتج تجريب برمجية ترنسلات لابز وفق حقلين المصطلح المستخلص ونسبة وروده.

	Extracted term	Score
1	aspirin	63%
2	coronary arteries	62%
3	relieve pains	56%
4	daily low dose	54%
5	blood clot	49%

نلاحظ من خلال الجدول (2) بالنسبة للغة الإنجليزية، تُركز هذه البرمجية على المصطلحات البسيطة التي تتكرر بكثرة وتُركز أيضا على المصطلحات المركبة والمتلازمات اللفظية التي تظهر مع بعضها داخل النص. وتقوم بترتيبها حسب أهمية نسبة ورودها في النص. كما تستبعد هذه البرمجية المصطلحات المرشحة البسيطة التي تظهر مرة واحدة في النص، على سبيل المثال analgesic (مُسَكِّن) ثم pains (الآلام) و strokes (السكتات الدماغية). ولا تُدرج بعض المصطلحات رغم تكرارها، مثل: heart attacks.

بالنسبة للغة الفرنسية قمنا بتجريب نص ينتمي إلى مجال الطب والبيولوجيا:

نص(2):

Le cœur est un organe musculaire creux, situé entre les poumons au milieu du thorax. C'est le moteur du système cardio-vasculaire, dont le rôle est de pomper le sang qu'il fait circuler dans tous les tissus de l'organisme. le cœur a besoin d'oxygène et de nutriments pour fonctionner correctement. Le sang lui parvient par les [artères coronaires](#). L'obstruction d'une artère coronaire peut entraîner une crise cardiaque.

القلب عضو عضلي أجوف يقع بين الرئتين في منتصف الصدر. وهو محرك الجهاز القلبي الوعائي. ويكمن دوره في ضخ الدم الذي يدور في جميع أنسجة الجسم. ويحتاج القلب إلى الأوكسجين ومواد القيت لكي يعمل بشكل سليم. ويصل الدم إليه عبر الشرايين التاجية. وقد يؤدي انسداد شريان تاجي إلى نوبة قلبية.

الجدول (3): ناتج تجريب برمجية ترنسلات لابز وفق حقلين المصطلح المستخلص ونسبة وروده.

	Terme	Score
1	artères coronaires	71%

2	organe musculaire creux	68%
3	cœur	62%
4	entraîner une crise cardiaque	50%
5	artère coronaire	44%
6	moteur du système cardio-vasculaire	40%
7	cœur est un organe	38%
8	cœur a besoin	38%
9	milieu du thorax	37%
10	nutriments pour fonctionner	36%
11	obstruction d' une artère	34%
12	pomper le sang	32%
13	poumons au milieu	31%
14	situé entre les poumons	31%

نلاحظ من خلال الجدول (3) بالنسبة للغة الفرنسية، تركّز هذه البرمجية على المصطلحات المرشحة البسيطة والمركبة والتي تكون مرتبة بشكل تنازلي تبعاً لنسب وورودها داخل النص. بالمقابل نلاحظ أن هذه البرمجية تُدرج ضمن لائحة المرشحات مقولات نحوية لا تنتمي إلى فئة المصطلحات، على سبيل المثال: poumons au milieu (الرئة في وسط) وكذلك situé entre les poumons (يوجد بين الرئتين) ثم coeur est un organe (القلب هو عضو) و coeur a besoin (القلب يحتاج). كما أن هذه البرمجية لا تعالج الخصائص الصرفية للمصطلحات، على سبيل المثال: coronlh aire artère (الشريان التاجي) و artères coronaires (الشريانات التاجية) التي تم تخصيص لها مدخلين في الجدول في صيغة المفرد وفي صيغة الجمع. وتُعتبر المرشحات التي تم استخلاصها ليست كلها مصطلحات بل هي عبارة عن أجزاء من مصطلحات، ماعدا المرشحات التي ظهرت في بداية الجدول والتي حصلت على أعلى نسب ورود.

3.4. برمجة ترمين Termine لاستخلاص المصطلحات

ترمين⁸ هي عبارة عن برمجة مخصصة لاستخلاص المصطلح، تدعم اللغة الإنجليزية وتعتمد في تحديدها للمصطلحات المرشحة على استراتيجية هجينة تجمع بين تحليل لساني وتحليل إحصائي. بالنسبة للتحليل اللساني تتوفر برمجة ترمين على واسم لتحديد المقولات النحوية حسب نوعية النصوص التي يتم وسمها. ويُمكنُ التحليل اللساني باستخلاص جميع المصطلحات المرشحة الموجودة في النص والتي تكون عبارة عن (سلاسل ألفاظ، صفات/أسماء). إضافة إلى ذلك يُوقَّر هذا التحليل

لائحة للاستبعاد⁹ (stop-list) وذلك لاستبعاد الألفاظ التي لا يمكن أن تُشكّل مصطلحات في اللغة الإنجليزية، مثل: الضمائر وحروف الجر وحروف العطف... الخ. أما بالنسبة للتحليل الإحصائي فيعتمد على وتيرة تردد المصطلح المرشح في النص ويعتمد أيضا على طول المصطلح المرشح.

تعتمد برمجية ترمين¹⁰ على واسم لتحديد المقولات النحوية وذلك حسب نوعية النصوص التي يتم وسمها. فبالنسبة للنصوص العامة يتم استخدام واسم تري تجر (Tree Tagger) نسخة 3.1. أما بالنسبة للنصوص المتخصصة في علوم الطب الحيوي، فيعتمد على واسم جينيا (GENIA Tagger) نسخة 2.1. وتستعين بمعجم أكرومين للتعرف على المختصرات بالنسبة للمصطلحات التي تُكتب بشكل مختزل مثل السيدا (SIDA).

قمنا بالاشتغال على النص (4) باللغة الإنجليزية:

الجدول (4): ناتج تجريب برمجية ترمين

Rank	Term	Score
1	Heart Attack	3
2	daily Low dose	1,584962
3	Low dose	1
3	Blood clot	1
3	Coronary artery	1

من خلال الجدول رقم (4) نلاحظ أن هذه البرمجية تُرتّب المصطلحات المرشحة التي يتم استخلاصها وفقا لأهميتها. كما أنها تتعرف على المصطلحات المرشحة من خلال تطبيق قواعد الاستخراج على جذع الكلمة وليس على الشكل الذي جاءت عليه في النص، على سبيل المثال:

جدول 5: تخليص المرشحات من الزوائد الصرفية

المصطلح كما جاء في النص	المصطلح الذي تم استخلاصه من طرف البرمجية
heart attacks	Heart Attack
coronary arteries	Coronary artery

تركز هذه البرمجية على طول المصطلح المرشح. وذلك، بإدراج جميع الإمكانات لكي تكون سلسلة الألفاظ المستخرجة مصطلحا مركبا أو تكون جزءاً من مصطلحات أخرى، مثل: Low dose (جرعة منخفضة) و Daily low dose (جرعة يومية منخفضة). كما أنها تأخذ بعين الاعتبار

طبيعة البنية التركيبية للمصطلحات المرشحة في اللغة الإنجليزية، مما يجعلها جيدة لاستخراج المركبات الاسمية والمتلازمات اللفظية، مثل:

جدول 6: المصطلح المرشح وطبيعة بنيته التركيبية في اللغة الإنجليزية

المصطلح	طبيعة بنيته التركيبية
aspirine	اسم
Heart attack	اسم-اسم
Coronary arteries	صفة-اسم

4.4. برمجة Alchemy API لاستخلاص المصطلحات

تعمل برمجة Alchemy API¹¹ على استخراج الكلمات المفتاحية والكيونات (entities) من نص وترتيبها وفقاً لأهميتها مع تحليل لدلالة الأحاسيس التي تحملها الجمل، وذلك للتعرف على الألفاظ التي تحمل معينة: دلالة إيجابية أو سلبية أو محايدة، لشخص أو مكان أو تُحِيل على شيء ما. مثلاً نقول: إن هذه السيارة جميلة (إحساس إيجابي) أو هذه السيارة ليست جميلة (إحساس سلبي).

تعتمد هذه البرمجة على لوغاريتمات إحصائية متطورة لاستخراج الكلمات المفتاحية وترتيبها، بالإضافة إلى توفرها على تقنيات متطورة لمعالجة اللغات الطبيعية في تحليل النصوص. وتدعم هذه البرمجة اللغة الإنجليزية والألمانية والفرنسية والإيطالية والبرتغالية والروسية والإسبانية ثم السويدية. وتبقى اللغة الإنجليزية هي اللغة التي تقدم خدمة متطورة في تحليل النصوص.

تدمج هذه البرمجة ضمنها خدمات أخرى يمكن للمصطلحي أن يستعين بها، كاستخراج المفاهيم من النصوص التي تتم معالجتها وتحديد المجالات التي تنتهي إليها المرشحات، مثلاً: أسبيرين (Aspirin) ينتمي إلى مجال العقاقير الأدوية (drugs). تقوم هذه البرمجة بتحليل النصوص في صيغة نصية، أو عناوين (URL)، وتأخذ المصطلحات التي يتم استخلاصها أشكالاً مختلفة من الصيغ النصية مثل: RDF و REL و XML و JSON.

قمنا بتجريب برمجة Alchemy API على النص (1) باللغة الإنجليزية، فحصلنا على اللائحة

التالية:

الجدول (7): ناتج تجريب برمجة Alchemy API لاستخلاص المصطلحات

KEYWORD	RELEVANCE	SENTIMENT
Daily low doses	0.973886	Negative
Heart attack	0.914901	Negative
Coronary arteries	0.785817	Negative
Blood clot	0.751577	negative

Low doses	0.71227	negative
Aspirin	0.669001	negative
Heart attacks	0.660123	negative
high risk	0.61927	negative
analgesic	0.483302	negative
pains	0.463671	negative
strokes	0.425351	negative

نلاحظ من خلال الجدول أن هذه البرمجية تستخرج المصطلحات وترتبها حسب أهميتها. ولا تعالج الخصائص الصرفية للمصطلحات بحيث تم تخصيص مدخلين في الجدول للمصطلح المرشح *heart attack* في صيغة المفرد وصيغة الجمع. كما أنها تعتبر جيدة لاستخلاص المصطلحات البسيطة والمصطلحات المركبة بحيث أنها تُدرج جميع الإمكانات لكي تكون المرشحات المستخرجة مصطلحا.

5. خلاصات

مهما كانت الاستراتيجية المتبناة، فإنه من الصعب تقييم مستخلصات المصطلحات خصوصا بسبب الفوارق الموجودة بين البرمجيات التي تقوم بالمعالجة الآلية للنصوص المتخصصة، لأنها من جهة لا تعتمد على نفس التقنيات لإيجاد المصطلحات، ومن جهة أخرى لا تعتمد جميعها على المؤشرات نفسها. وإن كانت بالفعل تساعد المترجمين والمصطلحيين، فهي في نفس الوقت تطرح مجموعة من المشاكل من بينها:

– إن اللوائح التي تنتجها هذه المستخلصات تتضمن مرشحات لا تهم المستخدم وتغفل إدراج مصطلحات مهمة.

– بعض البرمجيات تركز أكثر على المصطلحات المركبة، وتستبعد المصطلحات البسيطة التي لا تظهر سوى مرة واحدة، والبعض الآخر يركز على الأسماء، ويضع جانبا المعطيات المصطلحية الأخرى، كالصفات.

– بعض المستخرجات لا تتعرف على الخصائص الصرفية التي تلحق المصطلحات، ولا تعتبرها أشكالا مختلفة من المصطلح نفسه فتُخَصِّصُ لها مدخلين في لائحة المرشحات.

6. خاتمة

لقد قمنا في هذا البحث بعرض أهم برمجيات استخلاص المصطلح ومعاينتها من خلال نصوص لغوية باللغة الفرنسية والإنجليزية. وتوصلنا إلى أن برمجيات استخلاص المصطلح وإن كان يتم تحسينها بانتظام، تُعتبر وسيلة لتسهيل عمل المصطلحي أو المترجم لمصطلحات ومفاهيم مجال معين،

لأنها تقترح على المستخدم لوائح من المصطلحات المرشحة الممكنة، وعلى المصطلحي تصنيفها والتحقق من مقبوليتها وانتقاء المصطلحات المرشحة منها القابلة للاستخدام.

الهوامش

¹ الفهرسة الآلية هي عملية تُمكن آليا بتنظيم وتصنيف البيانات الرقمية في الأنترنت وفقا لترتيب معين، وتعتمد في ذلك على برمجيات يطلق عليها اسم المفهرسات الآلية، بحيث تقوم بتحليل النصوص والمدونات الضخمة للبحث عن الكلمات المفتاحية أو المصطلحات أو جذوع يحتاجها المستخدم وتقدمها مرتبة على شكل سياقات استعمالها المختلفة.

² الأنطولوجيا هو لفظ يوناني يشير إلى فرع من فروع الفلسفة التحليلية، وتعني علم الوجود بما هو موجود. ونقصد بها هنا: التمثيل الصوري لمجموعة من المفاهيم وتصنيف معانيها ضمن مجال معين، وإبراز خصائص كل مفهوم ورسم العلاقات الدلالية بين هذه المفاهيم باستعمال لغة المنطق.

³ ينظر: أعمال لوم (2012) l'homme التي قامت بعرض وضعية المصطلحات التي تنتمي إلى المقولات النحوية غير الاسمية وتقديم المعايير التي تُمكن التعرف عليها.

⁴ ينظر a Natural Language processing Tool for Terminology Extraction ص 774 على الرابط التالي:

https://www.euralex.org/elx_proceedings/Euralex1996_2/040_Didier%20Bourigault,%20Isabelle%20Gonzalez-Mullier%20&%20Cecile%20Gros%20-LEXTER,%20a%20Natural%20Language%20Processing.pdf

⁵ ينظر شيرش وهانكس (1990) Church. Kenneth ward et patrick hanks ص 22-29 .

⁶ ينظر برمجية fivefilters على الرابط التالي:

<http://fivefilters.org/term-extraction>

⁷ ينظر برمجية Translated labs على الرابط التالي:

<http://labs.translated.net/terminology-extraction/>

⁸ ينظر برمجية Termine على الرابط التالي:

<http://www.nactem.ac.uk/software/termine/#form>

⁹ قائمة الألفاظ المستثناة التي تتكون من مجموعة من الألفاظ الأكثر شيوعا في لغة ما، مثل: الضمائر وحروف الجر وحروف العطف. ويمكن أن تعطى للبرنامج قائمة بهذه الألفاظ ليستبعدا ويتجاهلها.

¹⁰ ينظر برمجية Termine على الرابط التالي:

<http://www.nactem.ac.uk/software/termine/#form>

¹¹ ينظر برمجية Alchemy language API على الرابط التالي:

<http://www.alchemyapi.com/api/keyword-extraction>