بناء نواة نظام مساعد على إنشاء أنطولوجية عربية انطلاقاً من النصوص

د. ندى غنيم*

م. وسيم صافي** **م. مؤيد السيد علي*****

الملخص

معظم تطبيقات معالجة اللغات الطبيعية تحتاج إلى التعامل مع المستوى الدلالي للغة. ولهذا السبب ظهرت عدة محاولات لبناء أدوات معالجة دلالية للنصوص من خلال بناء الشبكات الدلالية كالأنطولوجية وشبكة الكلمات وشبكة الأطر، وشبكة الأفعَّال.

نعرض في هذا البحث المنهجية المعتمدة في بناء منصة ArOntoLearn، وهي بيئة عمل تساعد على بناء أنطولوجية عربية اعتماداً على النصوص في الوب، وأهم سمات هذه البيئة أنها تدعم اللغة العربية وتستخدم المعرفة السابقة في إجرائيات التعلم، فضلاً عن أنها تمثل الأنطولوجية الناتجة باستخدام نموذج الأنطولوجية الاحتمالي (Probabilistic Ontology Model (POM، الذي يمكن ترجمته إلى أي صيغة تمثيل للمعرفة. يقوم النظام بتحليل الموارد النصية العربية، يقابلها مع نماذج مفرادتية-نحوية بهدف تعلّم مفاهيم وعلاقات جديدة. إنّ دعم اللغة العربية ليس سهلاً نظراً لكون أدوات المعالجة اللغوية المتوافرة غير فعَّالة كفاية لمعالجة النصوص العربية غير المشكولة التي كذلك نادراً ما تتضمن علامات الترقيم الصحيحة المساعدة على التحليل الصحيح للجمل. لذلك حاولنا بناء بيئة عمل مرنة يمكن إعدادها بسهولة بحيث تُعَدَّلُ أدوات التحليل المستخدمة فيها وتُستَبْدَلُ بأخرى أكثر تطوراً عند توافرها.

الكلمات المفتاحية: تعلم الأنطولوجية، المعالجة الآلية للغة العربية، استحصال المعرفة، نموذج الأنطولوجية الاحتمالي.

* المعهد العالي للعلوم التطبيقية والتكنولوجية- دمشق.

** المعهد العالي للعلوم التطبيقية والتكنولوجية- دمشق.

*** قسم الذكاء الصنعي-كلية المعلوماتية-جامعة دمشق.

1.مقدمة

تحتاج أغلب تطبيقات معالجة اللغات الطبيعية إلى التعامل مع المستوى الدلالي للغة. لهذا السبب ظهرت عدة محاولات لبناء أدوات معالجة دلالية للنصوص من خلال بناء شبكات دلالية كالأنطولوجية، وشبكة الكلمات "WordNet" [4]، وشبكة الأطر "FrameNet" [2][3]، وشبكة الأفعّال "VerbNet" [1].

نعرض في هذه الورقة المنهجية المعتمدة لبناء بيئة عمل ArOntoLearn، بهدف المساعدة على إنشاء أنطولوجية عربية اعتماداً على النصوص في الوب.

تتضمن النصوص التي توصّف مجالاً معيناً توصيفاً كاملاً أو جزئياً لهذا المجال، أي إنَّ النصوص هي تعبير المؤلف عن الأنطولوجية التي لديه عن المجال الذي يكتب عنه باستخدام اللغة التي يتقنها. فإذا استطعنا إيجاد طريقة تعكس هذه العملية، أي تقوم باستنتاج أنطولوجية من النصوص أو تعلّمها نكون قد أسهمنا بتسريع عملية بناء الأنطولوجية.

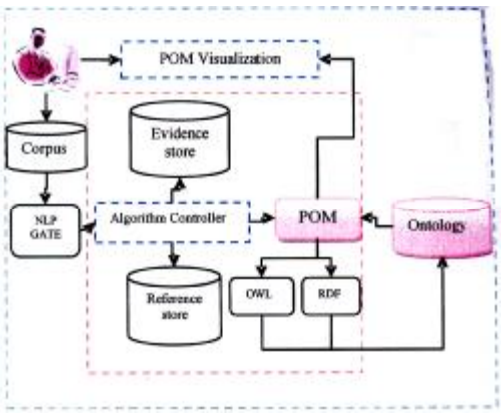
تتألف دورة حياة نظام تعلم الأنطولوجية عموماً من أربع مراحل رئيسة. تبدأ دورة الحياة باستيراد Import/Reuse أنطولوجيات مطورة مسبقاً أو إعادة استخدامها، ثم مرحلة استخراج الأنطولوجية Extract إذْ تُطبَّق خورازميات التعلم على المعطيات وتُعَالَجُ بهدف بناء نموذج تكراري وتزايدي يمثّل الأنطولوجية التي يجري تعلمها، وفي المرحلة الثالثة تجري عملية تشذيب Prune للأنطولوجية المستخرجة، إذْ قد تكون كبيرة الحجم وغير متسقة، أمَّا المرحلة الأخيرة فتجري فيها عمليات تصحيح Refine.

عند استعراض البحوث الجارية في هذا المجال، التي في مجملها لا تدعم اللغة العربية، نجد Mo'k workbench [8] الذي يعتمد طريقة التعلم التلقائي غير المشرف عليه لاستنتاج هرميات المفاهيم من مجموعات النصوص. من جهة أخرى، هناك أداة OntoLT [10]، المدموجة في محرر Protégé للأنطولوجية، وهي تعتمد بصورة كبيرة على التحليل اللغوي، وتستخدم أساساً البنية الداخلية للجمل الاسمية لاستنتاج معارف الأنطولوجية من النصوص. يركز إطار العمل OntoLearn [11] على مسألة إزالة اللبس في معاني الكلمات ويعرض خوارزمية جديدة SSI تعتمد على بنية الأنطولوجية. ينجز إطار العمل Text2Onto [5] العديد من الخوارزميات التي تتعلق بالمهام الجزئية لتعلم الأنطولوجية، مثل استخلاص المصطلحات term extraction، وإنشاء التصانيف taxonomy construction، وتقانات تعلم العلاقات بين المفاهيم.

نعرض في هذه الورقة بيئة العمل خاصتنا التي تستخدم نماذج مفرادتية-نحوية ومذيّلاً دلالياًsemantic annotator. لما كان دعم اللغة العربية ليس أمراً سهلاً، نظراً إلى أنَّ أدوات المعالجة اللغوية المتوافرة غير فعَّالة كفاية لمعالجة النصوص العربية غير المشكولة التي نادراً ما تتضمن علامات الترقيم الصحيحة، لذلك فقد حاولنا بناء بيئة عمل مرنة يمكن إعدادها بسهولة بحيث تُعَدَّلُ أدوات التحليل المستخدمة فيها وتُستَبْدَلُ بأخرى أكثر تطوراً عند توافرها.

2. بنيان النظام

يمكن تقسيم بنيان النظام إلى ثلاث وحدات: نموذج الأنطولوجية الاحتمالي Probabilistic Ontology Model (POM)، المتحكم بالخوارزمياتAlgorithms Controller، ووحدة معالجة اللغة العربية Arabic NLP tool (الشكل 1).



الشكل (1) بنيان نظام ArOntoLeam

يخزن POM نتائج مختلف خوارزميات تعلم الأنطولوجية التي يجري تهيئتها وتفعيلها من قبل المتحكم POM Controller الذي يقوم بتنظيم العمل ضمن نطاق النظام، إذْ يستخدم وحدة معالجة اللغات الطبيعية، باعتماد منصة العملGeneral Architecture for Text Engineering (GATE)، في عمليات المعالجة الأولية اللغوية للمعطيات، ومن ثم يقوم بتنفيذ خوارزميات تعلم الأنطولوجية وفق الترتيب الذي حدده مدير النظام. في النهاية، يجري عرض نتائج التعلم لمدير النظام ضمن واجهة بيانية، كما يمكن تصدير نموذج الأنطولوجية إلى لغة تمثيل معيارية مثل لغة أنطولوجية الوب Web Ontology Language (OWL) باستخدام وحدة كتابة الأنطولوجية.

a) نموذج الأنطولوجية الاحتمالي

يمكن التعبير عن نموذج الأنطولوجية الاحتمالي POM [5] بأنَّه مجموعة من كيانات النمذجة ذات القيمة instantiated modeling primitives المستقلة عن لغة تمثيل أنطولوجية معينة. كيانات النمذجة التي استخدمناها في نظامنا هي: (1) المفاهيم concepts، (2) المنتسخات instances، (3) الوراثة في المفاهيم concept inheritance (subclass-of)، (4) انتساخ المفاهيم concept instantiation (instance-of).

لا يعدُّ POM احتمالياً بالمعنى الرياضي، ولكن يأخذ كل كيانِ نمذجةٍ قيمةً تبيّن مدى ثقة الخوارزمية المستخدمة في إيجاد هذا المنتسخ. ويهدف هذا الاحتمال إلى تسهيل التفاعل مع المستخدم وتمكينه من ترشيح POM واختيار المنتسخات الصحيحة.

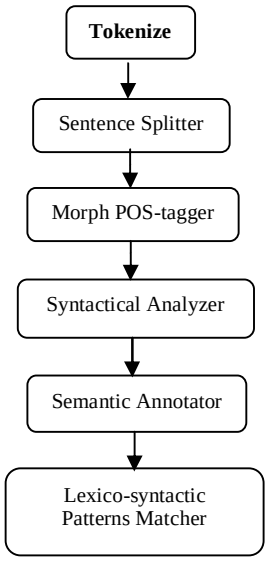
b) وحدة معالجة اللغة العربية

تعتمد هذه الوحدة على منصة العمل GATE [6]، وقد جرى اعتمادها لسببين: الأول مرونتها من حيث مجموعة الخوارزميات اللغوية المستخدمة، والثاني هو استخدام تمثيل Java Annotation Patterns Engine (JAPE) الذي يوفر محولاً منتهي الحالات لتمييز النماذج المفرداتية-النحوية، اعتماداً على التعابير النظامية regular expression.

يبدأ العمل في GATE (الشكل 2) بالتقطيع إذْ يقوم بتقسيم النص إلى مجموعة من الجمل عن طريق متابعة علامات الترقيم، ومن ثم تُقسَّمُ كل جملة إلى كتل بسيطة كالأرقام، وعلامات الترقيم، والكلمات، وذلك اعتماداً على مجموعة من قواعد التقطيع ليكون الناتج دخلاً إلى محلل صرفي ومحدد أنماط كلام Part of Speech Tagger [7] يحدد لكل مفردة نمطها ومجموعة تأشيرات annotation أخرى، فمثلاً لدينا لكلمة الدخل "الطفولة" الخرج الآتي:

token=الطفولة,
POS-tag='common_noun',
root='طفل',
pattern='الفعولة',
stem='طفولة',
stem pattern='فعولة'

هذه التأشيرات الناتجة ستكون دخلاً للمحلل النحوي (وقد استخدمنا لذلك المحلل النحوي Stanford [9]) الذي سيُسنِد الصنفَ النحوي إلى المفردات ويبني الشجرةَ النحوية للجملة. يجري بعد ذلك تنفيذ محول JAPE على المدونة التي جرى تأشير مفرداتها بهدف مطابقة مجموعة من النماذج المفرداتية-النحوية المطلوبة في خوارزميات تعلم الأنطولوجية.



الشكل (2) تسلسل التطبيقات ضمن GATE

مثال عن نماذج JAPE المستخدمة:

Pattern 1: 'is-a' relation
(syntaxnode.type =NP): instance
({Token.name in {هي،احدى ، احد، هو}})
(syntaxnode.type = NP): concept

يسعى هذا النموذج إلى مطابقة جملٍ مثل "دمشق إحدى المدن"، "سورية إحدى الدول"، "دمشق هي عاصمة"، "السيارة هي عربة"، "جامعة دمشق هي جامعة". وهو يعبّر عن علاقة is-a بين المنتسخ والمفهوم.

وقد قمنا بتطوير نواة من 15 نموذجاً مفرداتياً-نحوياً لالتقاط علاقات is-a وhas-a ضمن أشكال مختلفة من الجمل العربية.

c) المتحكم بالخوارزميات

هذه الوحدة تقود عملية التعلم وتنظم العمل ضمن كامل النظام، إذْ يقوم المتحكم باستدعاء المعالجة الأولية ضمن منصة العمل GATE، ثم تُنَفِّذُ خوارزميات التعلم وفق التسلسل الذي يحدده المستخدم، ويأخذ من هذه الخوارزميات طلبات التعديل الموجهة إلى نموذج الأنطولوجية ليقوم بتنفيذ هذه التعديلات، إذْ إنَّ الخوارزميات لا يسمح لها بالتعديل المباشر على نموذج الأنطولوجية، مما يسمح بتطبيق خوارزميات مركبة ومن ثم معالجة نتائجها قبل وصولها إلى نموذج الأنطولوجية.

إنّ تنفيذ أية خوارزمية يتم عبر ثلاث مراحل تبدأ بمرحلة التبليغ، حين تُبَلَّغُ الخوارزمية بتغييرات نصية حدثت ضمن المدونات مثل إضافة، حذف أو تعديل وثيقة، تليها مرحلة الحساب، حيث تقوم الخوارزمية بحساب تأثير هذه التغييرات في المفاهيم والعلاقات الموجودة لديها، وأخيراً تقوم الخوارزمية بتوليد طلبات تغيير في نموذج الأنطولوجية لتعكس التغييرات في المفاهيم والعلاقات، فمثلاً قد تطلب الخوارزمية إضافة مفهومٍ جديد، أو إحداث علاقة جديدة بين مفهومين في نموذج الأنطولوجية.

نستعرض فيما يأتي الخوارزمية المستخدمة في إجرائية التعلم، ونفصّل المنهجية المستخدمة لحساب احتمال كيان نمذجة ما:

1- خوارزميات استخراج المفاهيم والمنتسخات instances:

نفذنا مجموعة من الخوارزميات لحساب المعايير الآتية:

- RTF: وهو معيار لقياس تكرار المفاهيم المرشحة ضمن المدونة، له الصيغة الآتية: $Rtf(w) = \frac{rtf(i)}{rtf(all)}$

إذْ Rtf(i): تكرار المفهوم المرشح ضمن الوثيقة.

Rtf(all): تكرار المفهوم المرشح ضمن وثائق المدونة كلّها.

- TFIDF: وهو معيار إحصائي يعبّر عن أهمية المفهوم المرشح في الوثيقة ضمن المدونة [12]، إذْ تزداد أهمية المفهوم المرشح كلما ظهر أكثر ضمن الوثيقة. وله الصيغة للآتية: $tfidf(W) = tf(w).\log(\frac{N}{df(w)})$

إذْ Tf(w): عدد مرات تكرار المفهوم المرشح ضمن الوثيقة.

Df(w): عدد الوثائق التي تحتوي على المفهوم المرشح.

N: العدد الكلي للوثائق ضمن المدونة.

Tfidf(w): الأهمية النسبية للمفهوم المرشح ضمن الوثيقة w.

- الإنتروبي Entropy: وهو مؤشر لقياس كمية المعلومات [13]، له الصيغة الآتية:

$$E(w) = \frac{DR}{tR} * \log(\frac{DR}{tR})$$

- C-Value/NC-Value: مقياس إحصائي يعطي أهمية لتداخل المفاهيم المرشحة [14]، وله الصيغة الآتية:

$$c - value = \log/c/.(f(c)) - \frac{1}{nbc}\sum_{tce} f(Tc)$$

C: المفهوم المرشح لدخول الأنطولوجية.

F(c): تكرار المفهوم المرشح.

TC: مجموعة المفاهيم التي تحوي المفهوم المرشح c.

Nbc: عدد عناصر مجموعة المفاهيم التي تحوي المفهوم المرشح c.

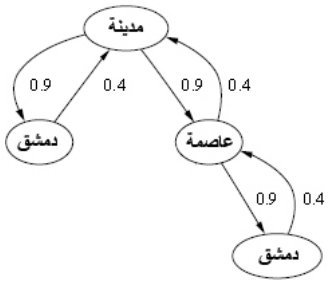
من أجل المعايير السابقة جميعها حُسِبَت القيمة النظامية normalized ضمن المجال [0..1]، واسْتُخْدِمتْ هذه القيمة كاحتمال للمفهوم ضمن نموذج الأنطولوجية الاحتمالي POM.

2-خوارزميات استخراج العلاقات:

لحساب وثوقية العلاقات المستخرجة، قمنا بتنفيذ عدة خوارزميات منها:

- التصنيف باعتماد علاقة التخصيص Hyponym في WordNet:

نستخدم أنطولوجية AWN باعتبار أنَ هيكليتها تعتمد على العلاقة IS-A، إذْ نقوم بتحديد مكان المفهومين طرفي العلاقة ضمن AWN ثم نقوم بحساب أقصر طريق يصل بين المفهومين عبر شجرة مفاهيم AWN وبناءً على عدد الوصلات في هذا الطريق تُسنَدُ قيمة الوثوقية إلى العلاقة بين المفهومين [15]، كما هو موضّح في الشكل (3)، فإن حساب وثوقية العلاقة بين "دمشق" و"مدينة"، يتحدد بحساب وزن أقصر طريق بين المفهومين مقسوماً على عدد الوصلات (في الشكل هو 0.9).



الشكل (3) مقتطف من أنطولوجية AWN

- التصنيف باعتماد نماذج مفهوم-منتسخ Concept-instance: تعتمد على التكرار النسبي للعلاقات التي اسْتُخرِجَتْ عن طريق النماذج المفرداتية- النحوية، فكلما ازداد تكرار العلاقة ضمن وثائق المدونة ازدادت معها الثقة بصحة العلاقة، إذْ نقوم بحساب تكرار العلاقة المرشحة ضمن الوثيقة N(rab i)، وحساب تكرار العلاقة المرشحة ضمن كاملة المدونة N(rab all)، ومن ثم نُطبِّقُ العلاقة الآتية:

$$Rtf(ra,b) = \frac{N(rabi)}{N(raball)}$$

- التصنيف باعتماد البحث في غوغل Google: تستخدم هذه الخوارزمية محرك البحث غوغل للتأكد من نتائج عملية التعلم، إذْ يُنَفَّذُ استعلام للعلاقة ضمن محرك البحث، ومن ثم يُستَخْدَمُ عدد النتائج التي يعيدها محرك البحث في حساب وثوقية العلاقة التي تزداد بزيادة عدد النتائج في محرك البحث. فمثلاً: لحساب وثوقية العلاقة:

(دمشق، عاصمة)Is-a

تُولَّدُ الاستعلامات الآتية: "دمشق هي عاصمة" (عدد نتائجها 19,700)، و"دمشق عاصمة" (عدد نتائجها 239,000).

ولحساب وثوقية العلاقة:

(دمشق، فاكهة)Is-a

يتم توليد الاستعلامات التالية: "دمشق هي فاكهة" (عدد نتائجها 0)، "دمشق فاكهة" (عدد نتائجها 4).

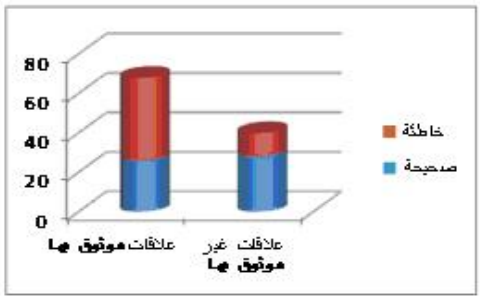
ولذلك فمن الواضح أن العلاقة الاولى أكثر موثوقية من العلاقة الثانية.

- تشابه السياق Context Similarity: إذْ نقوم بحساب تشابه السياق بين كياني العلاقة المرشحة لدخول الأنطولوجية. هذه الخوارزمية خاصة بعلاقة التشابه بين المفاهيم، وقد تستخدم أنطولوجية شبكة الكلمات AWN لحساب تشابه السياق، كما يمكن الاعتماد على التشابه المفرداتي بين المفهومين.

3. النتائج

لاختبار النظام، قمنا بتطبيقه على مجموعة من النصوص مستخلصة من الويكيبديا العربية. قمنا باختيار 125 وثيقة في مجال الدول والمدن في الوطن العربي. استخدمنا شبكة الكلمات العربية AWN كأنطولوجية مساعدة. وضعنا 15 نموذجاً مفرداتيا-نحوياً (5 من نمط instance-of، و10 علاقات subclass-of). ركزنا في الاختبارات على علاقة instance-of واعتمدنا على الخوارزمية (Pattern Concept-instance Classification) في حساب وثوقية العلاقات. لما كان تطبيق محلل ستانفورد التركيبي Stanford Syntactic Parser على الجمل العربية يتطلب أن تكون هذه الجمل جيدة الصياغة، وهذا ليس محققاً تماماً في جمل ويكيبيديا العربية، أجرينا نوعين من الاختبارات: الأول باستخدام محلل ستانفورد التركيبي على عدد محدود من الجمل (34 جملة)، والثاني باستخدام محلل صرفي فقط [7] على المدونة كلها (12779 جملة). جرى تقسيم النتائج وفقاً لعتبة ثقة تساوي 0.5، وكانت النتائج كما يأتي:

a) **باستخدام المحلل الصرفي فقط**: عدد العلاقات التي كانت الثقة فيها أعلى من 0.5 هو 68. بعد إجراء التحقق اليدوي فيها، تبيّن أن 26 علاقة منها صحيحة، و 42 منها مغلوط فيها. أمَّا عدد العلاقات التي كانت الثقة بها أقل من 0.5 فهو 40. بعد إجراء التحقق اليدوي منها، تبيّن أن 28 علاقات منها صحيحة، و12 منها مغلوط فيها. بالنتيجة فإن دقة precision هذه النتائج (أي نسبة النتائج الصحيحة من النتائج المعادة) هي 0.5.



الشكل (4) نتائج الاختبارات باستخدام المحلل الصرفي

b) **باستخدام محلل ستانفورد التركيبي أيضاً**: عدد العلاقات التي كانت الثقة فيها أعلى من 0.5 هو 10. بعد إجراء التحقق اليدوي منها، تبيّن أن 8 علاقات منها صحيحة، و2 منها مغلوط فيها. أمَّا عدد العلاقات التي كانت الثقة بها أقل من 0.5 فهو 2. بعد إجراء التحقق اليدوي منها، تبيّن أن العلاقتين صحيحتان. بالنتيجة فإن دقة هذه النتائج هي 0.83.

الشكل (5) نتائج الاختبارات باستخدام المحلل النحوي

نلاحظ أن استخدام محلل ستانفورد قد رفع من دقة النظام لأن بإمكاننا استخدام التأشيرات النحوية للجمل في النماذج المفرداتية-النحوية. لسوء الحظ، لم نستطع تنفيذ محلل ستانفورد على كامل المدونة لأنَّه يتطلب جملاً بسيطة جيدة الصياغة. سيجري في المرحلة المقبلة لاحقاً إجراء عملية تنقيح المدونة الكاملة السابقة وإجراء الاختبارات عليها.

	صرفي	نحوي
الجمل	12779	34
>= 0.5	68	10
True	26	8
False	42	2
< 0.5	40	2
True	28	2
False	12	0

الشكل (6) النتائج الكلية لاختبارات النظام

خاتمة:

يعدُّ تعلّم الأنطولوجية من النصوص الحل الأبرز لمشكلة بناء الأنطولوجية التي يحتاج إليها الوب الدلالي في تطبيقاته المختلفة، لكن عملية التعلّم بشكل آلي مهما بلغت دقتها فإنها لن تعطي نتائج نهائية كاملة وصحيحة، لذا لابدَّ من مرحلة لاحقة لتعديل هذه النتائج واستكمالها والتوافق عليها من قبل الخبراء.

تعدُّ النتائج التي تعطيها النواة كافية لبعض تطبيقات الوب الدلالي مثل العنقدة (Clustering) واستخراج المعلومات، ولكن هذه النواة تعدُّ مرحلة من مراحل مشروع متكامل لبناء تطبيقات الوب الدلالي للمحتوى العربي. لذا لابدَّ من استكمال بناء النواة واختبارها، ولابدَّ أيضاً من بناء باقي الأجزاء التي تحتاج إليها تطبيقات الوب الدلالي مثل أدوات تخزين الأنطولوجية والبحث فيها، وأدوات وسم الوثائق بالأنطولوجية ontology-based document annotation وتصنيف هذه الوثائق اعتماداً على هذا الوسم، وأدوات عرض الأنطولوجية، وغيرها.

المراجع:

1. VerbNet, from the site: http://verbs.colorado.edu/~mpalmer/projects/verbnet/downloads.html
2. John.son, C., Fillmore, C., Petruck, M. Baker, C., Ellsworth, M., Ruppenhofer, J., and Wood, E. 2002. FrameNet: Theory and Practice, from http://www.icsi.Berkeley.edu / framenet.
3. Josef Ruppenhofer, MichaelEllsworth, Miriam R. L. Petruck, Christopher R. Johnson, Jan Scheffczyk. "Frame Net II :Extended Theory and Practice", 2006.
4. WordNet. Retrieved June 2009, from http//:www.globalwordnet.org
5. P. Cimiano and J. Volker. Text2onto – a framework for ontology learning and data-driven change discovery, Proc. NLDB 2005, Lecture Notes in Computer Science, vol. 3513, Springer, Alicante, 2005, pp. 227-238.
6. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications, 2002, In Proceedings of the 40th Annual Meeting of the ACL.
7. R. Sonbol, N. Ghneim, M. Desouki, Arabic Morphological Analysis: a new approach, ICTTA 2008, Damascus, Syria.
8. G. Bisson, c. Nedellec, and L.canamero. Designing clustering methods for ontology building – The Mo'K workbench. In proceedings of the ECAI ontology Learning Workshop, pages 13-19, 2000.
9. Arabic Stanford parser, Retrieved October 2,2008, from : http//www.nlp.stanford.edu/software/parser-faq.shtml.
10. P. Buitelaar, D. Olejnik, and M. Sintek. OntoLT: A protégé plug-in for ontology extraction from text, 2003, In Proceeding of the international Semantic Web Conference, (ISWC).
11. P. Velardi, R. Navigli, A. Cuchiarelli, and F. Neri. Evaluation of ontolearn, a methodology for automatic population of domain Ontologies, 2005, In P. Buitelaar.
12. G. Salton. Developments in automatic text retrieval. 253:974–979, 1991.
13. R. M. Gray. Entropy and Information Theory. Springer, 1990.
14. J. Tsuji K. Frantzi, S. Ananiadou. The c-value/nc-value method of automatic recognition for multi -word terms. In Proceedings of the ECDL, pages585–604, 1998.
15. G. Miller. WordNet Hyponym Classification: A lexical database for English. Communications of the ACM, 38(11):39–41, 1995.

مسرد المصطلحات

semantic annotator	مذيّل دلالي
Prune	تشذيب
Refine	تصحيح
term extraction	استخلاص المصطلحات
taxonomy construction	إنشاء التصانيف
regular expression	التعابير النظامية
Probabilistic Ontology Model POM	نموذج الأنطولوجية الاحتمالي
Algorithms Controller	المتحكم بالخوارزميات
Arabic NLP tool	وحدة معالجة اللغة العربية
instantiated modeling primitives	كيانات النمذجة ذات القيمة

تاريخ ورود البحث إلى مجلة جامعة دمشق 2011/10/19.