استكشاف وتمثيل المعرفة من النص العربي

بطريقة هجينة

د.يسرى مالك ضمد

جامعة البصرة / كلية العلوم / قسم علوم الحاسبات

المستخلص:

يقدم هذا البحث طريقة هجينة في تمثيل المعرفة للجمل ذات التراكيب القواعدية المختلفة في النص العربي مستفيدين من خصائص كل منها في التمثيل الاكفأ من خلال قدرتها على التعبير بدقة عن المواقف وعلى تمثيل الترابط بين الكيانات ومتجنبين نقاط ضعف كل منها .

استخدمنا شبكات الدلالة (semantic network)لتمثيل المعرفة للجمل التي تحتوي على فعل في تركيبها بالاعتماد على قواعد الحالة(case grammar) في تحليل الجملة اما الجمل التي لاتحتوي على فعل فتمثل باطار (frame) بالاضافة الى استخدام الهيكل الشجرى للاسماء الذي تم ربطه مع كل اسم في الجملة لتسهيل عملية جلب السمات الدلالية له ثم تقدم المعرفة بشكل قاعدة بيانات دلالية غنية بالمعرفة يمكن الاستفادة منها في مختلف تطبيقات معالجة اللغات الطبيعية منها فهم النصوص، الترجمة الالية، تلخيص النصوص الالي ،الاجابة عن الاستفسارات وغيرها.

الكلمات المفتاحية:تمثيل المعرفة، شبكات الدلالة، قواعد الحالة، الاطر.

1 - المقدمة:

يعد مجال تمثيل المعرفة (knowledge representation) من الحقول المهمة في معالجة اللغات الطبيعية اليا اذ يسهل عملية محاكاة طريقة التفكير البشري بواسطة الحاسوب . ان عملية تمثيل المعرفة تسبقها عملية استكشاف لتلك المعرفة ثم استخراجها وبالتالي يمكن تمثيلها باحدى تقنيات التمثيل. وآلية تنقيب النصوص (text mining) هي أحدى الطرائق المستخدمة لاستخراج المعرفة المفيدة وغير الظاهرة بكميات كبيرة من النصوص غير المنتظمة ، و بتعبير آخر تحويل النص الحر إلى نص يمكن تحليله آليا تتم هذه العملية بمعالجة أولية للنص ، وذلك باستخراج الكلمات والمفاهيم ، و من ثم إيجاد العلاقات بين المفاهيم وتمثيل النص في قواعد ربطه وتصنيفه وإمكانية عرضه للمستخدم بطريقة سهلة الفهم [Witten ,2005].

لدى تصميم نظام لتمثيل المعرفة يجب أن نهتم بالكيفية التي يستخدم بها النظام ذلك التمثيل في التشغيل ليكون ناجحاً . ويجب أن نهتم بالطريقة التي تتغير بها فاعليّة النظام تبعاً لكمية المعرفة الموجودة فيه فيجب أن يكون في وسعنا إضافة حقائق جديدة من دون أن تقلقنا التفاصيل المتعلقة بكيفية ارتباطها بحقائق أخرى. فبعض التمثيلات تصلح للكميات

الصغيرة، لكنها تنفجر بصورة أسية عند إضافة معلومات أكثر وتمثيلات اخرى أقل حساسية للحجم[اديب،2010] لهذا يجب اختيار طريقة مناسبة لتمثيل المعرفة تلبي حاجة النظام الذي يستخدمها بالاضافة الى تحديد انواع العلاقات المطلوبة.

احدى الصعوبات في معالجة نصوص اللغة الطبيعية حاسوبيا هي كيفية تمثيل المعرفة بطريقة تصبح فيها مفيدة للتحليل[Christopher,2000] . من هنا كانت الحاجة الى دمج طرائق مختلفة للتمثيل مثل الشبكات الدلالية للاستفادة من قوتها بقدرتها على العثور على الارتباطات اي كيفية ايجاد مجموعة الحقائق بينما الانواع الاخرى تركز على كيفية تطبيق الحقائق حين العثور عليها.

سنقدم في الفقرات الاحقة تعريف للمعرفة وكيفية اكتشافها وتحليلها ثم استخراجها ومن ثم الطريقة المقترحة لتمثيل النص المكتوب باللغة العربية وذلك بدمج مجموعة من الاليات للوصول الى بناء قاعدة بيانات دلالية غنية بالمعرفة اللازمة لمختلف تطبيقات معالجة اللغات الطبيعية.

2 - **المعرفة** knowledge :

المعرفة كمفهوم ندركه عقليا له اربعة مواصفات اساسية هي:

1. المظهرية : وتظهر لنا المعرفة على شكل شفرة تؤثر بها على حواسنا وهذا الشكل قد يكون صوريا او صوتيا او رمزيا كاللغات وغيرها من الاشكال.
2. الادراك : الاشكال التي تظهر بها المعرفة يجب ان تؤثر في الانسان وتحفز الادراك لديه ليتعامل معها بعد ذلك .
3. التخزين : المعرفة المخزونة في الدماغ مسبقا مع المعلومات المستلمة كمؤثرات والاجهزة الحسية التي تدرك وتتفاعل مع الحالة نجد ان الاستجابات مختلفة من شخص الى اخر.
4. الاستمرارية : ان تاثير المؤثرات قد يكون وقتي ويختفي بعد لحظة التاثير وفي هذه الحالة يكون التاثير على الذاكرة القصيرة المدى وعكسه يكون التاثير على الذاكرة البعيدة المدى.

فالمعرفة تمثل العلاقات بين المعلومات وتكون غالبا على شكل تنقيبات واستنتاجات اي ان:

Knowledge= information + relation

3 - **اكتشاف المعرفة** knowledge discovery :

عملية الاكتشاف تعني استخلاص المعرفة من البيانات (خلاصات بيانات ،الموديلات، العلاقات بين البيانات،.....) وهي عملية معقدة وغير مباشرة تتألف من مجموعة من المراحل المترابطة وهي [Feldman , Ronen & Sanger , James , 2007]:

- تحديد البيانات واستخراجها بحسب الأهداف المنتظرة .
- معالجة البيانات وتنظيفها كإلغاءالمعلومات المتكررة والتصحيح الشكلي ومعالجة البيانات الناقصة .
- تعديل المعلومات بشكل يتلاءم مع هدف استخراجها.
- اختيار كيفية استخراج المعلومات ، إما من أجل دراسة الخصائص العامة للمعلومات المستخرجة أو من خلال دراسة تطوير المعلومات في المستقبل و تخمينها .
- التصنيف : وهو إيجاد مجموعات من المعلومات بناء على خصائص مشتركة .
- الربط والتسلسل : وهو استخراج العلاقة النسبية بين البيانات .
- التأكد من المعلومات المستخرجة .
- عرض النتائج بطريقة سهلة يمكن ان تساعد على تحليلها .

4 - **تحليل المعرفة** knowledge analyses :

يعرف تحليل المعرفة كمعرفة لسانية ماذا تشير اليه الكلمات اي معناها مجتمعة ضمن النص والسياق الذي ظهرت به . والمعرفة نوعان هما:

- المعرفة التحليلية القواعدية (syntactic) knowledge تتضمن :

1. تحليل الصنف (اسم، فعل، حرف،.....).
2. كيفية حدوث وتنظيم هذه الاصناف.

- المعرفة الدلالية (semantic knowledge) تعتمد على صفات الكلمات اي السمات الدلالية لها. [آلان،1993]

5 - تمثيل المعرفة knowledge representation:

يعد احد مجالات الذكاء الاصطناعي وهدفه ايجاد طريقة لتمثيل الحقائق والمعارف من خلال رموز بحيث يمكننا تعريف مجال المشكلة او عالمها (domain of discourse) ومن ثم تعريف عمليات تمكن من استنباط حقائق جديدة مشتقة من المعطيات ، وبالتالي فان تمثيل المعرفة يعني ايجاد طريقة موحدة تمكن الحاسوب من محاكاة طريقة التفكير البشري. [Elaine,1983]

وهنالك عدة اليات لتمثيل المعرفة تم استخدام بعض منها في البحث وهي:

1. **الشبكات الدلالية semantic network**:

اقترحت من قبل العالم النفسي كولين عام 1968 واستخدمت كهيكل بياني وكتمثيل معجمي لمعاني اللغة الانكليزية .وهي عبارة عن عقد(nodes)تمثل معلومات مخزونة ترتبط فيما بينها باقواس(arcs) معنونة تمثل العلاقات وتعتبر وسيلة طبيعية وبسيطة لوصف الاشياء [enn,2007].

المعلومات المتوفرة في الشبكة يمكن اجراء عمليتين اساسيتين عليهما الاستنتاج من خلال الروابط بين الكيانات و بحث التقاطع وفيه يتم عندما تتقاطع رابطتان يكون الممر بينهما اقصر مجموعة من الروابط تصل بين شيئين [Elaine,1983].

2. **الاطار fram** :

ان الاطار هو هيكل دلالي اكثر تعقيد في تصميمه حيث يسمح بتمثيل الصفات المشتركة للكيانات تم اقتراحه من قبل (العالم منسكي) . حيث تقولب الكيانات المتشابهه في اطار عام بينما توضع صفاتها الخاصة في اطر جزئية ترتبط بالاطار العام. [clive &Raymond,1991]

6 - الطريقة المقترحة:

اعتمدنا في بحثنا أستخدام عدة اليات لتمثيل المعرفة للتغلب على نقاط ضعف كل الية ولزيادة قوة طريقتنا في تمثيل المعرفة المستكشفة من النص العربي. تتضمن الطريقة عدة معالجات كما موضح المخطط العام لها في الشكل(1) وهي:

1. ***مهمة ادخال النص***..وفيها يتم ادخال النص المكتوب باللغة العربية مباشرة الى الحاسوب او عن طريقة غير مباشرة من خلال ملف مخزون بصيغة doc.

مثال:" ارتفع سعر الدولار الامريكي بعد الاحداث التي اجتاحت منطقة الشرق الاوسط....."

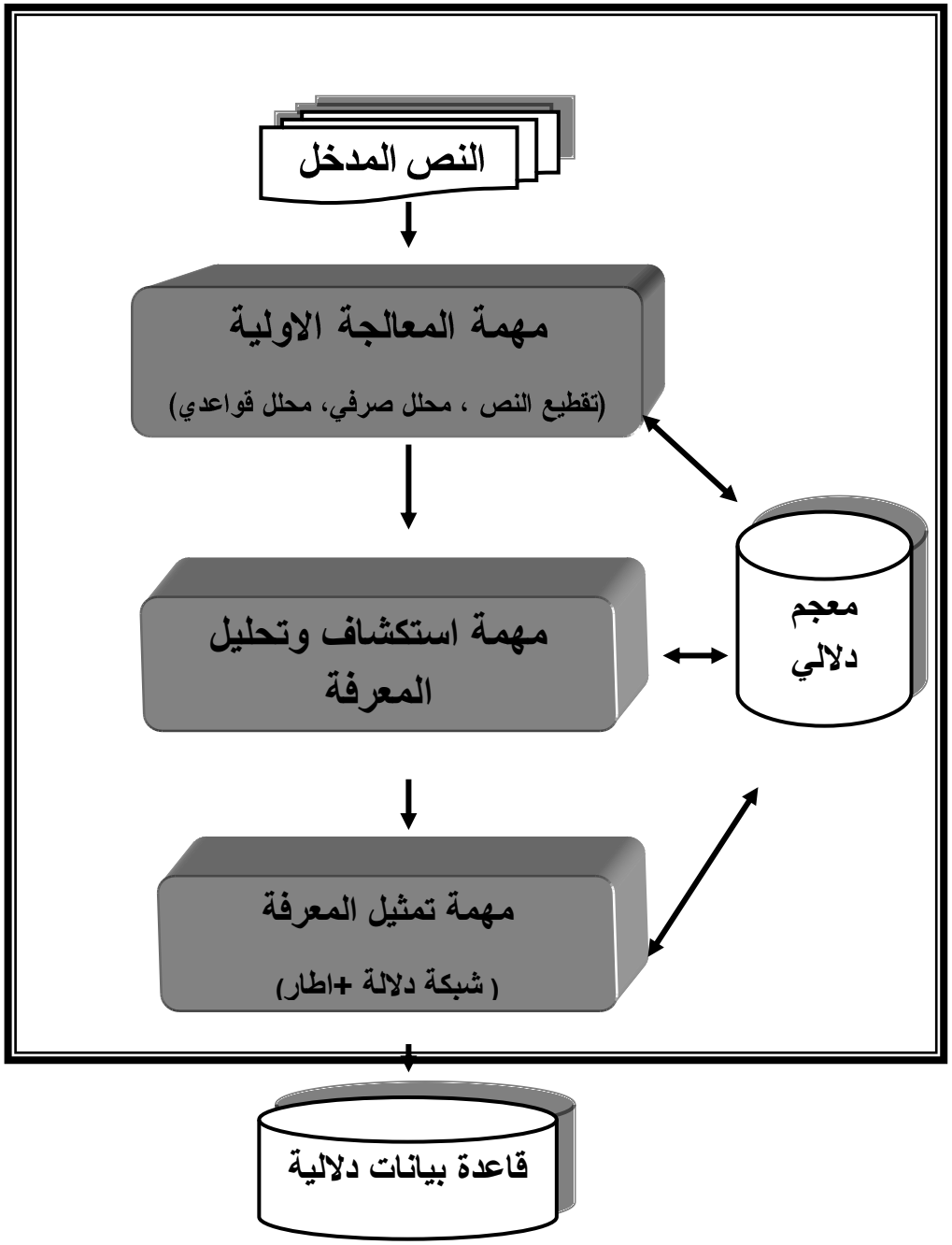
2. ***مهمة المعالجة الاولية*** ..هذه المهة تتضمن عدة معالجات تجرى على النص المدخل وهي:

- **معالجة تقطيع النص** .. وفيه يتم تقطيع النص الى كلمات وجمل. حيث يتحول الى مصفوفة من الكلمات واخرى الى جمل.

$$D=\{S_I\} \quad i=1..m$$

$$D=\{w_j\} \quad j=1..n$$

حيث m هي عدد الجمل في النص و n عدد الكلمات في النص.



الشكل(1) مخطط عام للطريقة المقترحة

- **معالجة التحليل الصرفي** ..هذه المعالجة ترجع كل كلمة تحتوي على لواصق (سوابق او لواحق) الى جذعها بالتعاون مع المعجم الساند للطريقة المقترحة.

- **المعجم الساند** : نظم المعجم بشكل جداول تحتوي على جميع اصناف الكلام (فعل،اسم،حرف،صفة،ظرف زمان، ظرف مكان)للغة العربية (لغة البحث) اذ وضع كل صنف في جدول خاص مع جميع المعلومات المرتبطة بكل مفردة من معنى وسياق وعلاقات اخرى لغوية.

نظم صنف الاسم بشكل هيكل شجري لضمان توارث الصفات المشتركة لها ليسهل عملية الحصول على جميع المعلومات الخاصة بكل اسم عند الحاجة.

- **معالجة التحليل القواعدي**.. وهي المعالجة المسؤولة عن تحديد نمط القاعدة لكل جملة في النص بالتعاون مع المعجم الذي يزودها باصناف الكلمات (1 فعل ،2 اسم ،3 حرف،4 صفة ،5 ظرف مكان ، 6 ظرف زمان). مخرجات هذه المعالجة قائمة بتسلسل الجمل مع النمط القواعدي لكل منها.وتكون على نوعين:
  - **جمل بسيطة** وتتكون من نمط واحد من الانماط القواعدية مثل:فعل فاعل مفعول به .
  - **جمل مركبة** وتتكون من مجموعة من الجمل البسيطة وتراكيب مختلفة مثل الجمل الظرفية ،الجار والمجرور،...وغيرها.

3. ***مهمة استكشاف وتحليل المعرفة*** .. في هذه المرحلة تم تصنيف الجمل التي رحلت من المعالجة السابقة الى قائمتين أعتمادا على وجود صنف الفعل في نمطها القواعدي هما:
   - ❖ **قائمة الجمل الفعلية** التي تحتوي على فعل او اكثرفي نمطها القواعدي.
   - ❖ **قائمة الجمل الاسمية** التي لاتحتوي على فعل في نمطها القواعدي.

ان تصنيفنا الى الجمل لم يكن اعتباطيا وانما لاجل تمثيل كل صنف منها بآلية مناسبة تمثل المعرفة فيها باحسن تمثيل كما سيوضح بمهمة تمثيل المعرفة اللاحقة.

بعد فرز الجمل الاسمية والجمل الفعلية كل منها في قائمة منفصلة جدول(1) و(2) هيئت جميع المعلومات الدلالية لكل مفردة فيها من المعجم الساند مثل (المعنى، المرادف، التضاد، السمة الدلالية للمفردة،.....الخ).

جدول (1) قائمة الجمل الاسمية في النص

ت	الجمل الاسمية
2	والجنود مستعدون للانقضاض على العدو.
3	ان جميع جنود الكتيبة مخلصون في تأدية واجبهم.
.	.....
.	.....

جدول (2) قائمة الجمل الفعلية في النص

ت	الجمل الفعلية
1	تحركت قطعات الجيش دفاعا عن الوطن .
4	حيث رصدت فعاليات اللارهابيين وحددت اماكن تواجدهم.
.	.....
.	.....

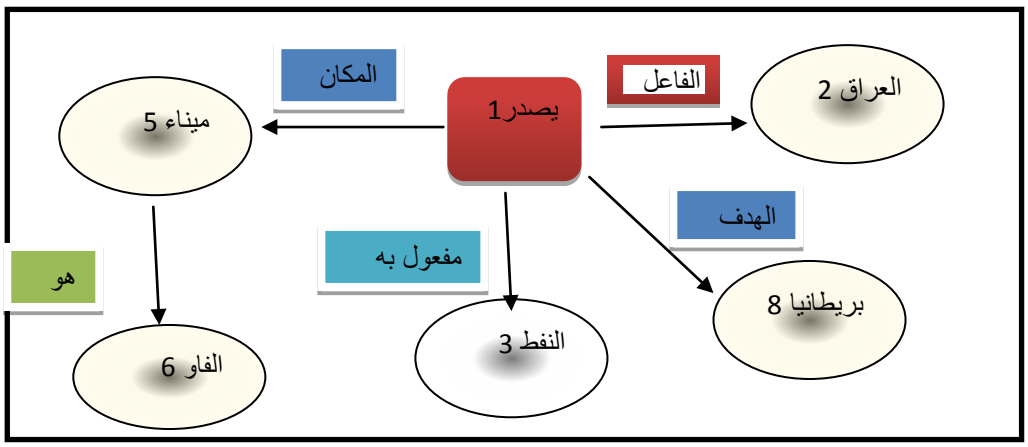
مهمة تمثيل المعرفة..

بعد تهيئة أصناف الجمل مع جميع المعلومات الدلالية والقواعدية لكل منها تبدا عملية اختيار آلية تمثيل المعرفة المناسبة لها تحقق افضل تمثيل مستفيدين من ميزات كل الية ومتجنبين ضعفها كما يلي:

4-1آلية تمثيل الجمل الفعلية:

ان خصوصية الفعل والذي يعتبر مركز ثقل الجملة الذي تتمحور باقي المفردات حوله، كانت الشبكات الدلالية هي افضل طريقة لتمثيل الجملة الفعلية . اذ يمكن وصف الجملة بشكل شبكة دلالية كما موضح بالشكل (2) بالاعتماد على قواعد الحالة له والتي تعتبر تمثيل بين المعرفة القواعدية والمعرفة الدلالية في اللغة وهي المعرفة اللسانية، اقترحت من قبل العالم فلمور. تعتبر هذه القواعد ان المعرفة اللغوية تنتظم حول الفعل او بدقة اكثر(معنى الفعل) حيث تربط مجموعة من الحالات مع اي معنى للفعل وقد تكون حالات اجبارية او اختيارية [Nicholas,1990] تستخدم في تحليل الجملة لتمثيل معناها بشبكة دلالية بالتالي يمكن الاستفادة منها في استرجاع المعلومات المتوفرة في الجملة للاجابة على اي تساؤل عن الحدث في الجملة مثل (من، ماذا، لمن، اين، ....).

الجملة: يصدر العراق النفط من ميناء الفاو الى بريطانيا.



الشكل(2)شبكة الدلالة للجملة

سنقوم بوصف العلاقات التي تربط بين العقد في الشبكة الدلالية بشكل جدول رقم (3) يحتوي الجدول على مجموعة من العلاقات التي تربط بين مفردات الجملة الفعلية وكما بينا سابقا هنالك علاقات اجبارية واخرى اختيارية ومن السهولة اضافة اي علاقة جديدة ممكن ان تظهر بين المفردات في النص الى الجدول.

جدول رقم (3)

code	Relation	*No.rel*
a	*Agent*	***1***
o	*Object*	***2***
b	beneficiary	**3**
i	Isa	**4**
d	*direction*	**5**
n	*Owner*	***6***
l	location	**7**

نظمنا الشبكة الدلالية لكل جملة بشكل ازواج مرتبة لمفرداتها تشمل العقدة المصدر snod مع العقدة الهدف onod والعلاقة التي تربطهما rel كما يلي:

Rel(snod, onod)

وبذلك استطعنا ان نمثل الشبكة بشكل جدول يسهل التعامل معه برمجيا واجراء جميع العمليات عليه مثل(البحث، الاضافة، التعديل،..... الخ).

ولتمثيل شبكة المثال في الشكل(2) بشكل جدول (4) حيث ترقم المفردات حسب تسلسلها في الجملة .

الجدول رقم (4)

Rel. code	**no. onod**	**no. Snod**	**No. se**
a	**2**	**1**	**S1**
o	**3**	**1**	
l	**5**	**1**	
D	**8**	**1**	
i	**6**	**5**	

تتوسع معلومات شبكة الدلالة في الشكل (2) من خلال الهيكل الشجري للاسماء في المعجم الساند لكل اسم يتمحور حول الفعل الذي يدعمها بقوة معرفية غنية بالمعلومات . وبهذا تمثل كل جمل النص الفعلية في جدول واحد.

رغم مميزات شبكات الدلالة في قوتها بتمثيل المعرفة وبساطتها في استخدام المنطق للاجابة عن التساؤلات وسهولتها في بناء التساؤلات الا ان مشكلتها في عدم مقدرتها التعبير عن الحقائق الاكثر تعقيد مثل تلك التي تتضمن قياسات كمية بالاضافة الى ذلك تصبح عملية السيطرة على الشبكة معقدة جدا عندما يكبر حجمها ولهذا السبب يجب دمج انواع تمثيل اخرى للحصول على تمثيل دلالي جيد لمعلومات النص .

4-2 آلية تمثيل الجمل الاسمية:

اما الجمل التي مفرداتها هي عبارة عن مجموعة من اسماء ،صفات ،ظرف زمان اوظرف مكان جميع هذه الاصناف من الكلام لها سمات دلالية خاصة بها . تتميز الجمل التي لاتحتوي على فعل بانها جمل وصفية لحال، زمان، مكان اواشخاص وعلى هذا الاساس يجب ان نختار هيكل تمثيل معرفي مناسب لها وهو الاطار وربط مفرداته

من نوع الاسماء بهيكلها الشجري كما وضحنا في التمثيل السابق.

تم اختيار مجموعة من الصفات واضيفت صفة اخرى في حالة اضافة صفة جديدة ومن الممكن اضافة عدد اكبر من

جدول (5) الاطار للجملة الاسمية

other	Quality	Color	Instrument	location	time	object	semantic	class

من الملاحظ ان هنالك بعض العلاقات المتشابه في الشبكة الدلالية والاطار وهذا شئ طبيعي لان التمثيلين يستخدمان الصفات الى الجدول ليشمل كل الحالات بسهولة وكما موضح في الجدول(5) قد تكون هنالك حقول فارغة وقد يملئ حقل بهيكل دلالي اخر حسب المعلومات الدلالية الممكن توفرها من المعجم. تمثل جميع الجمل في جدول واحد كما في الالية السابقة.

لوصف مفردات اللغة من اسماء وصفات وغيرها التي تكون سماتها الدلالية واحدة في اي تمثيل.

سوف نوضح التمثيل بشكل جدول (6) للمثال التالي :

الجملة: طلبة جامعة البصرة متميزون رغم الظروف الصعبة.

جدول(6)تمثيل الجملة الاسمية، *تدل على هيكل يربط في هذا الحقل

other	instrument	location	time	Object	semantic *	class	word	No. word	No. sen
						اسم	طلبة	1	S1
		1		1		اسم	جامعة	2	
		1		2		اسم	البصرة	3	
				1		صفة	**متميزون**	4	
						حرف	رغم	5	
						اسم	**الظروف**	6	
				6		صفة	صعبة	7	
			:			:		:	S2
			:			:		:	
								:	

من خلال هذه المجموعة من التمثيلات لكل جمل النص تتكون لدينا قاعدة بيانات دلالية غنية بالمعلومات توفر لاي نظام من انظمة معالجة اللغات الطبيعية التي تعتمد على المعرفة في تطبيقها جميع المعلومات المستكشفة من النص.

5.الاستنتاج والعمل المستقبلي:

كل نوع من انواع هياكل تمثيل المعرفة له نقاط قوة وضعف ولهذا عملية دمج اكثر من هيكل لتكوين طريقة هجينة حسن من تمثيل اكبر قدر ممكن من المعرفة يمكن استخدامها في التطبيقات التي تحتاجها. كما ان تعقيد القاعدة التركيبية للجملة يؤثر على كفاءة التمثيل.

وتعتبر البحث بداية في طريق تمثيل المعرفة للنص العربي حيث لم يعالج فيه عائدية الضمائر في النص ومشاكل الغموض القواعدي للتراكيب وترك كعمل مستقبلي ان شاء الله.

المصادر

Christopher s. butler , "**computation and language**",(2000), [www.pdffactory.com].

Clive l. dym &Raymond e. levitt, "**knowledge_based systems in engineering**", mcgraw-will, inc., (1991).

Elaine rich, "**artificial intelligence**", mcgraw-will, inc., (1993).

Enn tyugu, "**algorithms and architectures of artificial intelligence**", tallinn University of technology, Estonia, iso press (2007).

Feldman , Ronen & Sanger , James , ***"The Text mining Handbook : Advanced Approaches in Analyzing Unstructured Data "*** , Cambridge University PRESS , (2007).

Nicholas r. & john m. , "**using semantic networks for data base management**", university of Toronto, (1990)

Witten , Ian H. , ***"Text Mining ",*** University of Walkato , (2005).

اديب يوسف شيش، **"التفكير مطالعات في علم المعرفة"**، منشورات وزارة الثقافة-الهيئة العامة السورية للكتاب، (2010).

الان بونيه، **" الذكاء الاصطناعي واقعه ومستقبله "**، ترجمة علي صبري فرغلي، عالم المعرفة،(1993).

Knowledge discovery and representation from Arabic text by hybrid method

Yusra Malik Dummad

Computer department

science College/Basra University

Abstract:

This paper presents hybrid method of Knowledge representation for sentences with different grammar component in Arabic text, with taking the benefit from attributes of each one in the optimize represent by their ability in expression exactly about objects and associative representation between them and avoid debility points for each one.

We use semantic networks for Knowledge representation for the sentences which contain the verb in its component depending on case grammar for sentence analyses, other sentences that are not contain the verb represented by frame then the knowledge presented as semantic data base advantageous in different natural language processing applications such as language understanding, automatic translation, automatic text summarization, answering about the questions, etc.