أساليب مقاربة جماعية لإزالة الإلتباس في أسماء الكيانات

مقدمة من : أيمن عنتر الهلباوي

الأطروحة تم تقديمها للحصول علي درجة دكتوراة الفلسفة في علوم الحاسب

قسم علوم الحاسب

جامعة شيفلد

المملكة المتحدة

يوليو٢٠١٣

ملخص الرسالة

أصبحت الأنترنت أحد أهم مصادر المعلومات. غالب هذه المعلومات موجودة في صورة نصوص لغوية طبيعية (كما يستخدمها الإنسان في الطبيعة وليست نصوص مخلقة) . واحد من أهم مكونات هذه النصوص الطبيعية هو أسماء الكيانات مثل أسماء الأشخاص والمؤسسات والأماكن...إلخ. ولذلك كان التعرف علي أسماء الكيانات وتصنيفها في النصوص اللغوية الطبيعية محور اهتمام الكثير من الباحثين لعدة سنوات. وكما أن الأسم النصي الواحد قد يشير إلي عدة كيانات فإن الكيان الواحد قد يشار إليه بعدة أسماء نصية. وهنا يكون هناك مشكلة في تمييز الكيان عند الإشارة النصية لأسمه وهذا النوع من المشاكل معروف في معالجة اللغات الطبيعية بمشكلة فك الألتباس. وفك الألتباس لأسماء الكيانات تشير إلي مهمة ربط الإشارات النصية لأسماء الكيانات في النص المكتوب بسجل التعريف لهذا الكيان داخل قاعدة معلومات معينة. وتعد مسألة فك الألتباس لأسماء الكيانات في النصوص المكتوبة في غاية الأهمية للكثير من تطبيقات اللغة المحوسبة مثل محركات البحث النصي علي شبكة الأنترنت وكذلك للوحدات البرمجية التي تعمل كعميل يهدف إلي جمع معلومات عن الكيانات الحقيقية من جميع المصادر المتاحة علي الأنترنت. ويعد الهدف الرئيسي من هذا البحث هو تطوير مجموعة من الآليات الجديدة لفك الألتباس للإشارات النصية لأسماء الكيانات المبهمة استنادا علي الأرتباط الضمني بين أسماء الكيانات المختلفة داخل نفس الوثيقة.

وتركز هذه الأطروحة علي مشكلتين وثيقتي الصلة بعملية فك الألتباس. المشكلة الأولي هي عملية توليد جميع الأحتمالات الممكنة لأسم الكيان الملتبس بحيث تكون هذه الأحتمالات أقل ما يمكن وأن يكون السجل الصحيح – سجل الكيان داخل قاعدة المعلومات – موجود داخل قائمة الأحتمالات الممكنة. أما المشكلة الثانية فهي مشكلة فك الألتباس باستخدام المقاربة الجماعية بين جميع الأحتمالات الممكنة لجميع أسماء الكيانات الملتبسة في نفس الوثيقة واستخدام علاقات التراكب المنطقي بين أسماء الكيانات المختلفة وقد تم استخدام ويكيبيديا كقاعدة معلومات مرجعية لفك الألتباس.

تم استخدام آليات ونظريات استرجاع المعلومات لتوليد أو تخليق قائمة من أسماء الكيانات المحتملة لكل إشارة لأسم كيان في النص. كما تم تقديم دالة جديدة (NEBSim) لقياس تشابه الوثائق إعتمادا علي الظهور الآني لأسماء الكيانات بفرض وجود إشارة إلي أسم كيان معين. وقد تم إستخدام هذه الدالة الجديدة بالتوازي مع استخدام دالة انحراف جيب الزاوية لتعليم نموذج لترتيب أسماء الكيانات المقترحة. وقد تم أستخدام نموذجي التعليم (مصنفات)Naıve Bayes وSVM لإعادة ترتيب قائمة المقترحات. تم تقييم هذا النموذج بإجراء مجموعة من التجارب باستخدام حالات قياسية تم نشرها من خلال TACKBP 2011 وقد أوضحت التجارب أن الدالة الجديدة قد حققت تقدم ملموس في دقة نتائج إعادة الترتيب وذلك مقارنة باستخدام الدالة القياسية دالة انحراف جيب الزاوية.

تم تطوير منهجين أو آليتين جديدتين لإزالة الإلتباس بأسلوب المقاربة الجماعية لربط إشارات أسماء الكيانات بقاعدة معلومات ويكيبيديا وقد تم إختبار كلا الآليتين باستخدام مجموعة البيانات القياسيةAIDA. المنهجية

الأولي تقوم بتمثيل الإعتمادات المشروطة بين أسماء الكيانات المختلفة في قاعدة المعلومات ويكيبيديا كشبكة ماركوف، بحيث تعامل أسماء الكيانات كمتغيرات مخفية بينما الإشارات المختلفة لأسماء الكيانات تعامل كملاحظات. وحيث أن عدد الحالات (أسماء الكيانات) والملاحظات (إشارات أسماء الكيانات)كبير جدا ، فإن استخدام طريقة فيتربيViterbi لإيجاد أفضل تسلسل للحالات المخفية بمعرفة الملاحظات الواردة في الوثيقة يعد مستحيل من الناحية الحسابية نظرا لكبر حجم فضاء التجربة .واستنادا علي أحد الظواهر الخاصة بهذه المشكلة نطاق البحث وهي أن لكل ملاحظة عدد معين من الحالات المخفية الممكنة واستحالة حدوث باقي الحالات. فقد قمنا بتطوير منهج جديد يعتمد علي التقديرات التقريبية لتخفيض حجم فضاء التجربة مما يساعد استخدام طريقة فيتربي في نطاق الممكن تنفيذه حسابيا. وقد أظهرت النتائج تقدما ملحوظا في دقة فك الألتباس مقارنة باستخدام المنهج الأساسي وكذلك الأساليب المتعارف عليها حاليا لحل هذا النوع من المشكلات. كما أن المنهج المقدم يوضح كيفية استخدام نموذج ماركوف المخفي (Hidden Markov Model) مع استخدام التقدير التقريبي المناسب لحل هذا النوع من المشكلات عندما يكون فضاء التجربة كبير جدا.

الأسلوب الثاني للمقاربة الجماعية لفك الألتباس يستخدم نماذج المخططات بحيث تمثل جميع الإحتمالات الممكنة كعقد داخل المخطط وجميع الأرتباطات الممكنة بين العقد المختلفة كروابط وصل بين هذه العقد. لكل عقدة درجة ثقة مبدئية مثال ذلك مدي شعبية أو رواج هذه العقدة أو الكيان. نستخدم نظام ترتيب الصفحات (Page-Rank) لإعطاء درجة لكل عقدة ويتم تجميع هذه الدرجة مع درجة الثقة المبدئية لأختيار أفضل مرشح من قائمة الأحتمالات. وقد أوضحت الأختبارات مدي فاعلية استخدام نظام ترتيب الصفحات (Page-Rank) مع درجات الثقة حيث حققت مدي دقة يتجاوز٨٧٪ متخطيا بذلك الدقة الناتجة عن استخدام المنهج الأساسي وكذلك الأساليب المتعارف عليها حاليا لحل هذا النوع من المشكلات.