Design a Computerized Lexicon for Machine Translation from Arabic to English

Dr. Suhad Malalla Khadim
Computer Sciences Department, University of Technology/ Baghdad
Email: suhad-malalla@yahoo.com
Dr. Matheel emaduldeen Abdulmunim
Computer Sciences Department, University of Technology/ Baghdad

Received on: 27/2/2013 & Accepted on: 11/6/2013

ABSTRACT

Lexicon plays an essential role in natural language processing systems and specially the machine translation systems, because it provides the system's components with the necessary information for the translation process. Although there have been a number of researches in natural language processing field, not enough attention has been given to the importance of the lexicon and specially the Arabic lexicon. The aim of this paper is to determine the basic requirements to build a lexicon used for machine translation purpose from Arabic to English language and organize lexicon contents in such form to provide efficient memory usage and minimize the retriever time, and also determine how can we retrieve these tokens from the lexicon after analyzing them by using morphology analyzer.

The proposed lexicon based on the Arabic root in storing its words, it use the logical terms to represent its knowledge and uses B^+tree data structure to organize this knowledge. And the proposed lexicon has been built by using the visual logical programming language (Visual Prolog).

Keywords: Lexicon, Machine Translation, Natural Language Processing.

تصميم معجم حاسوبي للترجمة الآلية من العربية إلى الإنكليزية

الخلاصة

يؤدي المعجم دورا مهما في أنظمة معالجة اللغات الطبيعية وخصوصا أنظمة الترجمة الآلية، فهو الذي يزود أجزاء المنظومة بالمعلومات الضرورية لعملية الترجمة، وعلى الرغم من وجود العديد من البحوث في إطار معالجة اللغات الطبيعية، لم يكن هناك اهتمام كاف في المعجم وخصوصا المعجم العربي.

يهدف البحث إلى تحديد المتطلبات الاساسية لبناء معجم يستخدم لأغراض الترجمة الآلية من اللغة العربية إلى اللغة الإنكليزية وتنظيم مواد المعجم بشكل يوفر الكفاءة في الخزن والسرعة في الاسترجاع وكذلك كيفية استرجاع المفردات من المعجم بعد تحليلها من قبل المحلل الصرفي .

https://doi.org/10.30684/etj.31.7B.15
2412-0758/University of Technology-Iraq, Baghdad, Iraq
This is an open access article under the CC BY 4.0 license http://creativecommons.org/licenses/by/4.0

يعتمد المعجم المقترح على الجذر العربي في خزن مفرداته ويستخدم الحدود المنطقية في تمثيل معارفه والهيكل الشجري +B في تنظيمها. ولقد تم بناء المعجم المقترح باستخدام اللغة البرمجية المنطقية المرئية(Visual Prolog).

الكلمات المفتاحية: المعجم، الترجمة الآلية، معالجة اللغات الطبيعية.

المقدمة

إن حقل معالجة اللغات الطبيعية يهتم بكيفية ايجاد السبل والوسائل التمثيلية والعلاجية الحاسوبية للغات الطبيعية أي بمعنى التفاعل مع الحاسوب باللغة الطبيعية واسقاط الحاجز اللغوي بين الحاسوب ومستخدميه. هذا وان الدوافع التي تكمن وراء هذا الاهتمام هو التطور السريع في مجال نظم الحاسوب وشبكات الانترنت زيادة على ذلك فان اللغة تمثل النواة الاساسية لوحدة العلوم والفكر والقاسم المشترك للعلوم الانسانية[١], واللغة هي وسيلة من وسائل الاتصال الفكري بين أبناء المجتمع وهي أداة التفاهم ونقل الأفكار والحضارة من جيل إلى آخر, لذا فإن إمكانية فهم اللغات الطبيعية من قبل الحاسوب هي حاجة ضرورية ومطلوبة من الحاسوب لعمل فعاليات مهمة تصب تطبيقاتها في مجال فهم اللغات الطبيعية والتي بدأت تتزايد مع الزمن ومن هذه التطبيقات الترجمة الآلية، فهم الكلام المنطوق، تحليل النصوص، تنقيح النصوص، إعراب النصوص وغيرها[٢]. وفي جميع هذه الانظمة يؤدي المعجم دورا اساسيا وخصوصا في مجال الترجمة الآلية.

معالجة اللغات الطبيعية

برز موضوع معالجة اللغات الطبيعية كواحد من أهم فروع الذكاء الاصطناعي، وإن أهم أهداف معالجة اللغات الطبيعية هو جعل الحاسوب يفهم اللغة الطبيعية.

إن الاتجاه العام في معالجة اللغة الطبيعية يتحدد في مراحل رئيسية وهي النظام الصوتي، الصرف، التركيب، الدلالة، المقاميات والمعجم[٣].

وفي بحثنا هذا سوف نختص في كيفية تصميم معجم لاسناد الترجمة الآلية لنص مكتوب في اللغة العربية إلى اللغة الإنكليزية وكيفية استرجاع المفردات من المعجم بعد تحليلها من قبل المحلل الصرفي .

مراحل الترجمة الآلية[٤].:

تمر الترجمة الآلية على غرار الترجمة البشرية بمراحل ثلاثة هي :التحليل ثم النقل ثم التوليد :فبعد أن يتم إدخال النص المراد ترجمته في الحاسوب، تبدأ عملية الترجمة.

-المرحلة الأولى :التحليل و يشمل الخطوات التالية:

يشرع الحاسوب في التحليل الصرفي للنص بأن يتعرف على أجزاء الكلمات و مكوناتها، عندئذ، يبحث الحاسوب في المعجم عن هذه الكلمات، ليقوم بعدها بالتحليل النحوي للنص كأن يستخرج الفعل و الفاعل...إلخ.

-المرحلة الثانية :النقل و يتم على مستويين اثنين هما:

*المستوى المعجمي :يقوم الحاسوب بنقل الكلمات من اللّغة المصدر إلى اللّغة الهدف، مستعينا في ذلك بالمعاجم الثنائية اللّغة المتاحة له.

*المستوى التركيبي : هنا يقوم الحاسوب بوضع المقابلات النحوية في اللّغة الهدف، مثلا الفاعل و المفعول به...إلخ.

-المرحلة الثالثة :التوليد أي التركيب (تكوين الجمل في اللّغة الهدف) و يتم هو الآخر على مستويين اثنين هما:

*المستوى النحوي : تطبيق قواعد النحو في اللّغة الهدف(كترتيب الكلمات) حسب مميزات ومقتضيات اللّغة المترجم إليها، كوضع الفعل في مقدمة الجملة العربية...إلخ.

*المستوى الصرفي :تطبيق القواعد النحوية و الصرفية في اللّغة الهدف، كالعدد و الجنس و الزمن...إلخ .

المعجم [٥،٦]

المعجم هو الذي يزود أجزاء المنظومة بالمعلومات الضرورية لعملية الترجمة، وإن مفردات المعجم ليست أسماء فقط نطلقها لتمييز الأشياء والمفاهيم بل رموز تتلقى مدلولاتها

[من علاقة المصاحبة التي تربطها من خلال السياق ومفاهيم أخرى يتضمنها المعجم، مثال ذلك كلمة "يصعد" يختلف مدلولها وفقا للسياق التي ترد فيه، فمثلا "يصعد إلى الطابق الثاني" ليس كما في "يصعد إلى السماء"، أي إن مجموعة مفردات المعجم ليست مجرد قائمة من الكلمات المنعزلة التي لا رابط بينها.

إن قواعد وأساليب واضحة يجب أن تحدد كيفية تمثيل المعلومات داخل المعجم، وعلى أي حال، فقد تم الإدراك أن المعجم يجب أن يحتوي على أكثر المعلومات تفصيلا كلما أمكن ذلك. ويجب أن ترمز هذه التفاصيل باعتماد طرق قياسية انموذجية وبما يسهل استعمال المعجم من قبل أي نظام للترجمة.

يختلف عدد المعاجم المطلوبة لأي نظام ترجمة آلية تبعا إلى التقنية المستخدمة في الترجمة فمثلا نظام الترجمة الآلية الذي يستخدم الصيغة الوسيطة للترجمة بين ثلاث لغات يحتاج إلى أربعة معاجم، لكل لغة من هذه اللغات الثلاثة معجم خاص بها ومعجم إضافي للغة الوسيطة. بينما في نظم الترجمة الآلية التي تستخدم تقنية النقل فإن النظام سوف يحتاج إلى ثلاثة معاجم، الأول يستعمل في تحليل لغة المصدر ويسمى معجم لغة المصدر، والثاني يستعمل لعمليات الترجمة من لغة المصدر إلى لغة الهدف ويسمى معجم لغة المصدر- معجم لغة الهدف، والآخر يستعمل لعملية توليد النص بلغة الهدف ويسمى معجم لغة الهدف، وبالطبع فقد تتداخل هذه المعاجم مع بعضها لتتلائم مع طبيعة عمل المنظومة.

المحلل الصرفي

يقصد بالتحليل الصرفي قيام النظام المحوسب بتحليل ما يغذى إليه من كلمات إلى عناصرها الصرفية الأولية، وتحديد سماتها.

لنأخذ مثالا عن التحليل الصرفي لكلمة من كلمات اللغة العربية:

إذا كانت الكلمة المدخلة للمحلل الصرفي الآلي هي" والاستيطاني"، فإن مخرجات المحلل الصرفي الآلي ستكون: عناصر الكلمة: "و"+"ال"+"استيطان"+"ياء النسب"، جذع الكلمة:"استيطان"، جذر الكلمة:"وطن"، الصيغة الصرفية: استفعال، العمليات:إبدال الواو ياء.

لنأخذ مثالا آخرا ولكن لكلمة من كلمات اللغة الإنكليزية:

لو كانت كلمة الإدخال المغذاة للمحلل الصرفي الآلي هي (unaccepted)، فإن الإخراج للمحلل الصرفي الآلي سيكون: عناصر الكلمة un+accept+ed، جذع الكلمة: accept، لواصق سابقة: un، لواصق لاحقة: ed، زمن الفعل: الماضي

يوضح المثالان بعض أوجه الاختلاف في طبيعة التحليل الصرفي الآلي بين "العربية"و"الإنكليزية"، وعلى رأس تلك الأوجه اعتماد تكوين الكلمات في اللغة العربية على الجذور والأنماط الصرفية في حين اعتماد اللغة الإنكليزية على جذوع الكلمات واللواصق.

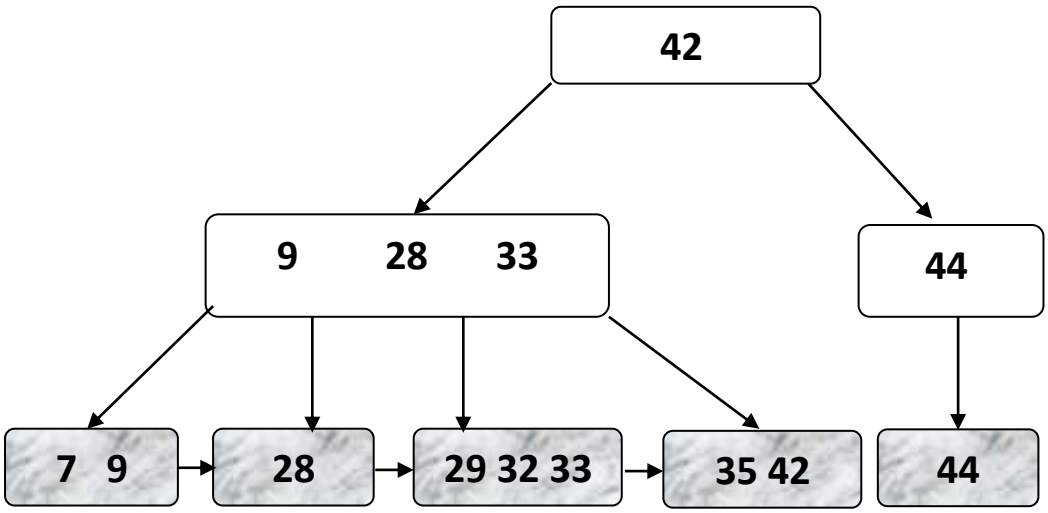
الهيكل الشجري +B [٧]

شجرة +B هي واحدة من طرق هيكلة البيانات (data structure) والتي تستخدم في التطبيقات التي تحتاج إلى ترتيب كمية كبيرة من البيانات أو للتطبيقات التي تحتاج إلى طريقة بحث كفوءة داخل بيانات كبيرة الحجم. نستطيع اعتبار شجرة +B كطريقة فهرسة

لقاعدة البيانات (index to database) لذلك تسمى في بعض الأحيان الفهارس (indices).

عند خزن قيد معين في قاعدة بيانات معينة سوف يخزن موقع القيد داخل شجرة +B وكذلك يخزن مفتاح ذلك القيد (key)، فلذلك فإن أوراق (leaves) الشجرة ستكون متكونة من مفتاح كل قيد موجود في قاعدة البيانات علاوة على موقع ذلك القيد (reference number) داخل القاعدة. وعند البحث عن قيد معين داخل قاعدة البيانات، سوف نحتاج إلى معرفة مفتاح ذلك القيد، وشجرة +B ستوفر قيمة موقع ذلك القيد داخل قاعدة البيانات، وباستخدام هذه القيمة نستطيع الوصول إلى القيد في قاعدة البيانات بصورة مباشرة.

شجرة +B متوازنة، لاحظ الشكل(١)، والمقصود بهذا أن البحث عن أي مفتاح في أوراق الشجرة (leaves) له نفس العمق، وبسبب هذه الخاصية فإن البحث عن قيد معين داخل ملايين القيود اصبح مضمونا حتى في أسوأ الحالات. وشجرة +B تشبه الشجرة الثنائية (binary tree) ما عدا أن شجرة+B تخزن اكثر من مفتاح داخل كل عقدة (node) وهذه الخاصية هي من مزايا شجرة +B حيث أعطتها قابلية الوصول السريع إلى البيانات، أن مفاتيح القيود لا تخزن في أوراق الشجرة فقط وإنما تكرر هذه المفاتيح مرة ثانية في العقد غير الورقية وذلك لتحديد المسار الصحيح للوصول إلى موقع القيد المطلوب.



شكل رقم (١) كيفية الخزن داخل الهيكل الشجري . B^+

المعجم المقترح

المعجم هو من الأجزاء الأساسية في أي منظومة ترجمة آلية، وبدون المعجم لا يمكن بناء منظومة ترجمة آلية، إن المعجم هو المسؤول عن تزويد أجزاء المنظومة بالمعلومات الضرورية لعملية الترجمة الآلية. ونظرا للخاصية الصرفية العالية للغة العربية فلقد استخدمنا "الجذر" بدلا من "الجذع" كأساس لتنظيم المفردات العربية للمعجم، في هذا المعجم المقترح سنعتمد على الحقائق المنطقية في تمثيل المعرفة المطلوبة للترجمة الالية.

معجم المنظومة المقترحة يتضمن عدد من الملفات وهي: ملف الجذور العربية، ملف المفردات، ملف السوابق، ملف اللواحق، ملف البنى العميقة للواحق، ملف الأوزان الصرفية، ملف السمات الدلالية، ملف الأطر الدلالية، ملف التصنيفات النحوية الفرعية، ملف الكلمات العربية المبنية، ملف شواذ الأفعال الإنكليزية، ملف شواذ الأسماء الإنكليزية، ملف الكلمات المركبة والتعابير وملف البنى العميقة للأوزان الصرفية.

المعجم يحتوي على معطيات مهمة لعملية الترجمة من اللغة العربية إلى اللغة الإنكليزية، حيث يتم ربط كل مفردة عربية في المعجم بمجموعة من السمات مثل العدد،

النوع، النوع الدقيق، الجنس، وغيرها، ويوجد أيضا تحديد لبعض السمات النحوية (مثلا من السمات النحوية للأفعال والظروف هي سمة متعد أو لازم)، ويتم ربط المفردات أيضا بالسمات الدلالية.

للمفردة العربية عدة احتمالات فهي أما تكون كلمة مشتقة، كلمة أدواتية، كلمة مركبة أو تكون جذر ولكل احتمال ممكن أن يكون له عدة معاني ولكل معنى احتمال يكون له أكثر من سمة دلالية أو أكثر من إطار دلالي أو أكثر من تصنيف نحوي فرعي.

تصنف مفردات المعجم إلى خمسة أنواع وهي:اسم، فعل، أداة، مركب، صفة وظرف. تصنف كل مفردة في المعجم اعتمادا على هذه الأنواع المذكورة، مع تحديد لنوعها الدقيق فمثلا النوع الدقيق للأسماء التي تتعامل معها المنظومة هي:اسم علم، اسم آلة، اسم معرف، اسم جنس، اسم المكان والزمان، اسم فاعل، اسم مفعول، صيغة مبالغة، مصدر، مصدر ميمي، جمع تكسير.

قد تم حفظ المعارف في ملفات ذات هيكل بياني شجري من نوع +B وبذلك وفرت الخاصية إمكانية توسيع المعارف بدون التقييد بسعة الذاكرة وسرعة الوصول إليها. لأن الشجرة الثنائية المتطورة(B^+ tree) هي إحدى الهياكل البيانية الجديدة التي حافظت على مزايا شجرة B^+ في الوصول المباشر مع إمكانية التناول المتسلسل السريع.

تمثيل المعرفة المعجمية

يغذى المعجم أو يزود بالمعطيات المعجمية والتي تشمل المفردات العربية ومعانيها في اللغة الإنكليزية وجميع المعرفة الضرورية لعملية الترجمة والتي تشمل المعرفة الصرفية، النحوية والدلالية. وبعض المعرفة مخزونة في المعجم (أي لا تتطلب إدخالها من قبل المستخدم) مثل المعرفة حول الكلمات العربية المبنية (مثل: أسماء الإشارة، الأدوات، الضمائر المنفصلة)، المعرفة حول السوابق واللواحق العربية والمعرفة حول الأوزان والصيغ الصرفية.

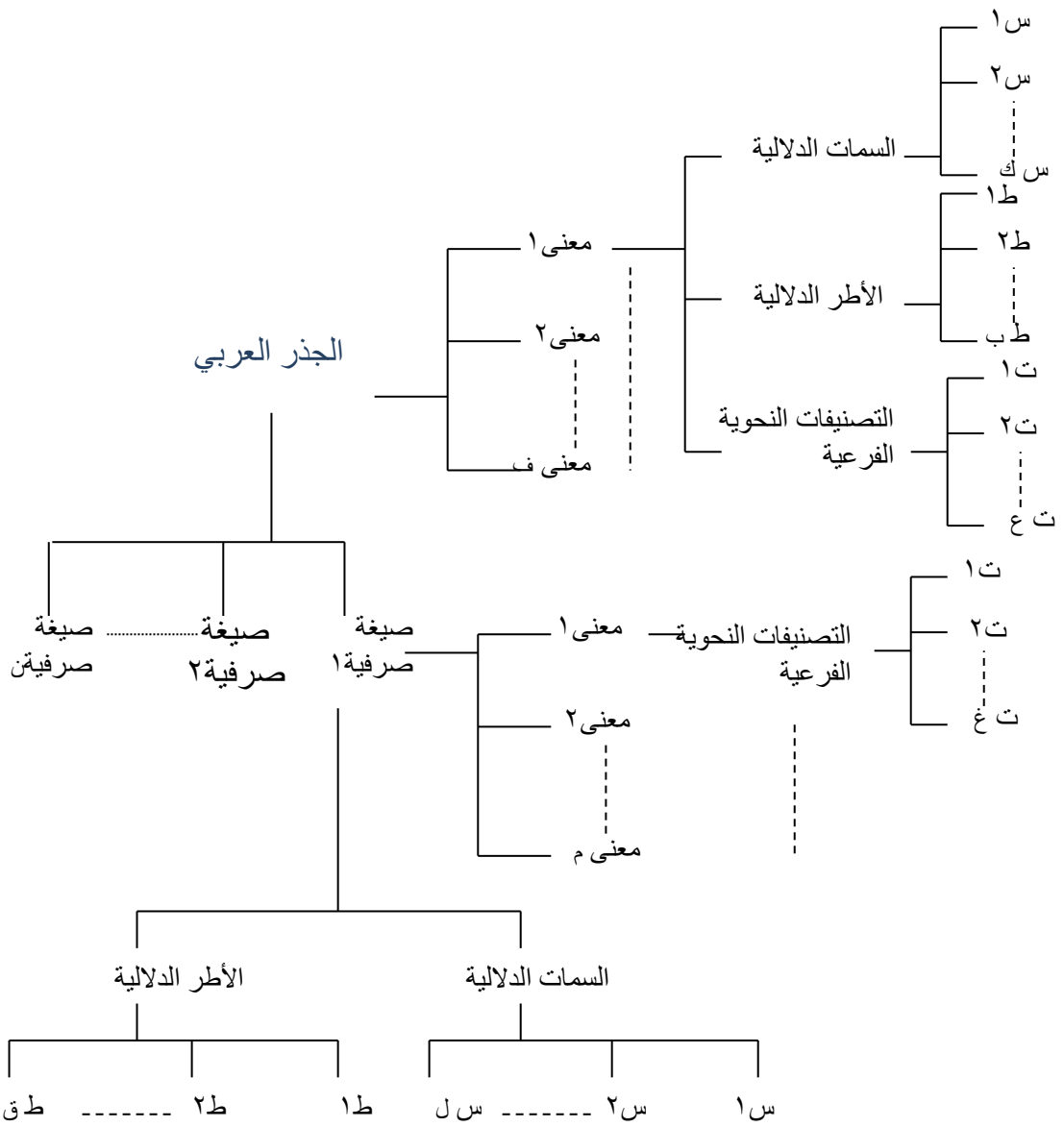
لقد تم تمثيل المعرفة على شكل حدود منطقية تحتوي على جميع المعلومات اللازمة لعملية الترجمة من اللغة العربية إلى اللغة الإنكليزية مثل:

الجذر العربي

إن المفتاح في الهيكل الشجري للمعجم الخاص بقاعدة معرفة الجذور هو الجذر العربي، والمعلومات الخاصة بالجذر العربي تتمثل بحد واحد يتكون من العناصر البيانية الآتية:

١. Reflist1: قائمة مؤشرات إلى المركبات البيانية التي تحتاجها المنظومة للوصول إلى المعنى الإنكليزي المقابل (المفردة الإنكليزية).
٢. Ref1: مؤشر إلى الحد الذي يصف الصيغ الصرفية التي تنطبق على هذه المفردة العربية، ويتكون هذا الحد من العناصر البيانية الآتية:
   ٢-١ Int: رقم صحيح يمثل رمز الصيغة الصرفية.
   ٢-٢ Ref: مؤشر إلى الحد الذي يتكون من عنصر بياني واحد وهو قائمة من السلاسل الحرفية التي تمثل الصيغ الصرفية التي ممكن أن تظهر المفردة العربية بها.
٣. Ref2: مؤشر إلى الحد الذي يتكون من عنصر بياني واحد وهو قائمة من المركبات البيانية والتي تتكون من العناصر البيانية الآتية:
   ٣-١ Int1: رقم صحيح يمثل رمز الصيغة الصرفية.
   ٣-٢ Reflist: قائمة مؤشرات إلى المركبات البيانية التي تصف الكلمات الإنكليزية التي تقابل هذه الصيغة الصرفية ، وهذه المركبات البيانية تمثل بحد واحد يتكون من العناصر البيانية الآتية:

٣-٢-١ Eng: سلسلة حرفية تمثل المفردة الإنكليزية.
٣-٢-٢ Weight: سلسلة حرفية تمثل الصيغة الصرفية للمفردة العربية.
٣-٢-٣ Vowel: قائمة من الأحرف التي تمثل تشكيل الكلمة.
٣-٢-٤ Reflist3: قائمة مؤشرات إلى التصانيف النحوية الفرعية للمفردة.



الشكل (٢) تشابك العلاقات الواصفة للجذر العربي.

المفردة الإنكليزية

تمثل بحد واحد يتكون من العناصر البيانية الآتية:

Entry: سلسلة حرفية (string) تمثل الجذع الإنكليزي أو المعنى الإنكليزي المقابل للمفردة العربية.

Class: رقم صحيح (integer) يمثل قسم الكلام الذي تنتمي إليه المفردة العربية.

spe_class: رقم صحيح يمثل النوع الدقيق لقسم الكلام.

Check: رقم صحيح لتحديد قابلية المفردة للعد.

Gender: رقم صحيح يحدد نوع الجنس.

Number: رقم صحيح لتحديد العدد.

Ref: مؤشر إلى السمات الدلالية.

Reflist1: قائمة مؤشرات الأطر الدلالية.

Reflist2: قائمة مؤشرات التصنيفات النحوية الفرعية.

Charlist: قائمة أحرف تمثل تشكيل الكلمة.

كيفية البحث عن الكلمات داخل المعجم المقترح

يقوم المحلل الصرفي باستخلاص الجذور من الكلمات العربية حيث تمر المفردات الواحدة تلو الأخرى بمرحلة التحليل إلى أولياتها الصرفية، وتتمثل بفصل السوابق واللواحق (إن وجدت) عن جذع الكلمة ومن ثم استخلاص الجذر والوزن المطابق من الجذع (في حالة كون الجذع المستخلص هو ليس جذر الكلمة) والبحث في المعجم عن جذر الكلمة لاسترجاع المعرفة المرتبطة بالكلمة مثل المعرفة النحوية والدلالية للكلمة. صمم المحلل الصرفي بحيث يتلائم مع متطلبات الترجمة الآلية للمنظومة المقترحة، إذ يقوم بالتحليل باستخدام طرق مختلفة وذلك للحصول على جميع الاحتمالات الممكنة لتحليل الكلمات، ففي البداية يبحث في معجم الكلمات الأدواتية، وبعدها ينتقل إلى مرحلة البحث في معجم الجذور ومحاولة تحليل الكلمة إلى سوابق ولواحق وإذا تم أيجاد الكلمة سيعود مرة أخرى للبحث محاولا تحليل الكلمة إلى جذر وسوابق ولواحق وباسلوب مختلف، وإذا لم يجد تحليلا آخرا للكلمة ينتقل إلى محاولة تحليل الكلمة إلى سوابق ولواحق وجذع واستخلاص الجذر والصيغة صرفية من الجذع وإذا وجد تحليلا للكلمة يعود مرة ثانية لاكتشاف تحليل آخر للكلمة المشتقة، مع الأخذ بنظر الاعتبار حالات الإعلال والإبدال وحالات الشواذ التي ممكن أن تظهر، لاحظ الخوارزمية الآتية:

خوارزمية التحليل الصرفي

المدخلات: إدخال قائمة كلمات الجملة مع تشكيلها.

المخرجات: قائمة (mylist) تتضمن المعلومات الصرفية لكل كلمة في الجملة المدخلة.

١. استرجع دالة معالجة الكلمات الأدواتية صرفيا (البحث في الكلمات الأدواتية عن الكلمة الحالية).

٢. هل الكلمة أدواتية؟ إذا كان الجواب:

٢-١ كلا: انتقل إلى الخطوة (٣).

٢-٢ نعم: اخزن المعلومات المطلوبة والناتجة من عملية تحليل الكلمة في قائمة mylist مع رقم الكلمة الحالية.

٣. استرجع دالة معالجة الجذور صرفيا (البحث في ملف الجذور أي إن الكلمة متكونة من لواصق وجذر).

٤. هل وجدت تحليلا مختلفا للكلمة الحالية؟ إذا كان الجواب:

٤-١ كلا: انتقل إلى الخطوة (٥).

٤-٢ نعم: اخزن المعلومات المطلوبة والناتجة عن تحليل الكلمة في قائمة mylist مع رقم الكلمة الحالية ثم انتقل إلى الخطوة (٣).

٥. استرجع دالة معالجة الكلمات المشتقة صرفيا (الكلمة متكونة من جذر وصيغة صرفية ولواصق).
٦. هل وجدت تحليل مختلف للكلمة المشتقة؟ إذا كان الجواب:
٦-١ كلا: انتقل إلى الخطوة (٧).
٦-٢ نعم: اخزن المعلومات المطلوبة والناتجة عن تحليل الكلمة في قائمة mylist مع رقم الكلمة الحالية ثم انتقل إلى الخطوة (٥).
٧. استرجع دالة معالجة الكلمات التي تتضمن لواحق ذات بنية عميقة.
٨. هل وجدت تحليلا صرفيا للكلمة الحالية؟ إذا كان الجواب:
٨-١ نعم: انتقل إلى الخطوة (١١).
٨-٢ كلا: إرجاع "الكلمة ليست معجمية" ثم انتقل إلى الخطوة (١١).
٩. اخزن المعلومات المطلوبة والناتجة من تحليل الكلمة في قائمة mylist مع رقم الكلمة الحالية.
١٠. هل توجد كلمات أخرى في القائمة المدخلة؟ إذا كان الجواب:
١٠-١ نعم: انتقل إلى الخطوة (١).
١٠-٢ كلا: إرجاع "قائمة mylist".
١١. النهاية.

معالجة الكلمات الادواتية صرفيا

هذا الجزء من المحلل الصرفي مسؤول عن البحث في معجم الكلمات الأدواتية مثل أسماء الإشارة، الضمائر، الأدوات وغيرها، ويقوم بالبحث عن الكلمة في الهيكل الشجري الخاص بكل ملف من ملفات الكلمات الأدواتية التي يتضمنها المعجم، فإذا لم يجدها في أي من الملفات ينتقل إلى معالجة سوابق الكلمة وذلك لأن بعض الكلمات الأدواتية قد تتضمن سوابق (مثل: فهم، فهذه، وهو،ولسوف...)، ثم يعود للبحث مرة أخرى عن الكلمة بعد فصلها عن السوابق، فإذا تم إيجادها يتم التدقيق حول مطابقة هذه السابقة مع الكلمة وذلك من خلال المعلومات المخزونة في الحد الخاص بالكلمة الأدواتية. وتوضح الخوارزمية التالية كيفية تحليل الكلمة الأدواتية مع الأخذ بنظر الاعتبار التشكيل إن وجد.

خوارزمية معالجة الكلمات الأدواتية صرفيا

المدخلات: الكلمة العربية المراد تحليلها مع تشكيلها (قائمة أحرف تمثل التشكيل).

١. هل الكلمة موجودة في الهيكل الشجري لملف الضمائر الشخصية؟ إذا كان الجواب:
١-١نعم: انتقل إلى الخطوة (٢).
١-٢كلا: انتقل إلى الخطوة (٦).
٢. استرجع بيانات الكلمة المخزونة في المعجم.
٣. تدقيق التشكيل إن وجد.
٤. هل توجد سوابق؟ إذا كان الجواب:
٤-١ نعم: انتقل إلى الخطوة (٥).
٤-٢ كلا: انتقل إلى الخطوة (٥-٢).
٥. هل هذه السوابق ممكن أن ترتبط مع الكلمات الأدواتية؟ إذا كان الجواب:
٥-١ كلا: انتقل إلى الخطوة (٨-٢).
٥-٢ نعم: إرجاع: بيانات الكلمة الأدواتية مع بيانات السوابق إن وجدت ثم انتقل إلى الخطوة (٩).
٦. هل الكلمة موجودة في الهيكل الشجري لملف أسماء الإشارة؟ إذا كان الجواب:
٦-١ كلا: انتقل إلى الخطوة (٧).
٦-٢ نعم: انتقل إلى الخطوة (٢).

٧. الكلمة موجودة في الهيكل الشجري لملف الأدوات؟ إذا كان الجواب:
٧-١ كلا: انتقل إلى الخطوة (٨).
٧-٢ نعم: انتقل إلى الخطوة (٢).
٨. هل تحتوي الكلمة على سوابق؟ إذا كان الجواب:
٨-١ نعم: معالجة سوابق الكلمة ثم انتقل إلى الخطوة (١).
٨-٢ كلا: إرجاع: الكلمة ليست من الكلمات الأدواتية.
٩. النهاية.

أمثلة:

- **الكلمة**: ولسوف **السابقة**: ولـ **الأداة**: سوف
- **الكلمة**: فَهُمْ **السابقة**: ف **الأداة**: هم
- **الكلمة**: فهم **السابقة**: ف **الأداة**: هم

معالجة الجذور صرفيا

هذا الجزء من المحلل الصرفي مسؤول عن إيجاد الكلمات إذا كانت متكونة من جذر ولواصق، ففي البداية يبحث عن الكلمة في الهيكل الشجري الخاص بالجذور فإذا وجدها يسترجع بيانات الحدود التي تنطبق مع تشكيلها. وإذا لم يجد الكلمة في الهيكل الشجري يقوم بإقصاء أطول لاحقة وفصلها عن الكلمة والبحث عنها في المعجم، وإذا لم يتم إيجادها يقوم بإقصاء أطول سابقة وفصلها عن الكلمة المتبقية والبحث عنها في المعجم، وإذا لم يجدها يستمر بإقصاء لاحقة أقصر والبحث في المعجم. فإذا لم يتم إيجاد الكلمة في المعجم يرجع إلى الكلمة الأصلية ويحاول تحليلها بنفس الطريقة ولكن هذه المرة يقوم بمعالجة السوابق ومن ثم اللواحق. وإذا لم يتم إيجادها أيضا فهذا يعني أن الكلمة غير موجودة في معجم الجذور. أما إذا وجدها فيقوم بتدقيق نوع اللواصق مع نوع الكلمة من خلال بيانات حدود اللواصق والكلمة ويقوم أيضا بتدقيق التشكيل. وتوضح الخوارزمية التالية كيفية تحليل الكلمة التي تتضمن جذر ولواصق.

خوارزمية معالجة الجذور صرفيا

المدخلات: الكلمة العربية مع تشكيلها.
المخرجات: المعلومات الناتجة من تحليل الكلمة والمطلوبة في عملية الترجمة.

١. هل الكلمة موجودة في الهيكل الشجري للجذور؟ إذا كان الجواب:
١-١ نعم: انتقل إلى الخطوة (٢).
١-٢ كلا: انتقل إلى الخطوة (٣).
٢. هل توجد لواصق في الكلمة؟ إذا كان الجواب:
٢-١ نعم: اختر المعاني التي تنطبق مع أنواع الكلمات التي ترتبط مع هذه اللواصق.
٢-٢ كلا: اختر المعاني التي تنطبق مع تشكيل الكلمة ثم انتقل إلى الخطوة (٧).
٣. هل أول حرفين: "الـ" ولم يتم معالجتهما مسبقا؟ إذا كان الجواب:
٣-١ كلا: انتقل إلى الخطوة (٤).
٣-٢ نعم: افصل أول حرفين عن الكلمة ثم انتقل إلى الخطوة (١).
٤. هل تمت معالجة لواحق الكلمة؟ إذا كان الجواب:
٤-١ نعم: انتقل إلى الخطوة (٥).
٤-٢ كلا: إقصاء أطول لاحقة وفصلها عن الكلمة ثم انتقل إلى الخطوة (١).
٥. هل تمت معالجة سوابق الكلمة؟ إذا كان الجواب:
٥-١ نعم: انتقل إلى الخطوة (٦).
٥-٢ كلا: إقصاء أطول سابقة وفصلها عن الكلمة ثم انتقل إلى الخطوة (١).

٦. هل طول اللاحقة المستخلصة أكبر من واحد؟ إذا كان الجواب:
٦-١ نعم: إقصاء لاحقة أقصر وفصلها عن الكلمة ثم انتقل إلى الخطوة (١).
٦-٢ كلا: إرجاع "الكلمة غير معجمية".
٧. النهاية.

أمثلة

الكلمة: تكتبون **السابقة**: ت **اللاحقة**: ون **الجذع**: كتب **الوزن المطابق**: فعل **الجذر**: كتب

معالجة الكلمات المشتقة صرفيا

الإدخال إلى هذا الجزء من المحلل الصرفي هو الكلمة المراد تحليلها مع قائمة أحرف التي تمثل تشكيل الكلمة، حيث يقوم بتحليل الكلمة إلى جذع ولواصق واستخلاص الجذر والصيغة الصرفية من الجذع الناتج، وتوضح الخوارزمية التالية كيفية تحليل الكلمة المشتقة:

خوارزمية معالجة الكلمات المشتقة صرفيا

المدخلات: الكلمة العربية مع تشكيلها Word.
المخرجات: المعلومات الناتجة من تحليل الكلمة والمطلوبة في عملية الترجمة.

١. إيجاد الميزان الصرفي والجذر للكلمة.
٢. هل الجذر موجود في الهيكل الشجري ؟ إذا كان الجواب:
١-١ نعم: انتقل إلى الخطوة (٣).
١-٢ كلا: انتقل إلى الخطوة (٨).
٣. هل الميزان الصرفي موجود ضمن الأوزان الصرفية التابعة للجذر؟ إذا كان الجواب:
٣-١ نعم: انتقل إلى الخطوة (٤).
٣-٢ كلا: انتقل إلى الخطوة (٨).
٤. هل توجد لواصق؟ إذا كان الجواب:
٤-١ نعم: انتقل إلى الخطوة (٥).
٤-٢ كلا: انتقل إلى الخطوة (٦).
٥. اختر المعاني التي تنطبق مع أنواع الكلمات التي ترتبط مع هذه اللواصق.
٦. اختر المعاني الإنكليزية التي تنطبق مع تشكيل الكلمة.
٧. انتقل إلى الخطوة (١٣).
٨. هل أول حرفين بالكلمة هما "ال" ولم يتم معالجتهما؟ إذا كان الجواب:
٨-١ كلا: انتقل إلى الخطوة (٩).
٨-٢ نعم: فصل "ال" التعريف عن الكلمة ثم انتقل إلى الخطوة (١).
٩. هل الميزان الصرفي موجود في الهيكل الشجري الخاص بالبنى العميقة للأوزان الصرفية؟ إذا كان الجواب:
٩-١ كلا: انتقل إلى الخطوة (١٠).
٩-٢ نعم: استرجع الصيغة الصرفية لهذا الوزن واستبدل الكلمة المراد تحليلها بكلمة جديدة حسب هذه الصيغة الصرفية ثم انتقل إلى الخطوة (١).
١٠. هل تمت معالجة لواحق الكلمة؟ إذا كان الجواب:
١٠-١ نعم: انتقل إلى الخطوة (١١).
١٠-٢ كلا: إقصاء أطول لاحقة وفصلها عن الكلمة ثم انتقل إلى الخطوة رقم(١).
١١. هل تمت معالجة سوابق الكلمة؟ إذا كان الجواب:

١١-١ نعم: انتقل إلى الخطوة (١٢).
١١-٢ كلا: إقصاء أطول سابقة وفصلها عن الكلمة ثم انتقل إلى الخطوة رقم(١).
١٢. هل طول اللاحقة أكبر من واحد؟ إذا كان الجواب:
١٢-١ نعم: إقصاء لاحقة أقصر وفصلها عن الكلمة ثمانتقل إلى الخطوة رقم(١).
١٢-٢ كلا: إرجاع "الكلمة غير موجودة".
١٣. النهاية.

أمثلة:
- الكلمة: **الأولاد** السابقة: ال اللاحقة: - الجذع: أولاد الوزن المطابق: أفعال الجذر: ولد
- الكلمة: **وباستخراجهم** السابقة: وب اللاحقة: هم الجذع: استخراج الوزن المطابق: استفعال

الجذر: خرج
وفي حالة عدم أيجاد تحليل للكلمة ينتقل إلى المطابقة مع الأوزان حيث نتعامل في هذا الجزء مع الهياكل الآتية:
- هيكل١ الصورة السطحية للمفردة أو ما يسمى الميزان الصرفي.
- هيكل٢ الصورة الوسطية للمفردة.
- هيكل٣ الصورة العميقة للمفردة أو ما يسمى بالصيغة الصرفية.

حيث تقوم بتبديل المفردة التي هي بالهيكل١ إلى الهيكل٢ ومن ثم البحث عن الهيكل٣ في الصيغ الصرفية الخاصة بالجذر، لو نأخذ المثال الآتي:
لتحليل كلمة "بائع" سنحصل على الميزان الصرفي "فائل" وعند البحث عن هذا الميزان الصرفي في الهيكل الشجري الخاص بالأوزان الصرفية واسترجاع بيانات الحد الخاص بها سنستنتج أن أصل الميزان الصرفي هو أما "فاول" أو "فايل" (وهي احتمالات الهيكل٢) وأن الصيغة العميقة لهذا الميزان هي "فاعل"، وبعد تبديل الكلمة الأصلية بما يقابلها في الصيغة الوسطية سنحصل على كلمة "بايع" وعند استرجاع جذرها وإيجاد أن "فاعل" هي من ضمن صيغها الصرفية، نحصل على تحليل للكلمة.

أمثلة:
- الكلمة: **بائع** السابقة:- اللاحقة:- الميزان الصرفي: فائل الصيغة الصرفية: فاعل الصيغة الوسطية: فايل أو فاول الكلمة بعد مطابقتها مع الصيغة الوسطية: بايع الجذر: بيع
- الكلمة: **قال** السابقة: - اللاحقة: - الجذع: قال الميزان الصرفي: فال الصيغة الصرفية: فعل الصيغة الوسطية: فول الكلمة بعد مطابقتها مع الصيغة الوسطية: قول

الجذر: قول

معالجة الكلمات التي تحوي لواحق ذات بنى عميقة

في حالة عدم إيجاد تحليل للكلمة، نتوجه إلى هذا الجزء من المحلل الصرفي، لكي يقوم بإقصاء أطول لاحقة والبحث عنها في ملف البنى العميقة للواحق ، ومن ثم تبديل الكلمة بالبنية العميقة والرجوع للبحث عن الكلمة الناتجة في المعجم، لاحظ المثال الآتي:
لتحليل كلمة فتيان، يقوم بفصل أطول لاحقة عن الكلمة وهي"يان" ومن ثم يبحث عن هذه اللاحقة في الهيكل الشجري الخاص بالبنى العميقة للواحق واسترجاع الحد الخاص بها حيث احتمالات البنى العميقة لهذه اللاحقة هي القائمة ["ان"،"يان"] وعند ربط كل احتمال من هذه الاحتمالات مع جذع الكلمة والبحث عنه في المعجم ككلمة جديدة يستنتج الجذر وهو "فتى". وتوضح الخوارزمية الآتية كيفية تحليل الكلمة التي تحتوي على لواحق تتغير عند اتصالها بالكلمات.

خوارزمية معالجة الكلمات التي تحوي لواحق ذات بنية عميقة

المدخلات: الكلمة المراد تحليلها (Word).
المخرجات: المعلومات الناتجة من تحليل الكلمة والمطلوبة في عملية الترجمة.

١. افتح ملف البنى العميقة للواحق.
٢. *إقصاء أكبر لاحقة وفصلها عن الكلمة Word ولتكن الكلمة الناتجة Word1.*
٣. هل اللاحقة موجودة في الهيكل الشجري لملف البنى العميقة للواحق؟ إذا كان الجواب:
   ٣-١ نعم: انتقل إلى الخطوة (٤).
   ٣-٢ كلا: انتقل إلى الخطوة (١٠).
٤. *استرجع قائمة البنى العميقة للواحق Suffix_list.*
٥. هل القائمة Suffix_list فارغة؟ إذا كان الجواب:
   ٥-١ كلا: انتقل إلى الخطوة (٦).
   ٥-٢ نعم: إرجاع "الكلمة غير معجمية" ثم انتقل إلى الخطوة (١٣).
٦. اربط اللاحقة التالية Suffix مع الكلمة Word1 ولتكن الكلمة الناتجة Word2.
٧. احذف اللاحقة Suffix من القائمة Suffix_list.
٨. استدعي دالة المحلل الصرفي للبحث عن الكلمة Word2.
٩. هل وجدت تحليل للكلمة؟ إذا كان الجواب:
   ٩-١ كلا: انتقل إلى الخطوة (١٠).
   ٩-٢ نعم: انتقل إلى الخطوة (١٣).
١٠. إقصاء لاحقة أقصر، وفصلها عن الكلمة Word ولتكن الكلمة الناتجة Word1.
١١. انتقل إلى الخطوة (٣).
١٢. إرجاع "المعلومات الناتجة من تحليل الكلمة والمطلوبة في عملية الترجمة".
١٣. اغلق ملف البنى العميقة للواحق.
١٤. النهاية.

الاستنتاجات

١. إن حفظ المعرفة المعجمية في ملفات ذات هيكل بياني شجري من نوع +B يوفر إمكانية توسيع المعارف بدون التقييد بسعة الذاكرة وسرعة الوصول إليها.
٢. إن استخدام الحدود المنطقية لتمثيل المعرفة المعجمية ملائم لبناء منظومة ترجمة.
٣. ضرورة اعتماد المعجم على الجذور لتقليل المساحة الخزنية المطلوبة للمعرفة اللغوية الخاصة بالمفردات العربية.
٤. إن استخدام رموز رقمية لكثير من الحقول البيانية بدلاً من المعطيات الأصلية المطولة ضروري للتقليل من حجم المعجم.

المصادر

[1].Kadhem, Suhad Malallah, "Design a machine translation system from Arabic to English", (2003), Ph.D. thesis, Technology University.
[2].Zainab Abrahim Athman., Ali F.M., Salma Abdul Albaki M., " Machine translation from Arabic to English for some kinds of verbs Search breaching the doctoral thesis ", Thi-Qar Journal of Scienec (A) Vol.2(1),86-99, 2010.
[3].Zainab Abrahim Alfahad, Jenan Abd Al_Kareem Abd Al-Hasan, "Machine Translation of Arabic Verb Sentences into English", Basrah Journal of Scienec (A) Vol.26(1),25-32, 2008.
[4].Hutchins J., 2006, " The State of Machine Translation in Europe " , university of east Anglia.
[5].Attia, M. A., 2008. Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. PhD thesis, University of Manchester.

[6].Salem, Y. and Nolan, B., 2009b. UNIARAB: An universal machine translator system for Arabic Based on Role and Reference Grammar. In Proceedings of the 31st Annual Meeting of the Linguistics Association of Germany (DGfS 2009).

[7].Goetz Graefe ,"B-tree indexes, interpolation search, and skew", Chicago Illinois, USA, 2006.

[8].Yasser Salem, Arnold Hensman and Brian Nolan, "Towards Arabic to English Machine Translation", ITB Journal, 2008.