

حل اللبس الدلالي لظاهرة الاشتراك اللفظي في اللغة العربية باستخدام تصنيف Naive Bayesian

يسرى مالك ضمد

قسم علوم الحاسبات / كلية العلوم / جامعة البصرة

Yus_malik@yahoo.com

الخلاصة :

يعد الغموض الدلالي (sense ambiguity) من أكبر التحديات التي تواجه معالجة اللغة الطبيعية وتطبيقاتها آليا عموما واللغة العربية خصوصا وظاهرة الاشتراك اللفظي إحدى هذه المشاكل لما تسببه من غموض في اختيار المعنى المناسب للكلمة في الجملة. التوجه الحالي في معالجة اللغات الطبيعية هو إحصائي لما لهذا المجال من فعالية في حل بعض مشاكل التطبيقات.

قدم هذا البحث حل اللبس الدلالي (word sense disambiguation) لتعدد المعنى للكلمة في اللغة العربية بإحدى الطرق الإحصائية المستخدمة في هذا المجال وهي طريقة تصنيف (naïve Bayesian) بالإضافة إلى تصميم مدونة تحتوي على جميع السياقات المختلفة التي تظهر فيها الكلمات الغامضة. حققت الطريقة نسبة نجاح عالية في تحديد واختيار المعنى الصحيح للكلمة الغامضة في الجملة بلغت 90.98% مع مدونة الجمل و 95.99% مع مدونة السياقات.

الكلمات المفتاحية: معالجة اللغات الطبيعية ، الاشتراك اللفظي، اللبس الدلالي ، تصنيف naïve Bayesian.

1- المقدمة:

كل اللغات الطبيعية تحتوي على الألفاظ المشتركة المعنى (polysemy) أي الكلمات التي تتعدد معانيها ويتحدد لفظها حسب موقعها في الجملة أو السياق، ويشكل عامل مهم من عوامل الغموض في النصوص [د.احمد، 2000].

وتعد مشكلة الغموض أو اللبس في المعنى وتسمى أيضا باسم (اللبس الدلالي) من أكبر التحديات التي تواجه معالجة اللغة العربية وتطبيقاتها آليا وخاصة الترجمة الآلية [eneko، 2007]. وهناك غموض آخر يسمى (اللبس النحوي) ويحصل بسبب بناء أو تركيب الجملة، حيث يكون للجملة أكثر من معنى أو يمكن تفسيرها بأكثر من طريقة مثل (كاتبة القصص الفلسطينية) حيث يمكن أن تعود صفة الفلسطينية على كاتبة أو القصص. فسيعاني من اللبس في الوحدات اللغوية مزدوجة التفسير أو متعددة المعاني، لذلك ينبغي تحويل القيود اللغوية - النحوية والدلالية - إلى حلول تساعد على اختيار المعنى المقصود. واللبس من أهم أسباب تأخر الحاسبات في فهم اللغات المكتوبة والمنطوقة.

وهناك أيضا (اللبس المعجمي) عندما تنتمي الكلمة لأكثر من فئة معجمية (فتكون اسما وفاعلا وظرفا وصفة) وفي هذه الحالة يسبب لبسا تركيبيا ودلاليا ومثال ذلك كلمة (ذهب) التي تعني (رحل) عندما تكون فعلا وتعني (معدن اصفر لماع) عندما تكون اسما.

مشكلة اللبس في اللغة مشكلة ليست سهلة أو هينة، وتمثل عصب التحدي الذي يواجه المعالجة الآلية للغة الطبيعية، ولحل مشكلة اللبس ينبغي حصر جميع الوحدات

اللغوية (الأفعال والأسماء وحروف الجر والمصادر) اللبسية بجميع صور اللبس ووضع التفسير المناسب لكل سياق، ثم عمل القاعدة التي تحدد القيود المعجمية والنحوية والتركيبية والدلالية اللازمة لفك اللبس. فينبغي بحث كيفية فك حالات اللبس سواء أدركها الإنسان أم لا من أجل تمثيل الحلول حاسوبيا ليتمكن الحاسب من فك اللبس ومن ثم تحليل النصوص تحليلًا دقيقًا. إن التوجهات الحالية في معالجة بعض مشاكل معالجة اللغة الطبيعية هي تقنيات تعليم الآلة (machine learning) والطرق الاحصائية ودمجها مع الطرق التجريبية للحصول على نتائج أفضل وهذا ما اثبتته البحوث في الوقت الراهن [nitin، 2010].

بحثنا الحالي يصب في هذا المجال لحل اللبس الدلالي للغة العربية التي يقل أو يندر في بعض المجالات الخوض في غمارها مقارنة مع بقية اللغات الطبيعية.

2- الاشتراك اللفظي polysemy:

تعد ظاهرة الاشتراك اللفظي من بين أهم الظواهر التي تمتاز بها اللغة العربية كما في باقي اللغات الطبيعية. وتعرف على أنها كلمة واحدة تعبر عن مفاهيم متعددة. والكلمات المتعددة المعاني تكون أكثر الكلمات رواجًا واسهلها للانتاج والفهم. إن الآثار السلبية لهذه الظاهرة هو ما تسببه من تشويش يعوق التفاهم ويلقي ظلالة من الغموض على المعنى. واختيار معنى ملائم من بين قائمة من المعاني التي يعبر عنها المشترك اللفظي يتطلب مجهودا معرفيا ويتسبب أحيانا في أخطاء ويرفع اللبس عن طريق السياق اللغوي المباشر وكل مصادر

المعلومات المتوفرة لرفع اللبس. ان لكل كلمة معنى اساسي وتعدد المعاني يرجع الى المعلومات السياقية التي ليست جزءا من معنى المفردة [د.عبد القادر، 1982].

ان اختيار المعنى الصحيح للكلمة ذات المعاني المتعددة ضمن السياق بالنسبة للانسان يكون سهلا بينما عند الحاسوب ليس كذلك ، في بعض الحالات يختلف معنى جدول(1) مشتركات لفظية (أسماء)

الكلمة	عدد المعاني	المعنى-1	المعنى-2	المعنى-3	المعنى-4	المعنى-5
يد	5	جزء من جسم الانسان	قدرة وسطان	الجود	جناح	مقبض
خال	2	اخو الام	شامة			
عين	5	عضو البصر	نبع	جاسوس	نقد	الشيء نفسه

جدول(2) مشتركات لفظية (أفعال)

الكلمة	عدد المعاني	المعنى-1	المعنى-2	المعنى-3	المعنى-4	المعنى-5
عمل	5	فعل	طبق	سعى	صنع	اشتغل
ذهب	5	مضى	زال	ترك	توجه	اختلط
فتح	5	هيا	خلاف اغلق	احتل	نشر	هدى

ملاحظة: قد يكون للكلمات في الجدولين(1)،(2) معاني اكثر حسب ماهو موجود في المعاجم اللغوية ولكن حصرت هنا كمثال.

اللغات الطبيعية من خلال استخدام الخوارزميات والتقنيات التي جاءت من تعليم الالة (machine learning) في مختلف مهام معالجة اللغات اذ حققت نجاح كبير [Tom M. Mitchell, 2010].

وبما ان معالجة اللبس هي بتعريف المعنى الصحيح للكلمة في سياقها الخاص استخدمت انظمة التعليم الموجه

3-حل لبس الكلمة اليا word sense disambiguation automatically

:-

خلال العقود الاخيرة ازدادت اهمية الدمج بين الاحصاء والتجريب في معالجة

(supervised wsd) والتي تعتمد السياق في حلها [eneko، 2007].

يعتبر حل اللبس من المهام الاساسية في nlp لما لها من تأثير في زيادة معدل نجاح تطبيقاتها مثل (الترجمة الالية، فهم اللغات الطبيعية، استخلاص المعلومات وغيرها) [nyein، 2011].

4- طريقة التصنيف الإحصائية naïve Bayesian :

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \dots\dots (1)$$

العالية في التصنيف الصحيح مقارنة بغيرها من الطرق .

حيث يفترض تأثير قيمة خاصية معينة على اعطاء صنف مستقل عن قيم الخواص الأخرى ويدعى صنف الاستقلال الشرطي (class conditional independence) Nitin] indurkhyia & fred j. damerau , 2010]..

وتصنيف naïve Bayesian كما في معادلة (2) خوارزمية تصنيف خطية اعتمادا على نظرية Bayes وتستعمل فرضية الاستقلال وقد استخدمت بقوة في عمليات التصنيف في مجال تنقيب البيانات Data Maining واثبتت جدارتها وتستخدم ايضا في تعليم الآلة الموجه (machin learning supervised) لفعاليتها ودقتها

$$P(y|x) = \frac{p(y)p(x|\theta^{(y)})}{\sum_{l=1}^k p(y=l)p(x|\theta^{(l)})} \dots\dots(2)$$

$$\theta = [\theta^{(l)}]_{l=1,\dots,k}$$

الاشتراك اللفظي في معالجة اللغات الطبيعية، من خلال حساب احتمالية كل معاني الكلمة ثم يمكنه اختيار المعنى الصحيح للكلمة من مجموعة المعاني حسب السياق الذي ترد فيه الكلمة المعادلة(3). [Eneko,2007]

حيث θ معلمة النموذج .وتستخدم معلمات مختلفة $\theta^{(l)}$ لكل صنف l .

5- حل الغموض بأستخدام تصنيف naïve Bayesian :

يستخدم تصنيف naïve Bayesian في حل اللبس والغموض الذي تسببه ظاهرة

$$\begin{aligned} \arg_s \max p(s|x_1, \dots, x_m) &= \arg_s \max \frac{p(x_1, \dots, x_m|s)p(s)}{p(x_1, \dots, x_m)} \\ &= \arg_s \max p(s) \prod_{i=1}^m p(x_i|s) \quad \dots(3) \end{aligned}$$

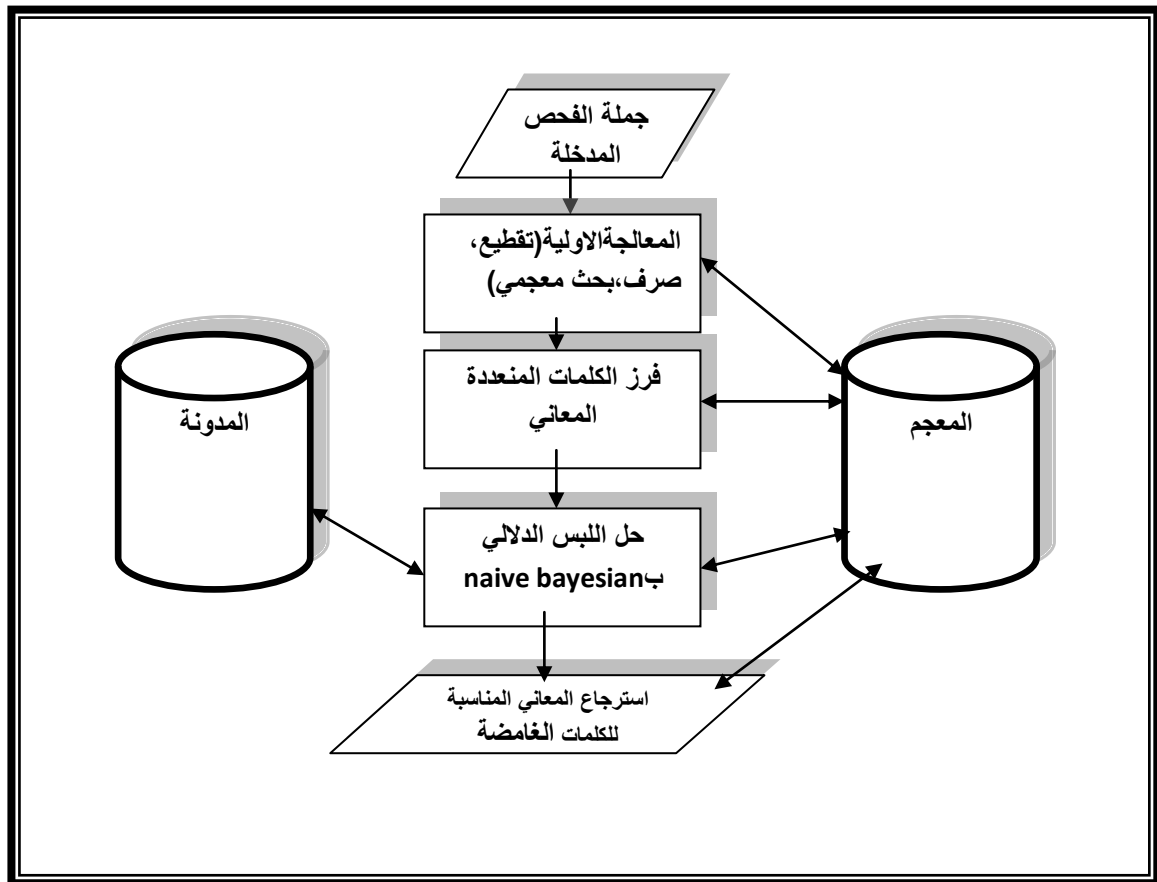
تناول بحثنا الحالي استخدام تصنيف naïve Bayesian لحل غموض الكلمات المتعددة المعاني في الجمل المكتوبة باللغة العربية .

حيث x هي متجة الصفات للكلمة و s هي معنى الكلمة الذي تم اختياره بأعلى احتمالية.

6- طريقة العمل:

تضمنت الطريقة مجموعة معالجات رئيسية يوضحها المخطط العام لطريقة العمل الشكل (1) هي:

نتيجة لكل ما تقدم ولكون ظاهرة الاشتراك اللفظي هي احدى التحديات التي تواجه تطبيقات المعالجة الآلية للغات الطبيعية من فهم وترجمة في عدم الدقة باختيار المعنى المناسب للكلمة المتعددة المعاني في الجملة وخصوصا في معالجة اللغة العربية ليا اذ تعتبر من اللغات العالمية ولكن العاملين في مجالها قليلون مقارنة مع باقي اللغات .



الشكل (1) المخطط العام لطريقة العمل

1. الادخال .. وفيها يتم ادخال الجملة العربية التي تشتمل على كلمة متعددة المعاني.
2. المعالجة الاولى.. وتشتمل عدة مراحل هي:
 - تقطيع الجملة الى الكلمات المكونة لها.
 - التحليل الصرفي للكلمات التي تحتوي على سوابق او لواحق للحصول على جذع الكلمة .
 - البحث المعجمي للحصول على صنف الكلمة وحالتها (ذات معنى واحد او متعددة المعاني).
3. فرز الكلمات المتعددة المعاني ليتسنى ادخالها لمعالجة حل اللبس كما في الخطوة اللاحقة.
4. حل اللبس وازالة الغموض باستخدام تصنيف naïve Bayesian كما في الخوارزمية (1).. في البداية تهئية المدونة (corpus) التي تحتوي على مجموعة من الجمل التي بدورها تحتوي على كلمات غامضة بسياقات مختلفة لجميع المعاني المحتملة لها، تستخدم هذه المدونة كمجموعة تدريب (training set) للطريقة ثم الاجراءات التالية:

$$\begin{aligned} & \arg_s \max p(s|x_1, \dots, x_m) \\ & = \arg_s \max p(s) \prod_{i=1}^m p(x_i|s) \end{aligned}$$

5. استرجاع المعنى المناسب الذي تم اختياره للكلمة الغامضة من المعجم.

- البحث في المدونة عن الكلمة الغامضة الموجودة في جملة الفحص وحساب عدد مرات تكرارها (n).
- استرجاع المعاني التي تظهر فيها الكلمة وعدد تكرار كل معنى c(si).
- حساب احتمالية كل معنى p(si).
- تكوين متجه لصفات الكلمة $f = \{f1, f2, \dots, fn\}$.
- حساب احتمالية وجود صفات الكلمة في كل معنى p(fj|si).
- في حالة وجود $p(fj|si)=0$ يستبدل بـ $p(si)/n$ لتجنب تأثير الصفر في معادلة (3) حيث انها قيمة صغيرة جدا لاتؤثر على القرار.
- يتم فك اللبس الدلالي من خلال تصفية المعاني بعيدة الاحتمال واختيار المعاني عالية الاحتمال أي اختيار المعنى الصحيح ذو الاحتمال الأكبر (likelihood) معادلة (3).

Input: ambiguous word w , words of sentence f .

Output: correct sense for ambiguous word si .

Calculating the total word ambiguous count in corpus n .

For all sense si of w do

$$P(si) = c(si)/n$$

End for

For all senses si of w do

For all words fj in the corpus do

$$P(fj|si) = c(fj, si)/c(si)$$

If ($p(fj|si) == 0$) then $p(fj|si) = p(si)/n$

End for

End for

For all sense si of w do

$$Score(si) = p(si)$$

For all words fj in f vector do

$$Score(si) = score(si) * p(fj|si)$$

End for

End for

Correct sense = max score (si)

الخوارزمية (1) خوارزمية تصنيف naïve Bayesian لإزالة الغموض

مثال لتوضيح طريقة العمل:

$$P(s_i) = c(s_i)/n$$

جملة الفحص: " يد الله مع الجماعة. "

الكلمة الغامضة: " يد "

$$P(s_2) = 3/8 = 0.375$$

لكلمة " يد " معاني عديدة حسب السياق الذي ترد فيه، وفي هذا المثال ولتقليل الحسابات اخذنا ثلاث معاني فقط وهي:

$$P(s_3) = 1/8 = 0.125$$

المعنى الاول (s1) : عضو من اعضاء الجسم.
• متجه الصفات للكلمة :
 $f = \{f_1, f_2, f_3, f_4\}$

المعنى الثاني (s2) : القدرة والسلطان.
= { " الجماعة"، "مع"، "الله"، "يد" }

المعنى الثالث (s3) : الجود.

• لو فرضنا تكرار كلمة " يد " في المدونة 8 وكانت موزعة حسب معناها كالتالي:

$$P(f_j | s_i) = c(f_j, s_i) / c(s_i)$$

$$S_1 = 4, S_2 = 3, S_3 = 1$$

• حساب احتمالية كل معنى:

Fi	p(fi s1)	p(fi s2)	p(fi s3)
يد	1	1	1
الله	0.0625	1	0.0156
مع	0.0625	0.333	0.0156
الجماعة	0.0625	0.0469	0.0156

• حساب تصنيف naïve Bayesian لكل معنى:

$$p(s | f_1, \dots, f_m) = p(s) \prod_{i=1}^m p(f_i | s)$$

$$P(s_1 | f) = 0.5 * 1 * 0.0625 * 0.0625 * 0.0625 = 0.00012$$

$$P(s_2|f)=0.375*1*1*0.333*0.0469=0.0058$$

✓

$$P(s_3|f)=0.125*1*0.0156*0.0156*0.0156=0.00000047$$

• اختيار أعلى احتمالية معنى وهي

$$Correct\ sense = \max score (s_i) = s_2 = 0.0058$$

وهي الاختيار الصحيح لمعنى كلمة "يد" حسب سياقها في جملة الفحص.

توجد السياقات الأخرى لها بعدد كبير وبعد تنفيذ الطريقة على مدونة السياقات أظهرت النتائج دقة عالية في الاختيار الصحيح للمعنى المناسب .

7- المدونة corpus :

احتوت المدونة على عدد من الجمل التي تحتوي على أفعال وأسماء متعددة المعاني و تم إضافة عدد من الجمل لتحتوي على جميع السياقات المختلفة لأي كلمة متعددة المعنى تم استخدامها وكانت النتائج جيدة جداً كما سيوضح من خلال التقييم اللاحق .

8- التقييم:

تم تقييم الطريقة وبإسناد مدونة تحتوي على مجموعة من الجمل المختلفة والتي تحتوي على مجموعة من الكلمات المتعددة المعاني للتجربة وبسياقات مختلفة لتعطي معاني مختلفة باستخدام مقياس الدقة **precision** والاسترجاع **Recall** ومقياس **F-score** كانت النتيجة كالتالي:

تم اقتراح مدونة أخرى تحتوي على جميع السياقات المختلفة لكل كلمة متعددة المعنى تم إدراجها في المدونة فقط أي كل سياق يدرج مرة واحدة لحل حالة الخطأ في الاختيار الصحيح عند وجود جملة واحدة تحتوي على سياق معين لكلمة غامضة بينما

$$P = \text{correct answers provided} / \text{answers provided}$$

$$R = \text{correct answers provided} / \text{total answers provided}$$

$$F\text{-score} = 2(P * R) / (P + R)$$

precision	90%
Recall	92%
F-score	90.98%

وقيمت بإسناد مدونة السياقات فكانت النتيجة كالتالي:

precision	95%
Recall	97%
F-score	95.99%

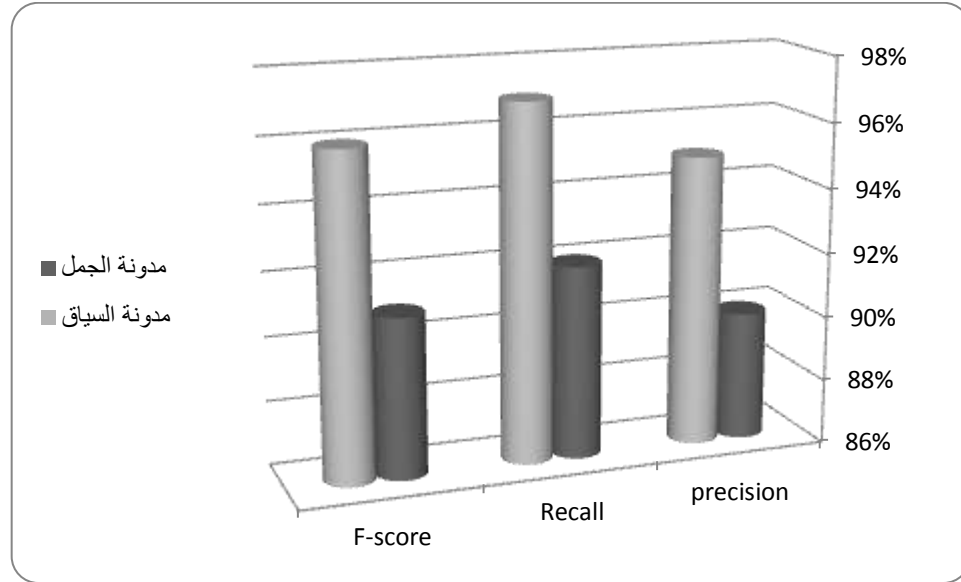
9- المناقشة والاستنتاج:

بعد إجراء تجارب متعددة على المدونة تبين ان زيادة عدد الجمل بسياقات مختلفة للكلمات المتعددة المعاني تأتي بنتائج جيدة جدا وقد بلغت دقة عالية في إعطاء القرار بالاختيار الصحيح للمعنى المناسب للكلمة .

لاحظنا ان حالات فشل الاختيار الصحيح جاء في حالة وجود سياق لكلمة ظهر في جملة واحدة فقط في المدونة بينما باقي السياقات كان لها تكرار كبير او في حالة عدم وجود السياق في المدونة. ولهذا صممنا مدونة اخرى تحتوي على جميع السياقات

المختلفة للكلمات المتعددة فقط اي كل سياق يظهر مرة واحدة فقط فكانت النتائج ممتازة ولم تحصل حالة فشل في الاختيار الصحيح للمعنى المناسب للكلمات المتعددة .

من هذا نستنتج ان محاولة توحيد تكرار السياقات في المدونة اعطى نتيجة جيدة في تقليل او عدم احتمال خطأ الاختيار الصحيح للمعنى كما موضح في المخطط(2).



مخطط (2) يوضح تقييم نتائج التصنيف على المدونتين

المصادر

for Myanmar language", international journal of scientific & engineering research v2, issue 9, September, issn 2229-5518.

Tom M. Mitchell, 2010, "Machine Learning", Copyrightc 2005, 2010.. All rights reserved. DRAFT OF January 19, 2010.

Tong Zhang, 2010, "Fundamental Statistical Techniques", taylor and francis group.

جبران مسعود، 1981، "الرائد - معجم لغوي عصري"، الجزء الاول والثاني، دار العلم للملايين، بيروت، لبنان.

د. احمد محمد المعتوق، 2000، "الالفاظ المشتركة المعاني في اللغة العربية طبيعتها- اهميتها- مصادرها"، جامعة الملك فهد للبترول والمعادن، الظهران.

د. عبد القادر الفاسي الفهري، 1982، "اللسانيات واللغة العربية نماذج تركيبية ودلالية"، منشورات كلية الاداب والعلوم الانسانية.

Eneko agirre & Philip Edmonds, 2007, "word sense disambiguation algorithms and applications", 167-216 , springer.

Karl-Michael Schneider, "Techniques for Improving the Performance of Naive Bayes for Text Classification",:

<http://www.phil.unipassau.de/linguistik/schneider>

Nitin indurkha & fred j. damerau, 2010, "hand book of natural language processing", by taylor and francis group.

Nyein thwet & khin mar soe, ni lar thein, 2011, "ambiguous Myanmar word disambiguation system for Myanmar-English statistical machine translation", international journal of computer applications (0975-8887) v27-no. 8, august.

Nyein thwet & khin mar soe, ni lar thein, 2011, "wsd system using naïve Bayesian algorithm

Word Sense Disambiguation for polysemy in Arabic language using Naïve Bayesian classification

D. Yusra Malik Dhammed

Computer science\College Science

Basra University

Yus_malik@yahoo.com

Abstract

The semantic ambiguity is the biggest challenges facing the natural language processing and it's applications automatically in general and specially in the Arabic language, and of polysemy phenomenon is one of these problems that causes ambiguity in the selection appropriate meaning of the word in a sentence. The current trend in natural language processing is statistically because of its effectiveness in solving some of the problems of applications.

This research presented word sense disambiguation for multiple meaning of the word in the Arabic language by one of the statistical methods that used in this field which is the classification method (naïve Bayesian), in addition to a corpus designing that contains all the different contexts in which the ambiguous words appears. The approach achieved a high success rate in the identification and selection of the correct meaning of the ambiguous word in the sentence amounted to 90.98% with corpus of sentence and 95.99% with the corpus of contexts.