Handling Text Mining Problems in Arabic using Domain-Specific Approach

Madeeh Al-Gedawy
Information and Decision Support Center
Egyptian Cabinet

Osman Hegazy
Department of Information Systems
Cairo University

ABSTRACT

Latin-based languages work smoothly within the traditional text mining techniques due to being definite and the natural limited alternatives of words meanings. On the other hand, in the Arabic language, we are facing 2 main differences: 1) the way, the Arabic language is being written today without diacritics in 99% of the text will make the text interpretation at the level of two consecutive words and even in some cases at the level of sentences indefinite 2) even with diacritics, Arabic words are very loose; each word in Arabic may bear more than one meaning regarding the context. Hence handling text in Arabic in the same manner that Latin languages do, will be rather time wasting. We need to rely on different techniques in order to enrich the criteria which will be adopted in text analysis.

We propose a domain-specific approach that yielded excellent results with some of Arabic text analysis aspects.Several classifiers have been built and tested for this purpose. This approach was compared to others that don't use the domain-specific approach; the paper concludes that the results obtained from the adopted technique are more appealing and promising.

Keywords

Arabic domain-specific – Fuzzy Logic – Weighted Decision Trees– Classification – Word Sense Disambiguation – Query Expansion – WordNet.

1. INTRODUCTION

The Arabic language is one of the most difficult languages to Classification and Word Sense Disambiguation (WSD) on, because the Arabic Language has a loose and tricky Word Order (WO), and word diacritization are usually absent from modern Arabic so one can't use diacritization marks to help in understanding the sentence structure, moreover, a single word in Arabic may have dozens of meanings, some of them are closely related, and interestingly there are some words that can mean something and its opposite[3].

Adding a domain-based layer to the classification was successfully applied in English by Gerard Escudero,Lluis Marquez and German Rigau. After they applied a classification WSD on 2 different corpuses, they commented: "The previous experiment shows that classifiers Trained on the A corpus does not work well on the B corpus, and vice-versa. Therefore, it seems that some kind of tuning process is necessary to adapt supervised systems to each new domain."[10]

This indicates that domain-specific classification yields better analysis and better understanding for the cases. Moreover, it is more logical than cross-domain classification; the author really likes the comment of Sara Owsley: "We believe it's necessary to use domain specific language to classify the emotional content. Domain specificity is critical in making this system work, since the language used to describe automobiles (sleek, maneuverable, etc.) is different from the language used to describe vacation destinations (relaxing, adventurous, etc.)."[13], and the same can be inferred about Arabic; it is always correct to say: فتاة رقيقةbutit is to wrong to say: سيارة رقيقة.

Another problem that domain-specific approach solves is word ambiguity; the lake of diacritics can lead to a perfect Word Sense Disambiguation problem. For example, the root قبل can be written as kabl (before), kabbal (kiss), or kabela (agree) without diacritics will be the same word قبل.

Knowing the right sense for the word and knowing the correct domain of the text can be of great benefit for text retrieval issues; it can be used for filtering the results, giving suggestions to the users and sometimes, for query expansion purposes.

Domain-specific Classification gives the ability to extract the main entities (Most Frequent Words) in each domain, and extract the features associated with these entities. Thus, the features existence will differ as the entities differ, and the entities will differ as the domain differs.

The paper is structured as follows. In Section 2, we present a brief description of our Arabic domain-specific classifier components. Section 3 we illustrate some experiments and comparisons, and in section 4, we give some remarks and future directions.

2. SYSTEM DESCRIPTION

Our Arabic specific-domain classifier is built up out of4 components: Stemmer, Classifiers, Sense Inventory and Web Directory.

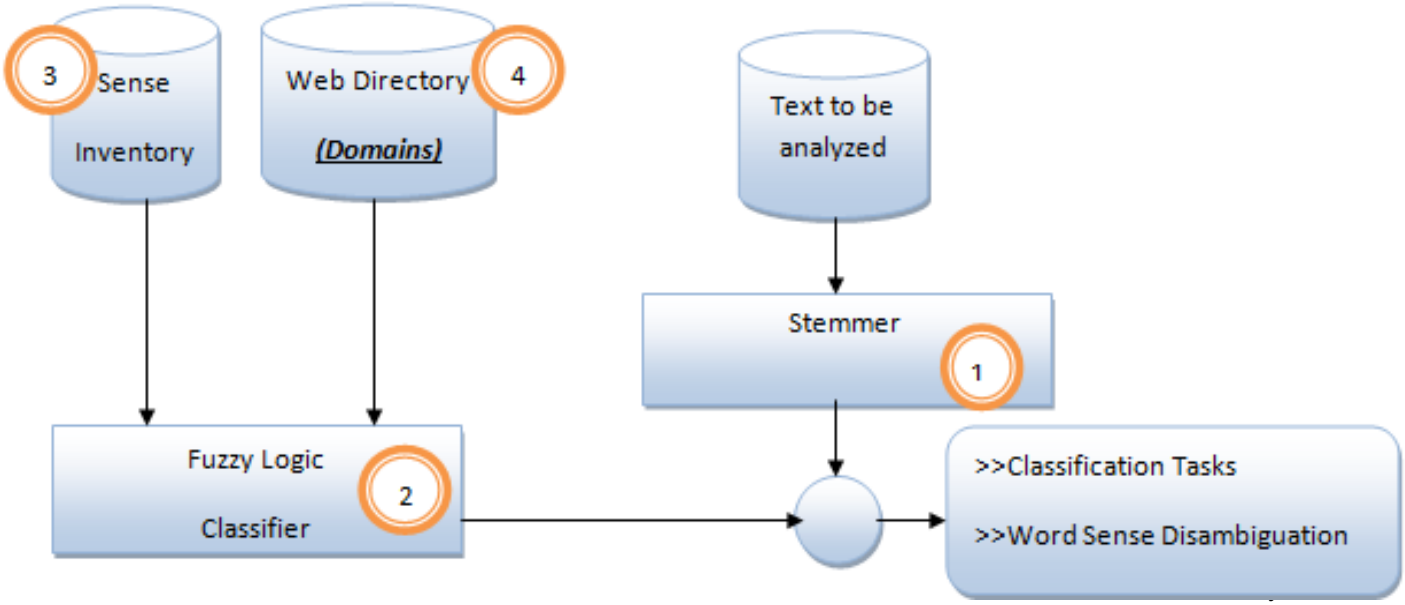

Table 1.Rooting Technique

Letter Position from right	Rank if word length is even	Rank if word length is odd
1	N	N
2	N-1	N-1
3	N-2	N-2
….	….	….
N/2	N/2+1	N/2
N/2+1	N/2+1-0.5	N/2+1-1.5
N/2+2	N/2+2-0.5	N/2+2-1.5
….	….	….
N	N-0.5	N-1.5

Each word represents a feature. The frequency of word occurrences in every document is recorded as Term Frequency (TF), the frequency of a word mentioning across all documents in recorded as Document Frequency (DF). The more the TF is bigger, the more the word is important, the

that the classifiers accuracy are slightly different from each other.

The author built a Fuzzy Logic Classifier and compared it 2 other classifiers: Decision Trees and Naïve Bayes .It was noticed thatFuzzy Logic classifier yields more significant results.

2.2.1 Decision Trees

A Decision Tree visually represent a decision; it describes the data and can be used in making decision by predicting the value of a dependent variable based on a combination of input independent variables (features or attributes). The tree is drawn by dividing the training set into subsets based on an attribute. This process is repeated recursively on these subsets (recursive partitioning). A criterion is set to stop the recursive process when no longer added value will be achieved by splitting the tree (pruning).It is based on the concept of Entropy (E) (amount of uncertainty associated with a variable), Assuming binary classes {p+,p-}. Entropy (information needed to classify a tuple) is defined as:

$$Entropy = -[(p+.(\log p+)) + (p-.(\log p-)]$$

The Entropy is used to specify the overall information needed to classify a tuple using this decision tree. To choose the variable upon which the next split will take place at, the node that most probably will improve the classification is selected, we calculate the feature information for every attribute and then we choose the attribute with the greatest information gain.

2.2.2- Naïve Bayes

Naïve Bayes is not a discriminant classifier; it is a generative classifier where every class is modeled and then the tested exemplars are exposed to these models. NB is extensively

used in text classification (TC) tasks. This is the basic pseudocode for the algorithm:

for all senses si of word (w) do

for all words tj in the vocabulary do

P(tj|si) = C(tj,si)/C(s)

end

end

for all senses si of word (w) do

P(si) = Frequency[Class(si)] /Frequency[Class(w)]

End

NB does no feature selection; it is combined from all the features. As stated in the pseudocode, NB applies Baysian rules to choose the relevant class where P(tj|si) is the perior probability of sense si and it is calculated as:

$$P(tj|si) = [P(si|tj))/P(si)] * P(tj)$$

2.2.3 Fuzzy Logic

"Fuzzy logic provides a means for encapsulating the subjective decision making process in an algorithm suitable for computer implementation" [Jan A Hazelzet- Can fuzzy logic make things more clear?].
"Fuzzy expert systems, in addition to dealing with uncertainty, are able to model common sense reasoning which is very difficult for general systems", "fuzzy logic is a relative reasoning logic not a precise multi-value logic" [Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization].

"In some applications fuzzy logic is an alternative to Bayesian inference. Fuzzy logic and Bayesian inference, however, are mathematically and semantically not compatible. You cannot, in general, understand the degree of truth in fuzzy logic as probability and vice versa; fuzziness measures "the degree to which an event occurs, not whether it occurs" [Wikipedia].

"manystatisticians are persuaded by the work of Bruno de Finetti that only one kind of mathematical uncertainty is needed and thus fuzzy logic is unnecessary. On the other hand, Bart Kosko argues that probability is a subtheory of fuzzy logic" [Wikipedia].

Actually the mathematical foundation of Fuzzy Logic is not as solid as that of probabilistic models but it made great success in reality and in business.

Algorithm Description

Suppose that the context 'c'contains the ambiguous word 'w' and that the context has a window of 'k' words and that w has 'n' senses so for fuzzification; fuzzy set is to be defined for every word sense 'S' So that the fuzzy set (FS) of sense 'y' is[7]:

$$FS\ (Sy) = U\ [D(w1), D(w2), D(wk)]\ (1)$$

Where D(w1) is a membership for the first word w1 in the context 'c' to be allocated to the sense 'y'; in other words, it works as a probability distribution or a density function but the summation of densities of w1 in respect to senses does not oblige to 1; and that shows the difference between possibility and probability theories. The U is the union operator; in general the union operator in fuzzy logic is equivalent to max operator. For example: U(a,b)=max(a,b) not a+b-ab but in this experiment the research is stuck to the ordinary a+b-ab to get advantage of the idempotency feature so the words of greater frequency will be highly weighed.

In this algorithm, the word 'w'will be assigned to the sense of the greatest order; and here is the pseudo-code that orders the different word's senses (this code combines the 2 phases of inference and defuzzification)[5]:

For y= 1 to sense 'n' do

Set membership of context 'c' to sense y 'D(Sy)'to 0

For i=0 to word k in context 'c'

Set flag=Exists (training set, wi)

If (flag=true) then

'D (Sy)'=U ('D (Sy)',Myi)

End if

End for

End for

Where Myi is the membership of word wi assigned to sense 'y'extracted from the training examples; it is calculated using a sigmoid function and takes into account the higher frequencies:

$$Myi = 0.3 + 0.7\ [\frac{1}{(1 + e - 2\ (f))}]$$

Where f is the frequency of the word in the training set for a specific sense. The training corpus is used as knowledge base for making inference and setting the memberships [6].

Measuring Performance

Given the F-measure as:

$$F = 1/(\left[\frac{b}{P}\right] + [(1-b)/R])$$

Where b=0.5

P (Precision) = X / (X+Y)

R (Recall) = X / (X+Z)

Where:

X=number of senses positively classified by both human and the algorithm=5

Y = number of senses positively classified by the algorithm while negatively classified by the human=0

Z = number of senses positively classified by the human while negatively classified by the algorithm=0

2.2.4Experiment

This experiment's training and test sets focuses on social. The ambiguous word to be investigated is "قاعدة" or "قاعده" which has several meanings:

1. Rule
2. EL-Ka3da (a terroristic group)
3. A database
4. A base
5. Mass of people
6. Sitting (colloquial)

The experiment covers 3 meanings: rule, El-Ka3da and sitting. Here is the social training set:

Table 2. Training Set

Context	Sense
الالفاظ النابية هى القاعدة الأساسية لدى هذه الطبقة وهم يقومون بالتمسك ببروتوكولات السباب بدقة متناهية لأنه يلبى أهداف نفسية نبيلة	Rule
القاعده الجديدة: الجريمة قد تفيد	Rule
أحمد بهجت: الحكومة المصرية عامله ذى أى عيلة مصرية ماشية بقاعدة: حط الكوز تحت البزبوز	Rule
أنا غير متزوج !! إذا أنا موجود هي قاعدة في مجتمعنا الخليجي وتوسعت لتصبح قاعدة يرددها شبابنا العربي كذلك	Rule
و عندنا فى المرور المصرى قاعدة البوز بيز يعنى اللى يفوت يبوز عربيته الاول هوا اللى يمر	Rule
وتقدر احصاءات رسمية نسبة الفقر في موريتانيا بـ 46% من مجموع السكان، 75% التفاوتات الاجتماعية والمجالية وتحسين ... قاعدة موارد الفقراء	Rule
سمية الخشاب ليحيا سعادة :"أنا مش قاعدة على بنك!"	Sitting
طلعت فوق السطوح زق الهوا كمى كل البنات اتجوزوا و انا قاعده جنب أمى	Sitting
قاعدة على قلبكم عشان خاطر الناس الحلوة	Sitting
و كانت اختها القبيحة قاعده وسطنا	Sitting
مرحبا عندي سؤال قاعده على اعصابي اني دورتي منتضمه كل 30 يوم وتئخرت الدوره علي يوم واحد ورحت المستوصف وحللت دم وطلع سالب	Sitting
هوة انتى علشان قاعدة لوحدك ... مانا جالى اخويا المحترم يطبق على مراوحى يعنى مش لوحدى ولا حاجه والماما والبابا بتوع الانا صحيوا كمان	Sitting
و في يوم كنت قاعده و لقيت مدام فتونه السكرتيره قاعد معاها مهندس جديد و هاتك يا رغي. ودودودو. قمت بقي بحاسه الفضول الي عند كل البنات	Sitting
البنات و الولاد دلوقتى بيعدوا فى الشارع للفجر و يرجعوا سكرنيين. مستقبل العيال ده مظلم	Sitting
وسائل لتدجين القاعدة و رفضها من 10 المجتمع المدني	El-Kad3a
اليمن: مشاكل الفقر والفساد أكبر من خطر "القاعدة""	El-Kad3a
انجذاب بعض الشباب للقاعدة يعد تهديدا شديدا يحتاج إلى دراسات نفسية و اجتماعية مكثفة من قبل الأكاديميين و صناع القرار	El-Kad3a
و أثار ضرب القاعدة للولايات – إن صح – كان فادحا على المسلمين المقيمين هناك	El-Kad3a

And here is the social test set:

Table 3. Test Set

Serial	Context
1	معلش بقا جوزى كان واخد أجازة اليومين دول ومكنتش بعرف ادخلكوا طابق على مراوحى انا قاعدة مستنياكو
2	خوف فى المدن الأوربية من احتمال ضربة ارهابية وشيكة من قبل تنظيم القاعدة
3	وانا كنت قاعده متنحه للى بيحصل وبما انى جديده ومعرفش اى حد فكنت قاعدة برطة واتفهم انى ال ايه خير الله ما اجعلة خير هادية وغلبانه ومرمية جنب الحيط
4	رايت اني كنت قاعده في غرفتي وقاعده على فراشي وكنت اقرى في كتاب
5	و من الملاحظ أن أخلاق الزحام لها قواعد خاصة بها تدعو للرثاء فى مصر

Sentences: 1, 3 and 4 are associated to the sense "Sitting" with a one hundred percent confidence. Sentences 2 associated to the sense "El-Ka3da" with a one hundred percent confidence and Sentences 5 associated to the sense "Rule" with a one hundred percent confidence. The F-measure for fuzzy logic is

1, the F-measure for Naïve Bayes is 0.89 and for decision trees is 0.91. This is why Fuzzy logic has been selected to be the Arabic domain-specific main classifier.

2.3 Sense Inventory

The author built an Arabic sense inventory based on the English WordNet Thesaurus. WordNet developed at the University of Princeton. It covers lexico-semantic categories called synsets. The synsets are sets of synonyms which gather lexical items having similar significances, for example the words “a board” and ”a plank” grouped in the synset {board, plank}. But ”a board” can also indicate a group of people (e.g., a Using WordNet for Text Categorization 17 board of directors) and to disambiguate these homonymic significances “a board” will also belong to the synset {board, committee}. There is a wealth of semantic relations available in WordNet through synonyms where bind equivalent or close concepts[4].

Every word synonym in WordNet is associated with several examples; these examples are complete sentences that demonstrate the usage of the word synonym within context. These sentences have been automatically translated into Arabic using 2 APIs packages: Google Translation API and Microsoft Bing; the Arabic translations for the word synonym is extracted from the translated examples and the words of the Arabic translated examples are used as keywords associated with the extracted Arabic word synonym. Here is an example for the Arabic words generated that have the root: و–ل-د:

Table 4.System Block Diagram

Arabic Word	English Word	Category	Sense_id
ولد	Infant	N	12122
ولد	Give birth	V	54111
ولد	generate	V	32511
مولد	generator	N	32655
مولد	Prophet's birthday	N	25412
مولد	birthday	N	23265
والد	Father	N	36521
والدة	Mother	N	21544
وليد	New born	N	36251
ميلاد	Birthday	N	2514
ولادة	birth	N	25654

Using this component with the fuzzy classifier, a good Arabic Word Sense Disambiguation sub-system was achieved as illustrated in figure 2[11], but no domain-specific features is plugged in yet.

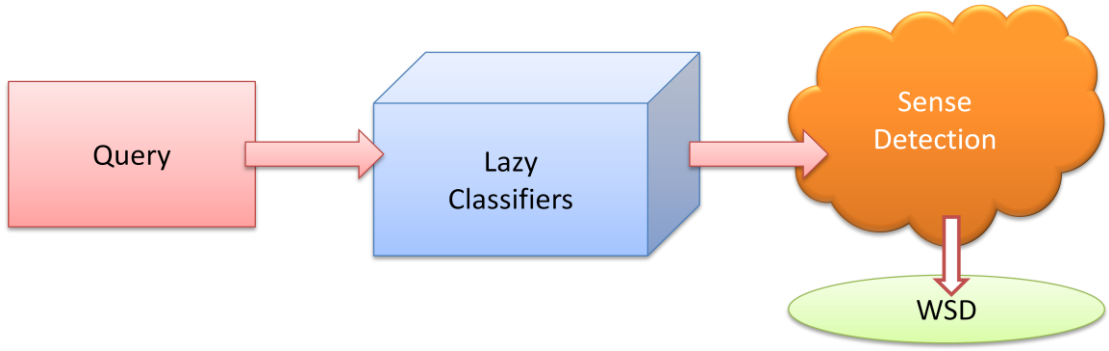

Fig 2: Solving Arabic WSD using generated sense inventory

This component was greatly enhanced by adding a domain-specific layer which enabled building a Domain Model (DM) that improves the ordinary vector space model (VSM) [12]; the main structure of a domain model is illustrated in table 5.

Table 5.Domain Matrix Example

	Domain 1	Domain 2	Domain 3
Sense 1	0.7	.18	0.12
Sense 2	0.35	0.4	0.25
Sense 3	0.09	0.9	0.01
Sense 4	0.1	0.09	0.81

2.4 Web Directory

The author built the domain categories by dumping the open source web directory "ODP". Here is a list of the main domains categories [13]:

Fig 3: ODP main categories- retrieved from: http://www.dmoz.org/ [8]

All the domains have been translated into Arabic using Microsoft Bing API services. The fuzzy classifier model has been built upon these domains. Whenever, a query is submitted, the classifier runs as allocate the text to the best domain. The advantages gained by taking the domain effect properties into consideration are limitless [9]; this feature has been enabled in the system as illustrated in the following application example.

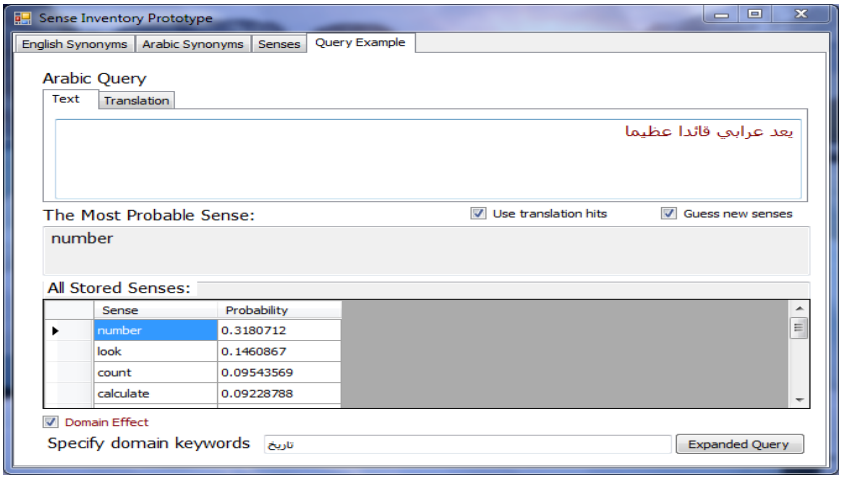

Fig 4: Domain-Specific Feature Enabled

Before running the classifier on the query, the system does a query expansion using the embedding keywords and it gets a larger set of keywords using the internet as a bigger corpus; the larger set of keywords are added to the original case but with small weights[2][16].

3 EXPERIMENTS AND ENHANCEMENTS

3.1 Examples

Here are some query examples that have been analyzed using our Arabic domain-specific fuzzy classifier:

Table 6.The Domain Effect enhancing the accuracy from 97.2% to 100%

Example 1- Domain: Politics			
" قال معلق عسكري صيني بارز يوم الاثنين ان من الواضح أن الولايات المتحدة ترغب في تطويق الصين في واحدة من أعنف الانتقادات للحملة الدبلوماسية التي يشنها الرئيس الامريكي باراك أوباما في منطقة اسيا والمحيط الهادي"			
Detected Domain	**Probability**	**بارز بمعنى ناتئ**	**بارز بمعنى ذو مكانة**
Regional/North_America/United_States/Government/Executive_Branch/President/Executive_Office_of_the	66.6% (8 levels depth)	1%	99%

President/Domestic Policy_Council			
Regional/Asia/China estic_Policy_Council	33.3% (3 levels depth)	2%	98%

Confidence	بارز بمعنى ناتئ	بارز بمعنى ذو مكانة
66.6% (8 levels depth)	1%	99%
33.3% (3 levels depth)	2%	98%

Table 7. The Domain Effect enhancing the accuracy from 62.8% to 74.15%

Example 2- Domain: Sex

" علياء المهدى: معظم المصريين متكتمين فيما العلاقة الجنسية ويصفونه بالقذر ، والبعض يعتبره رجل يستخدم المرأة . بالنسبة لي الجنس تعبير عن الاحترام وشغف للحب المتمثل في الجنس لإسعاد الطرفين.. أمارس الجنس الآمن ولا أتناول أقراص منع الحمل لأنني ضد الإجهاض.. فقدت عذريتي في سن 18 عاماً مع رجل يكبرني بـ40 عاماً"

Detected Domain	Probability	بارز بمعنى ناتئ	بارز بمعنى ذو مكانة
Society/Sexuality/Sacred_Sexuality	25% (3 levels depth)	97%	3%
Adult/Society/Sexuality/Fetishes/Body_Parts	50% (5 levels depth)	99%	1%
Health/Senior_Health/Sexuality	25% (3 levels depth)	31%	69%

Confidence	بارز بمعنى ناتئ	ز بمعنى ذو مكانة
25% (3 levels depth)	97%	3%
50% (5 levels depth)	99%	1%
25% (3 levels depth)	31%	69%

D(sex)=49.5/(49.5+17.25)= 74.15%

3.2 Optimizing sub-domain path using weighted decision trees

Some words in the analyzed paragraph or article are irrelevant and produce noise that gives unaccepted probability to misleading paths.Idea: we give an extra weight to the term according to the importance of the sentence mentioning.The same techniques used in text summarization are exploited to order the sentences.

3.2.1 Using Cosine Similarity

If there is a title provided for the article or paragraph, then it is assumed that the title reveals some disambiguation.A cosine similarity is conducted between every sentence and the title where s is sentence and t is the title vector.

3.2.2 Using TFIDF Technique

Term Frequency (TF) and Document Frequency (DF) are 2 basic measures that are used to rankdocuments importance. Combining both techniques, sentence score will be:

$$SentenceScore(S) = a * Sim + b * Cen$$

Where a and b are constants that controls the learning ratios. Then the weighted tf (WTF) of a term is:

4 CONCLUSION AND FUTURE WORK

The Arabic domain-specific classifier technique enables the researchers to accurately classify and disambiguate Arabic words. THE technique proved to yield better results in classification, Word Sense Disambiguation, Text Retrieval, Query Expansion and filtering results.

The Arabic domain-specific classifier technique depends on integrating several components to build an excellent Arabic classifier of multiple usages: a stemmer that guesses the roots of words heuristically using Al-Shalabi algorithm, a fuzzy logic classifier which proved to be excellent for the Arabic language due to the vagueness features, a sense inventory built based on the robust WordNet thesaurus; this component is used in Word Sense Disambiguation problems and a web directory dumped from the open project ODP which provides a wealth of domains and their keywords. All the translation tasks have been automated using both Google Translation and Microsoft Bing APIs.

Comparing the classifier in a Word Sense Disambiguation problem using specific-domain information provided by the web directory proved to yield much better results than the cross-domain approach that does not take into account the domain effect. The results was further improved and tuned using a weighted decision tree that assigns weights to nodes based on summarization techniques rather than the simple TF-IDF approach.

For future work, the research is eager to investigate the semi-supervised approach (usually bootstrapping and co-training). Because in supervised methods we subjectively choose the features for the classifiers, but now more examplars are fed into the classifiers thus the space will be well-inspected, and more features will be utilized. The auther likes to investigate in solving Arabic WSD using word order not only the context by using a mix of uni-directional association rules and Hidden Markov Models (HMM).

5 . REFERENCES

[1] Naïve Bayes Classifier for Arabic Word Sense Disambiguation.In Proceedings of the INFOS2008, Cairo-Egypt, March 27-29.Farag, A., and Andreas, N. 2008.

[2] Agirre, E.; Lopez de Lacalle, A.; Soroa, A. (2009) "Knowledge-based WSD on Specific Domains: Performing better than Generic Supervised WSD" Proc. of IJCAI.

[3] Roberto Navigli. Word Sense Disambiguation: A Survey, ACM Computing Surveys, 41(2), 2009, pp. 1–69.

[4] Navigli, R.; G. Crisafulli. Inducing Word Senses to Improve Web Search Result Clustering. Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010), MIT Stata Center, Massachusetts, USA.

[5] Y. Yuan and M.J. Shaw, Induction of fuzzy decision trees. Fuzzy Sets and Systems (1995), pp. 125–139.

[6] Arabacioglu, B. C. (2010). "Using fuzzy inference system for architectural space analysis". Applied Soft Computing 10 (3): 926–937.

[7] Jan A Hazelzet. Can fuzzy logic make things more clear? Critical care (London, England) 2009

[8] Open Directory RDF Dump, retrieved from: http://rdf.dmoz.org/

[9] David F. Prenatt, Jr., Life after the Open Directory Project, Traffick.com (June 1, 2000).

[10] An empirical study of the domain dependence of supervised word sense disambiguation systems, 2000 Article. Bibliometrics Data Bibliometrics.

[11] S. Mohammad and G. Hirst. Determining word sense dominance using a thesaurus.In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), 2006.

[12] LAPATA, M. AND KELLER, F. 2007. An information retrieval approach to sense ranking. In Proceedings of HLT-NAACL, Rochester, NY.348–355.

[13] Sara Owsley, Sanjay Sood ,Kristian J. Hammond. Domain Specific Affective Classification of Documents.In Proc. of the AAAICAAW (2006).

[14] bénIzquierdo, Armando Suárez, German Rigau. An empirical study on class-based word sense disambiguation.Proceedings of the 12th Conference of ACL EACL 2009.

[15] Sanderson, Mark. Ambiguous Queries: Test Collections Need More Sense. In Proceedings of the 31st annual international ACM SIGIR (SIGIR '08), pp. 499–506, New York, NY, USA, 2008.

[16] Wikipedia as Sense Inventory to Improve Diversity in Web Search Results C. Santamaria, J. Gonzalo, J. Artiles, Proceedings of ACL 2010.