Journal of Computer Science 7 (10): 1612-1618, 2011
ISSN 1549-3636

A Design and Development of Word Sense Disambiguation Algorithm for English Language Understanding for Database Access

M. Munusamy, G. Tholkappia Arasu, V. Palanisamy and S. Selvarajan
Department of Computer Science and Engineering,
Jayam College of Engineering and Technology,
Hogenakkal Falls Main Road, Nallanur, Tamil Nadu 636813, India

Abstract: Problem statement: This study attempts to present an object-net method for word sense disambiguation. It is proposed to model the elementary meanings which assist the machine to autonomously undertake the analysis and synthesis processes of meaning. **Approach:** In the proposed methodology, the disambiguation process was performed in context manner. Starting from natural text, the context of the sentence was identified, then the actual meaning identified using correlation of elementary object meanings existed in object-net database. It was because even ambiguous word will have only one meaning based on the context or object or domain on which the sentence was written. **Results:** This object-net approach disambiguates original text with high precision of 96% of the verbs and 97% of nouns for data extraction from the database and reporting in terms of graphs. **Conclusion:** The accuracy of finding the sense of a word and extracting data from the database and projecting into graphs was based on number of trained objects in object-net database. Due to this object-net database plays a major role in this proposed method.

Key words: Word Sense Disambiguation (WSD), Natural Language Processing (NLP), parse tree, object-net database, disambiguates original text, new methodology, shallow approaches, supervised methods, synthesis capability

INTRODUCTION

Word Sense Disambiguation (WSD) is the process of identifying which sense of a meaning is used in any given sentence, when the word has a number of distinct senses (Carpuat and Wu, 2005). For a long time the WSD is an open problem in natural language processing (NLP). The solution of this problem impacts other tasks such as discourse, engines, anaphora resolution, coherence, inference, information retrieval, machine translation and others.

There are two main types of approach for WSD in natural language processing called as deep approaches and shallow approaches.

Deep approaches: These approaches involve the intention to understand and create meaning from what is being learned, Interact vigorously with the content, make use of evidence, inquiry and evaluation, Take a broad view and relate ideas to one another and Relate concepts to every time experience. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer-readable format, outside of very limited domains. There is a long tradition in computational linguistics (Abney, 2004), of trying such approaches in terms of coded knowledge and in some cases; it is hard to say clearly whether the knowledge involved is linguistic or world knowledge. The first attempt was that by Margaret Master-man, at the Cambridge Language Research Unit in England, in the 1950s and Yarowsky's machine learning optimization of a thesaurus method in the 1990s.

Shallow approaches: These approaches are not concerned of learning the text instead they deal with the surrounding words of the ambiguous word and try to identify only parts of interest for a particular application. They just consider the surrounding words, using a training corpus of words tagged with their word senses the rules can be automatically derived by the computer (Mokhtar *et al*., 2002). This approach, while theoretically not as powerful as deep approaches, gives superior results in practice, due to the computer's limited word knowledge.

Corresponding Author: M.Munusamy, Department of Computer Science and Engineering,
Jayam College of Engineering and Technology, Hogenakkal Falls Main Road, Nallanur,
Tamil Nadu 636813, India

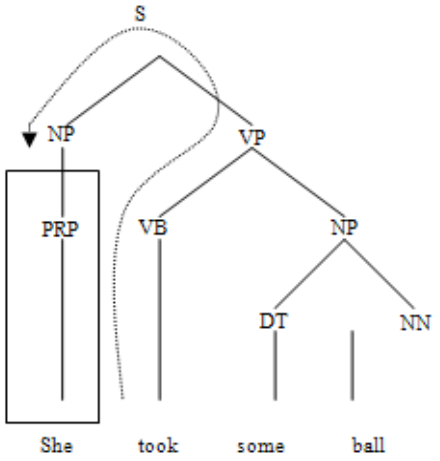

Fig. 1: An example parse tree path from the predicate "took" to the argument "She", represented as ↑VB↑VP↑S↓NP

In addition to deep approaches and shallow approaches, there are four conventional approaches to WSD.

Dictionary and knowledge-based methods: These approaches make use of dictionaries, thesauri and lexical knowledge bases, without using any corpus evidence.

Supervised methods: These approaches make use of sense-annotated corpora already been trained from semantically disambiguated corpus.

Semi-supervised or minimally-supervised methods: These approaches make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

Unsupervised methods: These eschew (almost) completely external information and works directly from raw corpora (i.e. not annotated) (Diab and Resnik, 2002).

The method proposed here is a semi-supervised method; it is called as object - net approach which uses the information dynamically gathered from user that is while machine finds any of untrained corpora or unable to solve the disambiguation then those information are reported to user or master, after user understand the problem the related corpora are trained. It differs from previous semi-supervised approaches: The algorithm has a set of disambiguated trained elementary objects (Rajaraman and Tan, 2001) and incrementally builds and resolves the untrained elementary objects; this approach allows identifying the semantic sense of input word with high precision of 96% of the verbs and 97% of nouns.

The algorithm presented here is an improvement over other existing algorithms in WSD and data extraction from database using English language instead of database query languages like SQL (Aziz *et al*., 2011); this algorithm can be incorporate into lager applications like machine translation, code generation, search engine, IR.

Resources: The algorithm does not dependant on any other existing WSD resources like WordNet, SemCor. Instead of that it uses separate database named as Object-Net Database which contains trained elementary objects. Initially the database is stored with limited data, this database updated when new untrained object found in the input text or when fine tuning is required on existing already trained element (Burges, 1998). The proposed algorithm finds all its required information (Tiun *et al*., 2010) to identify the meaning of the word on a particular context from this Object-Net database, so precision of word sense disambiguation of proposed algorithm mainly depends on data from this special Object-Net database.

MATERIALS AND METHODS

The algorithm presented in this study determines, in a given text, a set of nouns and verbs which can be disambiguated with high precision, the semantic tagging is performed using the sense defined in object-net database and actual meaning of the sentence is identified. But above mentioned task are completed in step by step using methods, so the various methods used to identify the correct sense of a word are presented first, Next presents Object-Net Database architecture, the main algorithm in which these procedures are invoked in an iterative manner and the method of updating, fine tuning the Object-Net Database.

Procedure #1: This procedure tokenizes the given sentence and creates a parse tree path for the given sentence. Parse tree paths were used for semantic role labeling. Predicates are typically assumed to be specific target words (verbs) and arguments are assumed to be spans of words in the sentence that are dominated by nodes in the parse tree. A parse tree path can be described as a sequence of transitions up from the target word then down to the node that dominates the argument span. The parse tree paths are particularly interesting for automated semantic role labeling because they generalize well across syntactically similar sentences. For example, the parse tree path in Fig. 1 would still correctly identify the "taker"

argument in the given sentence if the personal pronoun "she" were swapped with a markedly different noun phrase (Shaalan *et al*., 2004).

Procedure #2: Identify the words having only one sense (monosemous words) in Object-Net database and make them as having number of sense as #1.

Example: The noun subcommittee has one sense defined Object-Net database. So this is a monosemous word and marked as having sense #1.

Procedure #3: with this procedure, we are trying to get contextual clues regarding the usage of the sense of a word. For a given word Wi, at position i in the text, form two pairs, one with the word before Wi and the other one with the word after word Wi. Then we find out all the occurrences of these pairs found within the Object-Net database. If, in all the occurrences, the word Wi has only one sense as # Wis, then mark the word Wi as having sense #Wis.

Procedure #4: Find the words which are semantically connected to the already disambiguated words for which the connection distance is 0. The semantic distance is computed based on the ObjectNet hierarchy. Two words semantically connected at a distance of zero if they belong to same path of subnet.

Procedure #5: Find words which are semantically connected in object net and for which the connection distance length is zero. In this procedure none of the words considered by this procedure already disambiguated. We have to consider all the sense of both words in order to determine whether or not the distance between them is zero, this makes this procedure computationally intensive.

Procedure #6: Form the semantic network based on understanding made by the learning done from procedure #1 to procedure #5 and come to the final conclusion about the input sentence and action to be performed.

The procedures presented above are applied iterative; this allows us to identify a set of nouns and verbs which can be disambiguated with high precision. This object-net approach disambiguates original text with high precision of 96% of the verbs and 97% of nouns.

Object-net database architecture: The existing knowledge bases in machine readable formats are WordNet, OMCSNet, MindNet, CYC, Thought treasure, VerbNet, Semcor, Open Mind Word Expert, Frame Net and PropBank.

These knowledge bases are useful to serve the purpose of developing information retrieval systems and shallow semantic representation for an input text (Chen *et al*., 2004). They model their elementary meanings only with conceptual world properties and constraints and taxonomic relations between these words. They do not have synthesis capabilities, but rather their definitions are pre-programmed by humans. They do not make the machine creative enough to master its own language and to compose its own text based on its understood meanings. So a new methodology is required for machine to autonomously undertake the learning, analysis and of both the elementary and composite meanings of natural language and most importantly, it is to note that the robustness of proposed algorithm by machine relies not only on sophisticated algorithms for knowledge manipulation but also the kind of knowledge it has. (i.e., careful modeling of elementary meanings from an engineering point of view). The new methodology for maintaining trained elementary meaning is called Object-Net database and details of this database is explained in analytical and synthesis capability section.

Algorithm with an example: Consider for example to retrieve data from any of user database like "I need the student report that joined on 04 November 2010."

Procedure #1: Tokenize the given sentence as below:

"I + need + the + student + report + that + joined + on + 04 + November + 2010."

While categorizing these token words the below result is found:

"Pro+Ver+Art+Nou+ver+pro+ver
+adv+Num+Nov+Num"

Create the parse tree after tokenizing, the Fig. 2 shows the parse tree for above mentioned example sentence.

Table 1: Parsed tokens and its relation.

Pairs	Description
I + need	Whom->I
Need + the student	What->the student
The student + report	What->student
Report + that	Unable to correlate
That + joined	Which->joined
Joined + on	Unable to correlate
On + 04 November 2010	Which->date

Procedure #2: Find the words which are having unique sense of meaning and find object on which the action need to be performed:

"I (Sense#1) + need (Sense#1) + the (Sense#1) + student (Sense#1) + report + that (Sense#1) + joined + (on + (04 + November + 2010)) (Sense#1)"

In this example the word "I", "the", "student" "that" and "date (04 November 2010)" are having only one sense of meaning and student is the object on which the sentence related.

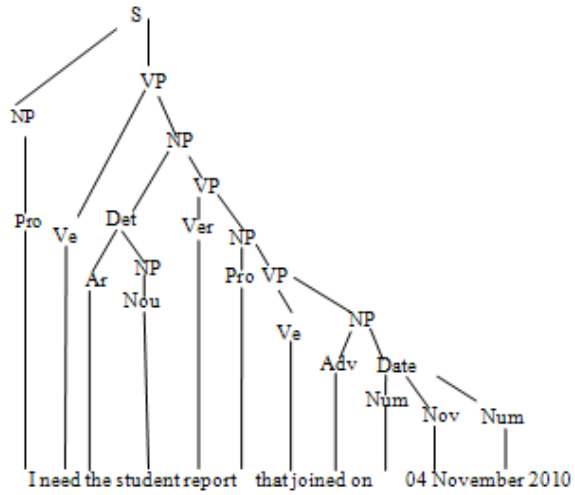

Fig. 2: Parsing of example sentence

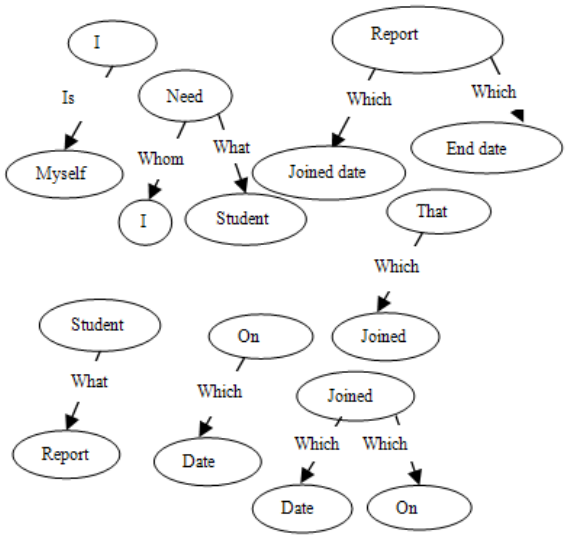

Fig. 3: Sample object-net database

Procedure #3: As per procedur#2 result, the related object or domain of sentence identified (i.e. as per example student), in Object-Net database search for the particular domain which is identified in procedur#2, from the identified object co-relate and identify meaning of the remaining words in sentence. Consider the network exist in object-net database as in Fig. 3.

While forming the two pairs one with the word previous to the current word and one next to the current word, for our example we will be arrived to 7 pairs as in below Table 1, the last column shows that understanding.

Procedure #4: From the procedure#3 we come to know that "need" is the action it required for "whom" is "I", "what" required is "student". From the student node "what" required is "report". But "report" is ambiguous word in English it is having many meaning and also by directly correlating words existing object-net is not giving correct path for the pairs "report + that" and "joined + on", as Date is already disambiguated and while considering pervious nodes it gives the meaning like "on" which is some date (ie. 04 November 2010). By node with connection distance of zero we will be arrived into the below mentioned paths:

- I->need
- I->need->the student
- I->need->the student->report
- On
- On->04 November 2010
- Joined-> on->04 November 2010
- That->Joined->on->04 November 2010

Procedure #5: The word "report" was not clear still Procedur#4, now the report is clear like on "join date" some report is required. The ambiguous word "report" semantically connected with other part of the sentence in three ways as mentioned below:

- Report->joined->04 November 2010
- Report->joined->on->04 November 2010
- Report->that->Joined->on->04 November 2010.

Here the path 2 and 3 are already occurred in Procedur#4 but path 3 is bigger than path 2, so this path is considered and now it is clear that report of joined date is required.

Procedure #6: From the procedure#5, the "need" node is connected to "student" node. "student" node is connected to "report" , "report" is connected to "joined date" and it is connected to "date", from this we have

form a semantic network which gives the meaning as "need" is the action required by "I" and what required is "student", from "student" what required are report and which report is "join date" report.

The Fig. 4 shows the semantic relation path which gives meaning of the sentence.

Analytical and synthesis capability in object-net database: The example sentence "I need the student report that joined on 04 November 2010" can be written in many as mentioned below to reference same meaning as above sentence says.

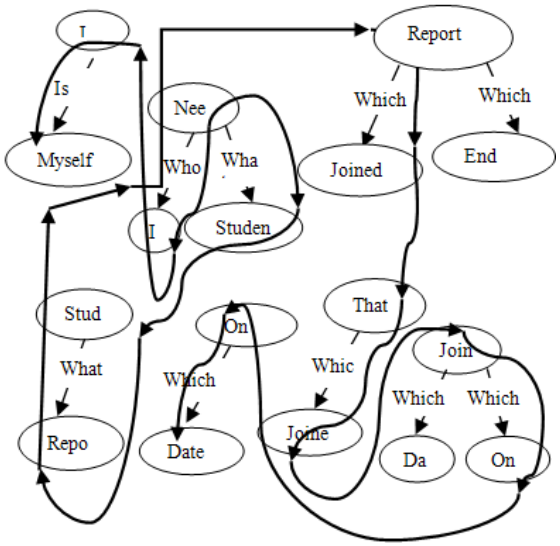

Fig. 4: Forming a semantic network in Object-Net Database

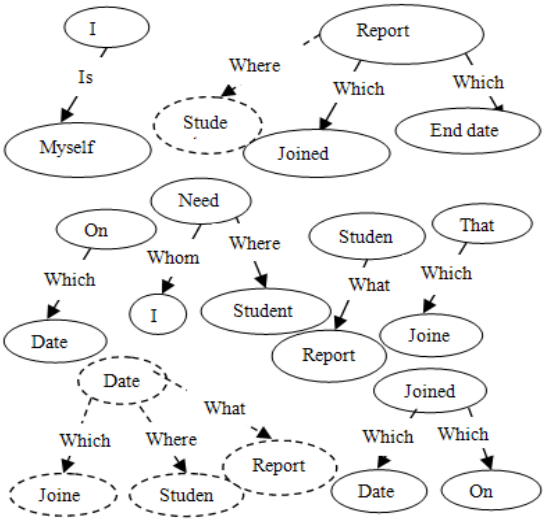

Fig. 5: Updating the active memory

The possible ways are:

- I need the student report that joined on 04 November 2010
- Need student report joined on 04 November 2010
- Report of student joined on 04 November 2010
- Student report joined on 04 November 2010
- On 04 November 2010 joined student report
- On 04 November 2010 joined student
- Joined on 04 November 2010 student report
- Joined on 04 November 2010 student
- Joined student report on 04 November 2010

The above mentioned sentences are giving same meaning as sentence#1, even though the sentences are not in corrected grammatical. But as a human can understand that meaning of all above sentence as "student report is required who are all joined on 04 November 2010". So similarly we have to make sure that our proposed algorithm is also capable understanding the meaning of sentence as human.

For example the above sentence # 3 "Report of student joined on 04 November 2010", in existing trained Object-Net network does not have direct relation from report student but already the "what" relation were existing so it makes the new understanding link between "Report" and "student" with relation of "what". Similarly consider the above sentence#5 "On 04 November 2010 joined student report", this sentence starts with a date and it does not have action part like a action verb "need", in existing Object-Net doesn't have any of node starts with "Date" but there is a "Which relationship exists between "Joined" and "Date" so system creates a new node as "Date" to "Joined" with relation of "Which", next for student report there are two relationship exist one is from "Report" and another one from "student" node, now it creates two relation from newly created date "Date" node to "Student" and "Report" with relation of "Where" and "What" respectively. The Fig. 5 show the updated Object-Net database which will be used for future purpose. So the system analyses and keeps updating its database memory there comes the system learning capability. If some words occurred in input text which is not exist in Object-Net database and also system is not able to resolve it internally then it will ask a master to train the relational network there come the human master into picture in order to correct and update the database.

RESULTS

Object-net approach for database extraction: We illustrate here the Object-Net disambiguation

algorithm with the help of previous example "I need the student report that joined on 04 November 2010". The system identifies the data meaning of the sentence and what is the command and what is action that user is expecting from the system. After identifying the meaning of the sentence, it maps the action to be done along with the trained internal actual database structure so that it can produce exact the SQL query for the input sentence or requirement.

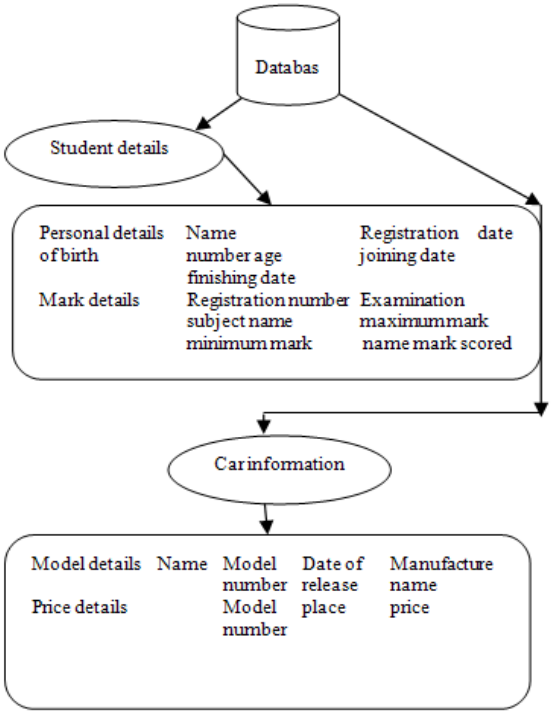

Fig. 6: Actual database information for mapping

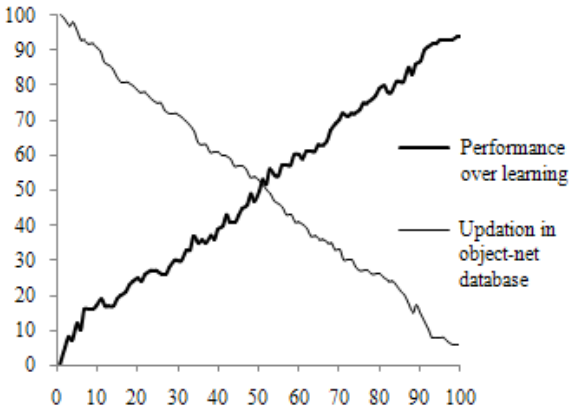

Fig. 7: Accuracy and number trained network

Object-net approach for business reporting: As per the above example ("I need the student report that joined on 04 November 2010") the system identifies the data meaning of the sentence and it produces the SQL query for the input sentence or requirement and executes the query by database engine and gets data in the form of table. If the input sentence says data to be projected in the form of graph such as "I need the number of student joined in between 20 January 2010 to 20 November 2010 and report it in terms graph where number of student in Y-axis and date in X-axis". Then the system will understand meaning of the input sentence and produces the SQL query , execute it in database engine, get the data from database and project it in terms of graphs as per its understanding from the input sentence.

The Fig. 6 shows that "student details" and "car information" databases are exist in a database; this mapping information is shared or trained to our system so that our system knows about where to fetch and which are to be fetched for a given sentence.

DISCUSSION

Performance of word sense disambiguation based on object-net database: The object-net data base consists of set of trained entity network along with their meaningful representation with their action/behavior/property. The performance of our word sense disambiguation algorithm mentioned as above from procedure 1-7 is mainly based on how many trained networks exist in Object-net database. If number of network data are high then number of hit ratio or number of occurrence of word in input text and trained network is high so it helps our algorithm to fetch correct object on which the input sentence is written and what is action or purpose of the sentence in order to give good accuracy on ambiguous words and sentence. When the number of trained network data of words in object-net database is less then number of hit ratio or number of occurrence of word in input text in trained network words is less so the active memory model of object-net database requires the help from master to train the non-trained words into database. Our algorithm will not come to the accurate result to user. The Fig. 7 plots the graph between accuracy of the result of our algorithm versus number trained network word exist in object-net database and the learning update required of object-net database in active memory model.

CONCLUSION

The algorithm identifies the meaning of sentence like human brain. It disambiguates ambiguous words based on object on which sentence is written as in above example the word “report” is ambiguous word but is giving clear meaning based on student object as it requires student report who have joined on 04th November 2010. In future we can train our Object-net data base to other object or domains where intelligent human-computer interaction is required. And also from understanding of natural text meaning to the actual database query generation process can be implemented for accessing data from user database as per the user requirement.

REFERENCES

Abney, S., 2004. Understanding the Yarowsky Algorithm. Computational Linguistics, vol.30, no.3, pp: 365-395. DOI: 10.1162/0891201041850876

Aziz, A.A., M.Y.M. Saman and M.P. Hamzah, 2011. Using metadata analysis and base analysis techniques in data qualities framework for data warehouses. Am. J. Econ. Bus. Admin., 3: 112-119. DOI: 10.3844/ajebasp.2011.112.119

Burges, C.J.C., 1998. A Tutorial on support vector machines for pattern recognition. Knowl. Discovery Data Min., 2: 121-167. DOI: 10.1023/A:1009715923555

Carpuat, M. and D. Wu, 2005. Word sense disambiguation vs. statistical machine translation, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, (ACL '05), Association for Computational Linguistics Stroudsburg, PA, USA, pp: 387-394. DOI: 10.3115/1219840.1219888

Chen, J., J. Yin, A.K.H. Tung and B. Liu, 2004. Discovering web usage patterns by mining cross transaction association rules. Proceedings of International Conference on Machine Learning and Cybernetics, Aug. 26-29, IEEE Xplore Press, USA, pp: 2655-2660. DOI: 10.1109/ICMLC.2004.1378232

Diab, M. and P. Resnik, 2002. An unsupervised method for word sense tagging using parallel corpora. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (ACL '02), Association for Computational Linguistics Stroudsburg, PA, USA, pp: 255-262. DOI: 10.3115/1073083.1073126

Mokhtar, S., J. Chanod and C. Roux, 2002. Robustness beyond shallowness: Incremental deep parsing. J. Natural Language 8: 121-144. DOI: 10.1017/S1351324902002887

Rajaraman, K. and A.H. Tan, 2001. Topic detection, tracking, and trend analysis using self-organizing neural networks. Adv. Knowl. Disc. Data Mining, 2035: 102-107. DOI: 10.1007/3-540-45357-1_13

Shaalan, K., A. Rafea, A.A. Mmonem and H. Baraka, 2004. Machine translation of English noun phrases into Arabic. Int. J. Comput. Proc. Orient. Languages, 17: 121-134. DOI: 10.1142/S021942790400105X

Tiun, S., R. Abdullah and T.E. Kong, 2010. Automatic topic identification using ontology hierarchy. Comput. Linguist. Intell. Text Proc., 2004: 444-453. DOI: 10.1007/3-540-44686-9_43