Automatic Discovery of Semantic Relations using MindNet

Zareen Syed[1], Evelyne Viegas[2], Savas Parastatidis[2]

[1]University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250

[2]Microsoft Research
One Microsoft Way
Redmond, WA 98052

E-mail: zarsyed1@umbc.edu, evelynev@microsoft.com, savasp@microsoft.com

Abstract

Information extraction deals with extracting entities (such as people,organizations or locations) and named relations between entities (such as "People born-in Country") from text documents. An important challenge in information extraction is the labeling of training data which is usually done manually and is therefore very laborious and in certain cases impractical. This paper introduces a new "model" to extract semantic relations fully automatically from text using the Encarta encyclopedia and lexical-semantic relations discovered by MindNet. MindNet is a lexical knowledge base that can be constructed fully automatically from a given text corpus without any human intervention. Encarta articles are categorized and linked to related articles by experts. We demonstrate how the structured data available in Encarta and the lexical semantic relations between words in MindNet can be used to enrich MindNet with semantic relations between entities. With a slight trade off of accuracy a semantically enriched MindNet can be used to extract relations from a text corpus without any human intervention.

1. Introduction

In this paper we present a new model to extract semantic relations fully automatically from text. For achieving this goal we rely on MindNet (Richardson, 1997) which provides a methodology for discovering lexical-semantic relations between words. An attractive feature of MindNet is that it can be built fully automatically from text without any human intervention. However, currently MindNet is restricted to finding lexical-semantic relations between words only. Enriching MindNet with semantic relations between entities would make it possible to extract semantic relations fully automatically from a text corpus. We employ a machine learning approach for enriching MindNet with semantic relations. Machine Learning approaches require labeled training data as input. Generating training data manually for all possible entities and relations is not feasible, therefore, we demonstrate how the structured data and content available in an online Encyclopedia, i.e. Encarta (http://encarta.msn.com), can be exploited for generating training data automatically. We first developed our approach for simpler tasks like Entity Classification and Clustering using MindNet and then designed our approach for Relation Extraction task. Entity Clustering and Classification are also often used as sub-tasks for Relation Extraction in order to constrain the relation arguments. This is the first time where the MindNet resource has been employed and evaluated for Information Extraction tasks.

In the next section we give a brief introduction to MindNet and then discuss our approach for entity classification, entity clustering and relation extraction tasks followed by evaluation and conclusions.

2. MindNet

MindNet is a knowledge representation methodology that uses a broad-coverage parser to build semantic networks from dictionaries, encyclopedias, and free text. MindNets are produced by a fully automatic process that takes the input text, sentence-breaks it, parses each sentence to build a semantic dependency graph (Logical Form), aggregates these individual graphs into a single large graph, and then assigns probabilistic weights to sub-graphs based on their frequency in the corpus as a whole. The process also encompasses a number of mechanisms for searching, sorting, and measuring the similarity of paths in a MindNet (Richardson, 1997). MindNet has different lexical semantic relations built between words and can be queried using a single word (to find out how different words are related to it) or a pair of words (to find out the different lexical-semantic paths between the input words). MindNet that has been built from dictionaries can be queried through a web interface (Vanderwende et al., 2005). Through MindNet, words are connected to other words within a sentence, across sentences and even across documents. For example, if the word "car" occurs in a particular sentence then it will be connected to words in the same sentence and also to words in other sentences present anywhere in the corpus wherever the word "car" is present. In this way one is able to retrieve how a particular word is related to other words in sentences across documents in a corpus. Unlike WordNet, MindNet is a methodology; MindNet can be created fully automatically from text.

3. Approach

There are two main approaches to Information Extraction systems: a knowledge engineering approach, which re-

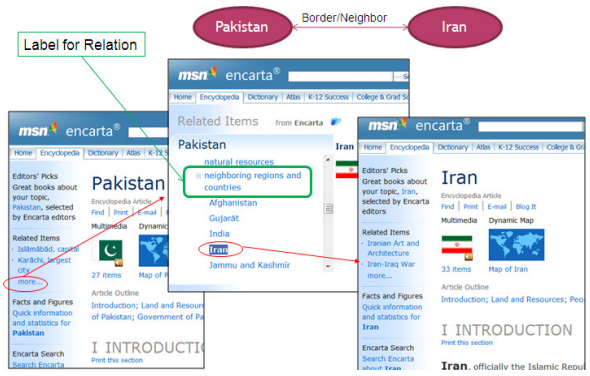

Figure 1: Relation Extraction using "Related Items" in Encarta

quires grammars to be hand crafted to express the rules for the system, a quite laborious process; an automatic training approach which requires the hand annotation of training data by a domain expert, and where sufficient volume of training data is required in order to get reasonable accuracy. Our approach focuses on generating training data automatically by exploiting the rich structure and content of an online encyclopedia Encarta. We also discuss in detail how we employ MindNet for feature extraction as well as for enriching MindNet itself with semantic relations between entities thus enabling fully automatic relation extraction from a text corpus.

3.1 Automatic Generation of Training Data

We exploited the rich structure and content of Encarta in order to generate training data automatically. Below we discuss our approach for generating training data for Entity Clustering and Classification task and for Relation Extraction using Encarta.

3.1.1 Training Data for Entity Classification and Clustering

For Entity Classification and Clustering tasks we used the associated categories with Encarta articles on entities in order to automatically label them with an Entity Class. For testing our approach we randomly selected 85 country and 85 city articles from Encarta under the category "Countries" and "World Cities, Towns & Villages" respectively. We automatically labeled the training data as "City" and "Country" and generated two feature sets, one feature set consisted of lemma from full text of articles and the second feature set consisted of first 10 nouns in articles on entities.

3.1.2 Training Data for Relation Extraction

For Relation Extraction task, we exploited the "Related Items" section in Encarta. Encarta articles are linked manually to related articles by experts through the "Related Items" section. The links are grouped under a hierarchy of labels up to three levels. We have exploited this rich semantic structure present in Encarta for automatically generating semantic classifiers for relation extraction (Figure 1). For example, there is a link from the article on "Pakistan" to the article on "Iran", and the link is grouped with other similar links under the label of "Neighboring Regions and Countries" in the "Related Items" section in Pakistan. This information can be used to identify that a relationship exists between the entities "Pakistan" and "Iran". To find which kind of relationship exists between these two entities, we can use the existing label ("Neighboring Regions and Countries") under which this link is grouped in the hierarchy in "Related Items" section. Further in this paper, we demonstrate how the training data can be generated automatically from Encarta by considering an example relation between countries i.e. "border/neighbor" relation between them. The approach that we used to label the training data automatically using Encarta is described in detail below.

For classifying the border relation between entities we generated the training data automatically. For that we generated the positive examples using the information present in Encarta "Related Items". We generated the negative examples from the positive examples (Figure 2). The approach we adopted is described in detail below.

For generating the positive pairs we selected specific labels in "Related Items" and further selected links grouped under those labels. The details are as follows:

Related Item Label Selection: Since the "Related Items" are listed manually by experts and do not follow a rigid assignment scheme, we can find a variety of labels for the same kind of related items, for example, for the border relation we selected the following two labels in "Related Items" (1) "neighboring countries" (2) "neighboring regions and countries".

Related Item Link Selection: The "Related Items" might have links to different kinds of Entities grouped under the same label. For example, the "Foreign Relations" related item has links to Afghanistan (Location) and Bandung Conference (Event).

Such links could be handled by Entity Classification to select the appropriate links for the relation under consideration. For example, if we define the "Foreign Relation" relation to be a relation between two locations/countries

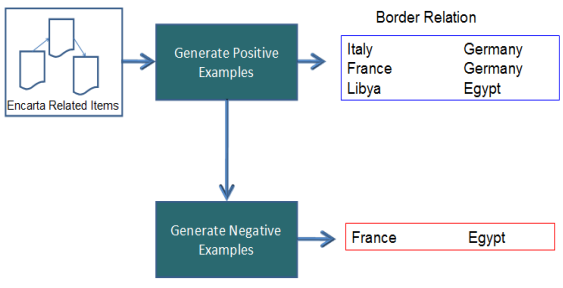

Figure 2: Automatic Generation of Positive and Negative examples for Relation Extraction using Encarta "Related Items"

then we can use Entity Classification to select such links from "Related Items". This step was also needed for the "border" relation between countries to ignore links to regions that are not countries and are defined under the "neighboring regions and countries" label. Another challenge in using the "Related Items" is that there are links that point to different sections of an article, this can add more complications. For example, the article on "Pakistan" has a link labeled as "Universities", one would expect to find a link to an article on some University however, it is linked to a section of the article on the city "Islamabad" which mentions the university in that city. Such links are difficult to interpret automatically. To avoid such difficult links we only select links that are to full articles rather than sections of articles.

After label selection and link selection we had 138 neighboring countries pairs from Encarta. We further performed some filtering steps to filter out symmetric pairs; names which were not countries such as United_States_(People); names that were not available as a single word in Encarta MindNet such as Bosnia_and_Herzogovina. We also reduced the multiword names of countries to single word names for example, we removed the phrase "Federal Republic of" that occurred in several country names. After the filtering steps we were left with 113 entity pairs that served as positive examples for classification.

We generated equal number of negative pairs i.e., 113 randomly using the Positive Pairs, to create each negative pair we randomly picked the first member of a positive pair, randomly picked second member of a positive pair and paired them together. We skipped all those pairs which existed in positive examples and which were symmetric to existing pairs and also manually verified that the pairs created were negative examples. Since the positive pairs list is not exhaustive we had to add this manual step. However, this step can be avoided completely by using classification algorithms that can work with positive examples only (Gang et al., 2007). However, since such algorithms generally find negative examples heuristically, the tradeoff will be slight reduction in accuracy, therefore the decision of selecting the appropriate classifier will depend on the level of accuracy required for the particular application and also the feasibility of applying the manual verification step.

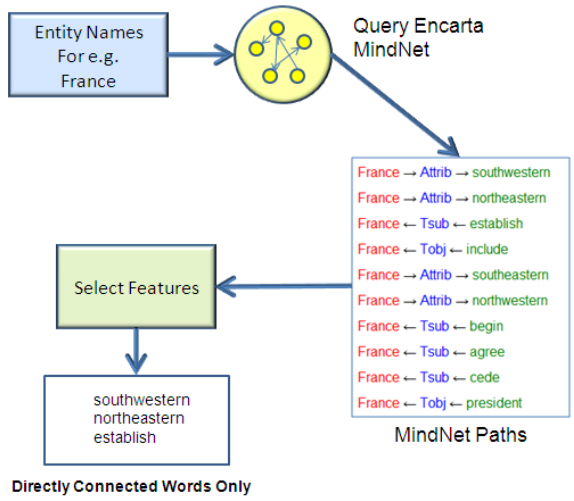

Figure 3. Feature Extraction for Entity Classification and Clustering using MindNet

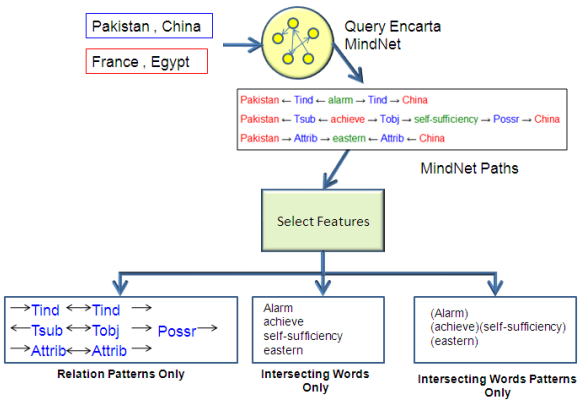

Figure 4. Method for Extracting "Relation Patterns Only", "Intersecting Words Only" and "Intersecting Words Patterns Only" Feature sets from MindNet for Relation Extraction.

3.2 Feature Set Construction using MindNet

For Entity Classification and Clustering tasks we queried MindNet built from Encarta using the Entity Names. We extracted all the directly connected words to the queried entity name in the top 100 paths returned by MindNet (Figure 3). We used term frequencies as weights for features in all Feature Sets i.e., features generated directly from Encarta and features generated using MindNet.

To generate features for Relation Extraction task, we queried MindNet using the pair of entity names and retrieved the paths connecting the pair of words. We generated three kinds of feature sets from the MindNet paths (Figure 4).

MindNet Relation Patterns: This feature set was constructed by extracting the relation patterns between the pair of words and ignoring the intersecting words.
Intersecting Words: This feature set was constructed using the intersecting words in the MindNet paths.
Intersecting Words Patterns: In this case, we extracted the intersecting word patterns from the MindNet Paths.

3.3 Enriching MindNet with Semantic Relations

We have already discussed how Encarta could be exploited to label training data automatically and how feature sets based on lexical-semantic relations could be constructed from MindNet. Among the different feature sets extracted using MindNet the feature set with "Relation Patterns Only" is very interesting as it can be used to enrich MindNet itself with semantic relations. We can use the lexical semantic relations present between words in MindNet to de-

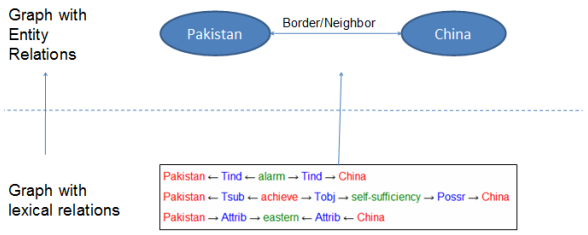

Figure 5: Deriving Semantic Relations between Entities from Lexical-Semantic Relations between words in MindNet

rive higher level semantic relations between Entities (Figure 5). For example "Pakistan" and "China" exist as words in MindNet built from Encarta, we can enrich MindNet relations by deriving the "border" relation between these two entities using automatically labeled data from Encarta and the relation patterns feature set from MindNet. As discussed earlier, MindNet can be built fully automatically from text; therefore enriching MindNet with semantic relations between entities will enable one to automatically generate the entity relations in addition to lexical relations between words. We can also associate the classification accuracy for a particular relation as weight on the edge corresponding to that relation in MindNet, thus building the graph of entity relations fully automatically.

4. Experiments and Results

In this section we discuss the experiments we de-signed for evaluating MindNet generated feature sets for simpler tasks like Entity Clustering and Classification and report their results. We then discuss in detail our approach for relation extraction and conduct its evaluation with the Border/Neighbor relation as an example.

4.1 Named Entity Classification

We created labeled training data for City and Country class as described in the previous section. We used SVM implementation in Weka (Witten and Frank, 2005) and 10 fold cross validation (Kohavi, 1995) to evaluate the accuracy of our automatically generated classifiers and feature sets. The results are shown in Table 2. The first 10 noun lemmas gave the highest accuracy for classification i.e. more than 96%. The accuracy obtained by MindNet and full text features was very close i.e. 92.3% and 91.1% respectively.

	Full Text Lemma Features	MindNet Features	First 10 Noun lemma features
Feature Set Size	41,457	4,446	760
Accuracy (%age)	91.1	92.3	96.4

Table 2. Comparison of different feature sets for Named Entity Classification

Since there was a great difference between the feature set sizes of the three feature sets we repeated the experiments by selecting the top 100 features through feature selection using Information Gain (Witten and Frank, 2005) (Table 3). It was again observed that the first 10 noun lemmas performed best with the accuracy of 97%. However, this time the difference was smaller and it was followed by full text features and then MindNet features.

	Full Text Lemma Features	MindNet Features	First 10 Noun lemma features
Feature Set Size	100	100	100
Accuracy (%age)	96.4	94.1	97.0

Table 3. Comparison of different feature sets for Named Entity Classification after Feature Selection

4.1.1 Evaluation on Wikipedia Articles

We used our automatically constructed classifier built using first 10 noun lemma features from Encarta on articles taken from Wikipedia. We randomly selected 100 city articles and 100 country articles from Wikipedia. We compared the accuracy obtained for Wikipedia Articles with Encarta Articles themselves (Table 4).

Test Sets	Accuracy
Encarta Articles	94.5%
Wikipedia Articles	88.5%

Table 4. Classification of Wikipedia and Encarta Articles using Classifiers constructed and trained on Encarta

The accuracy for Encarta articles is higher since our classifier was trained on Encarta. However, the accuracy for Wikipedia articles is also reasonably high i.e. 88.5% which indicates that classifiers constructed automatically from Encarta can be used on another similar resource such as Wikipedia with reasonably high accuracy.

4.2 Named Entity Clustering

For Named Entity Clustering we used the same training data as for Named Entity Classification however, this time we were interested in evaluating our feature set with respect to Entity Clustering task. Our goal was to evaluate and compare MindNet generated feature set with text based features to see which feature set gives more accurate cluster assignments. We used the K-Means algorithm with varying number of k (number of clusters) and evaluated the quality of our clusters using F-Measure for clustering (Toral and Munoz, 2006). The results are shown in Figure 6. It was observed that the cluster quality for full text lemma features was the least and deteriorated with increasing number of k, whereas for MindNet features the cluster quality was almost similar to text features for $k < 7$. However, for $k >= 8$ the cluster quality significantly improved to

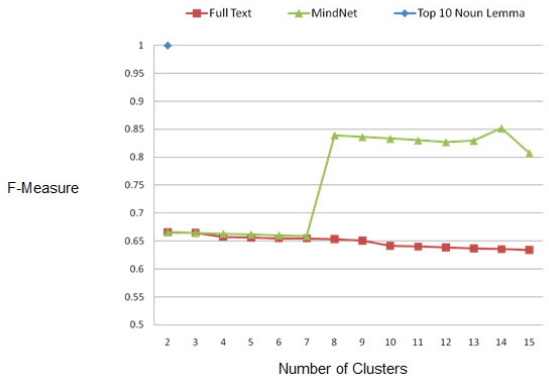

Figure 6: Comparing cluster quality for different feature sets extracted from Text and MindNet

F-Measure > 0.8. After close examination of the data it was observed that there were 6 outliers in the data and each of the 6 outliers was assigned a separate cluster. Among the remaining two clusters, one had majority of city and the other had majority of country class instances. The feature set with first 10 noun lemma gave the best cluster quality with F-measure equal to 1 at k=2 (100% accuracy). We also performed clustering after feature selection but again the first ten noun lemma gave the best results.

4.3 Relation Extraction

For Relation Extraction task we compared the performance of the three feature sets extracted from MindNet for classifying the border relation between countries. The results are shown in table 5. The classification was done using the SVM implementation in Weka (Witten and Frank, 2005) and the accuracy was computed using 10 fold cross validation (Kohavi, 1995). The features were extracted using the top 100 paths returned on querying MindNet for the pair of entities. Increasing the number of paths to 500 reduced the accuracy.

	Relation Patterns Only	**Intersecting Word Patterns Only**	**Intersecting Words Only**
Accuracy (%age)	80.53%	75.22%	78.76%

Table 5. Comparison of the Accuracy obtained using different feature sets extracted from MindNet

Among the three feature sets, the feature set with the relation patterns gave the highest accuracy followed by intersecting word patterns and then intersecting words features. We were most interested in the performance of the relation patterns feature set this feature set would enable enriching MindNet with semantic relations. We were interested in evaluating how the relation patterns would perform in comparison to baseline text features. If their performance is reasonably close to baseline text features and not too low then they can be used to directly enrich MindNet with semantic relations. With this in mind we carried our experiments further and compared the accuracy obtained by simple text features and MindNet Relation Pattern features.

4.3.1 Comparison of MindNet features and Text Features

The simple text features were constructed using the Bag of Words approach (BOW). In MindNet, we selected the top 5 paths. Using > 5 paths reduced the accuracy. We compared both feature sets by classifying the border relation using SVM with 10 fold cross validation (Table 6).

	Text Features	**MindNet (Relations)**
Accuracy (%age)	92%	82%

Table 6. Comparison of Text features and MindNet Features

The accuracy obtained by MindNet features was lower than the text features but still reasonably acceptable i.e. 82% given the fact that this feature set could be used to add semantic relations to MindNet and enable fully automatic relation extraction. Therefore by trading off a little accuracy, this feature set offers fully automated relation extraction which could be an attractive choice to make in certain cases.

5. Discussion

We have developed an approach for extracting semantic relations fully automatically from text and demonstrated and evaluated our approach with the Border/Neighbor relation as an example. We investigated the performance of MindNet generated features for simpler tasks like Entity Clustering and Classification to guide us in crafting our approach for Relation Extraction. In this section we discuss the results of our experiments and propose possible ways to further improve accuracy.

For the Entity Classification and Clustering tasks, the first 10 noun lemma gave the best results. A key reason for this is due to the fact that Encarta is an encyclopedia and the first one or two sentences of articles on entities essentially define the entity. Therefore, using just the features in the first few sentences helps in distinguishing that entity from other types of entities and also avoids the noise that is introduced by considering full text feature sets. For example the articles on countries have some typical words in the first few sentences such as “country”, “monarchy”, “republic” whereas, the articles on cities have words like “city” or “town” in the first few sentences. Similar observations have been reported on Wikipedia articles that the first sentence often defines the Entity and the words in first sentence provide very informative features related to the Entity (Sumida et al.,2008; Nguyen et al., 2007; Wang et al., 2007). However, in the case where we are dealing with non-encyclopedic articles, the top ten noun lemma might not give reasonable performance and exploiting full text features might be the only option, in such a case, MindNet

generated features might be a good option to use instead of full text lemma features for Clustering. We evaluated the automatically generated classifier on Wikipedia articles and observed that it was able to classify with 88.5% accuracy. One might question the need of classifying Wikipedia articles using Encarta Categories because Wikipedia articles already have categories associated with them. However, there are certain complexities and challenges associated with using the Wikipedia Categories directly as the articles within Wikipedia are associated with multiple categories and the categories are not in a strict hierarchy i.e. a single category can have multiple super categories. Secondly, the category hierarchy does not strictly hold the subsumption relationship. For example, one would expect to find geographical regions under the category "Geography", whereas, we find the category "Dances of Middle East" in the hierarchy below Geography through the following path: Geography-> Geography by place -> Regions-> Regions of Asia->Middle East -> Dances of Middle East. Wikipedia Categories have been generally used in research after filtering the irrelevant categories and then ranking the relevant ones through different algorithms (Kliegr, 2008; Syed et al., 2008). In spite of these challenges a very attractive feature of Wikipedia in this regard is its size and coverage which is increasing on regular basis. It has numerous articles on Named Entities, whereas Encarta is limited in its scope. Therefore, to classify Named Entities that are not present in Encarta we can look up Wikipedia articles on those entities and then classify those articles into Encarta Categories through our automatically generated classifiers. This would avoid the challenges related to filtering and ranking Wikipedia categories themselves.

For the Relation Extraction task, we demonstrated how the rich structure of Encarta could be used to automatically label training data for relations already present in Encarta, as shown with the "border" relation. We also compared different kinds of features extracted from MindNet built from Encarta; the feature sets included MindNet Relation Pattern, MindNet Intersecting Word Patterns and MindNet Intersecting words only. The three feature sets gave almost similar performance. However, we were most interested in the MindNet Relation Pattern feature set, as it could be used to advance the MindNet methodology to fully automatically identify semantic relations between entities. To further evaluate the performance we compared it with text features for Relation Extraction. Even though text features gave better accuracy, the accuracy given by MindNet features was also reasonable, with over 80% accuracy. This reduced accuracy trade-off might be a very practical option for tasks involving fully automated procedures. More work can be done to improve the accuracy of MindNet relations such as focusing on specific lexical relations in MindNet, handling word sense disambiguation and using a MindNet constructed from a domain specific corpus.

6. Related Work

With respect to Entity Classification, Wikipedia has been used for real time hypernym discovery in the work proposed by Kliegr (2008). He employed an unsupervised algorithm which expressed Entities and Classes as WorldNet synsets and Wikipedia was employed for real time hypernym discovery to map uncommon entities to WordNet. Toral and Munoz (2006) proposed to automatically build and maintain gazetteers for Named Entities by analyzing the entries of Wikipedia with the aid of a noun hierarchy from WordNet. For every noun in the first sentence of the Wikipedia article on a NE they follow the hypernymy branch of that synset until the root is reached or the considered synset is reached i.e. Person, Location or Organization. In the later case they consider the entity as belonging to that synset or class. They also apply a weighing algorithm and certain heuristics to improve their results. Watanabe et al. (2007) categorized Named Entities in Wikipedia by structuring anchor texts and dependencies between them induced by the HTML structure (for example, a list or table and anchor text of inter-article links) into a graph and trained Graph Based Conditional Random Fields for obtaining models for classifying Named Entities in Wikipedia. Beneti et al. (2006) used Wikipedia Categories for fine grained Named Entity Classifications. They applied several heuristics to filter out unimportant categories from Wikipedia such as "Cleanup from December 2005", "1946_births". After filtering they devised a ranking system to give a higher rank to the more relevant categories.

With respect to Relation Extraction, Wikipedia structure and content has been exploited in various ways for facilitating this task. Wang et. al. (2007) used Wikipedia to generate selectional constraint features for entities in the relations by incorporating the Definition words, Wikipedia Category words, Disambiguation text and relation predicates taking the entity at the subject and object position. Wu and Weld (2008) treated entries in the Wikipedia Infoboxes as attributes and combined that with WordNet to generate an Info-box ontology using machine learning for inferring WordNet mappings and ISA taxonomy. Bloehdorn and Blohm (2006) used Self Organizing Maps for Structured data by incorporating additional hyperlink information from Wikipedia. They applied it for clustering and then for extracting relations between created clusters. Nguyen et al. (2007) presented an approach for relation extraction from Wikipedia by extracting features from subtrees mined from the syntactic structure of text.

Our approach is different as compared to the above mentioned approaches, firstly because we rely on Encarta as our knowledge source which is structurally different from Wikipedia. For example, the category hierarchy in Encarta is restricted to two levels and each article is associated with a distinct category by an expert, on the other hand, Wikipedia category hierarchy is a thesaurus allowing categories to have multiple parents and a single article can be linked to multiple categories at the same time. Moreover, the depth of the category hierarchy in Wikipedia is not restricted and also contains many administrative categories

interlinked with informative categories. Approaches based on Wikipedia category hierarchy require additional steps for filtering and ranking the associated categories to find the most relevant ones (Syed et al., 2008), whereas our approach can be used directly because of the nature of Encarta category structure. Wikipedia has much more coverage than Encarta. However, we have discussed earlier in section 5 how we can handle the case where an Entity is missing in Encarta but present in Wikipedia.

For the task of relation extraction, our approach differs significantly from other approaches in the way we exploit the lexical-semantic relations discovered by MindNet. We cannot compare our work to the work on WordNet because MindNet and WordNet (Miller et al., 1990) are two different resources. For example, WordNet is a static resource built manually whereas MindNet is a methodology; MindNet can be created fully automatically from any text corpus without any human intervention; unlike WordNet, MindNet does not directly disambiguate between discrete senses of words and the distinction and interpretation of different senses is left to the application to handle (Dolan et al., 2000).

Earlier research work related to MindNet focused on determining similarity and inferring relations between words (Richardson, 1997), metaphor interpretation (Dolan, 1995), handling ambiguity in dictionary texts (Vanderwende et al., 1995), automatic analysis and interpretation of Noun Sequences in un-restricted text (Vanderwende et al.,1994), exploiting lexical information for visual processing (Dolan, 1994) and clustering related senses of words (Dolan, 1994). This is the first time where MindNet has been employed and evaluated to generate features for entity classification and relation extraction tasks.

7. Conclusion

In this research, we presented a new model to extract semantic relations fully automatically from text. We developed an approach to enrich MindNet with semantic relations by generating training data from Encarta automatically. We evaluated our approach through our experiments with a sample relation; the same approach could be used to add other semantic relations between entities in MindNet. The proposed methodology does not require hand labeling of data while delivering accuracy above 80% as seen on encyclopedic articles. This is the first study in which the MindNet resource has been employed and evaluated for information extraction tasks. MindNet holds the promise of being a lexical-semantic resource for the community of researchers who could use it for information extraction or semantic relation discovery. With a slight trade off of accuracy a semantically enriched MindNet can be used to extract relations from a text corpus without any human intervention.

8. Future Work

In the future, we plan to exploit the lexical semantic graph of words in MindNet. For this research we have used the top N paths returned by Querying MindNet for feature extraction which are ranked based on average vertex probability (Richardson, 1997). Presently, we are only able to query using a single word or a pair of words. A multi word query might be able to return more relevant paths by querying for context words along with the pair of entities, for finding the relation between entities. Graph algorithms could be employed for selecting the most relevant paths given the multiple words. For example we can use the initial set of multiple words (entities and context words) and run the Network Spreading Activation algorithm to find important nodes. The relevant paths could be those which pass through the highly activated nodes. More work can be done in this direction.

Another promising application of MindNet is Machine Translation. MindNet can be used to extract lexical semantic relations from corpora belonging to different languages. Since the lexical-semantic relations do not tend to change significantly across different languages it would be interesting to compare relation extraction performance using the lexical-semantic relations feature set across different languages in the future.

9. References

Aspasia Beneti, Wael Hammoumi, Eric Hielscher, Martin Mueller and David Persons. 2006. Automatic generation of fine-grained named entity classifications. Technical report, University of Amsterdam, February 2006.

Asuka Sumida and Kentaro Torisawa. 2008. Hacking Wikipedia for Hyponymy Relation Acquisition. In: Proc. of IJCNLP, 2008.

Antonio Toral and Rafael Munoz. 2006.A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In Proc. Of the workshop on New Text at the 11th EACL'06, Trento, Italy, 2006.

David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes 30(1), 3-26, 2007.

Dat P. T. Nguyen, Yutaka Matsuo and Mitsuru Ishizuka. 2007. Subtree Mining for Relation Extraction from Wikipedia. In Proc. of NAACL/HLT 2007, 2007.

Fei Wu and Daniel S. Weld. 2008. Automatically refining the Wikipedia Infobox Ontology. In: Proceedings of the Seventeenth International World Wide Web Conference (WWW-2008), Beijing, China, April 2008.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. 1990. Wordnet: An on-line lexical database. International Journal of Lexicography, 3(4):235-312. 1990.

Gang Wang , Huajie Zhang , Haofen Wang and Yong Yu. 2007. Enhancing Relation Extraction by Eliciting Selectional Constraint Features from Wikipedia. In: NLDB 2007, 2007.

Ian H. Witten and Eibe Frank. 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Lucy Vanderwende, Gary Kacmarcik, Hisami Suzuki and Arul Menezes. 2005. MindNet: An Automatically-Created Lexical Resource In Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, Vancouver, British Columbia, Canada, October 2005.

Lucy Vanderwende. 1994. Algorithm for Automatic Interpretation of Noun Sequences. In: Proc. 15th Int'l. Conf. Computational Linguistics, ACL, Morristown, NJ, 782-788, 1994.

Lucy Vanderwende. 1995. Ambiguity in the Acquisition of Lexical Information.In: Proc. of the AAAI 1995 Spring Symposium Series, symposium on representation and acquisition of lexical knowledge, 174-179, 1995.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12): 1137–1143, (Morgan Kaufmann, San Mateo), 1995.

Stephan Bloehdorn and Sebastian Blohm. 2006. A Self Organizing Map for Relation Extraction from Wikipedia using Structured Data Representations. In: Proc. of the International Workshop on Intelligent Information Access (IIIA-2006), Helsinki, Finland, 2006.

Stephen D. Richardson. 1997. Determining Similarity and Inferring Relations in a Lexical Knowledge Base. PhD. dissertation, City University of New York, 1997.

Tomas Kliegr. 2008. Unsupervised Entity Classification with Wikipedia and WordNet. Proc. of the 2nd K-Space PhD Jamboree Workshop, TELECOM Paris-Tech, Paris, France, July 25th 2008.

William Dolan, Lucy Vanderwende, and Stephen D. Richardson. 2000. Polysemy in a Broad-Coverage Natural Language Processing System. In Polysemy: Theoreti-cal and Computational Approaches, Ravin, Y. and Leacock, C., eds., Oxford University Press.178-204. 2000.

William Dolan. 1994. Exploiting Lexical Information for Visual Processing. In Proceedings of AAI-94 Workshop on the Integration of Natural Language and Vision Processing, Seattle, Washington, 185-188, 1994.

William Dolan. 1995. Metaphor as an Emergent Property of Machine-Readable Dictionaries. In: Proc. of the AAAI 1995 Spring Symposium Series, 27-32, 1995.

William Dolan. 1994. Word Sense Ambiguation: Clustering Related Senses. In: Proc. 15th Int'l. Conf. Computational Linguistics, ACL, Morristown, N.J., 712-716, 1994.

Gang Wang, Yong Yu and Haiping Zhu. 2007. PORE: Positive-Only Relation Extraction from Wikipedia Text. The Semantic Web: 6th International Semantic Web Conference, Springer-Verlag New York Inc, 2007.

Yotaro Watanabe, Masayuki Asahara and Yuji Matsumoto. 2007. A Graph-based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields. In Proc. of EMNLP-CoNLL 2007, 649-657, 2007.

Zareen Syed, Tim Finin and Anupam Joshi. 2008. Wikipedia as an Ontology for Describing Documents. Proc. of the International Conference on Weblogs and Social Media (ICWSM-08).