International Journal of Future Generation Communication and Networking
Vol. 13, No. 2, 2020 pp.1606-1613

A Comprehensive Study on Different Methodologies and Features in Synonym Identification for Language Processing

Vinutha C K and Jagadish S Kallimani
Department of Computer Science and Engineering
M S Ramaiah Institute of Technology (MSRIT)
Bangalore, India
vinuthashobha5@gmail.com and jagadish.k@msrit.edu

Abstract

Choosing the wrong word may convey unintended connotations, meanings or attitudes in a machine translation or natural language generation system. Identifying near synonyms like near, closer, almost and close by -- words that share the same core meaning but differ in their nuances— can be made only if knowledge about their differences is available. Identifying such synonym of a word/entity in the given context is a critical and trending concept in Natural Language Processing (NLP) which has immense application in various fields like word sense disambiguation, text summarization, document retrieval etc. There are wide variety of technique and methodologies have been proposed for identification of synonyms in a given context by utilizing various dataset or corpus. Identifying synonym in a given context has become more trending topic in a research field of NLP. In this paper we try to discuss various technique and works that has been used to solve automatic synonyms retrieval problem.

Keywords:*Distributional Semantic Model (DSM), Pattern-based Model, Supervised Learning, Hard Synonyms, Latent Semantic Analysis (LSA), Random Indexing (RI), Knowledge Base*

1. Introduction

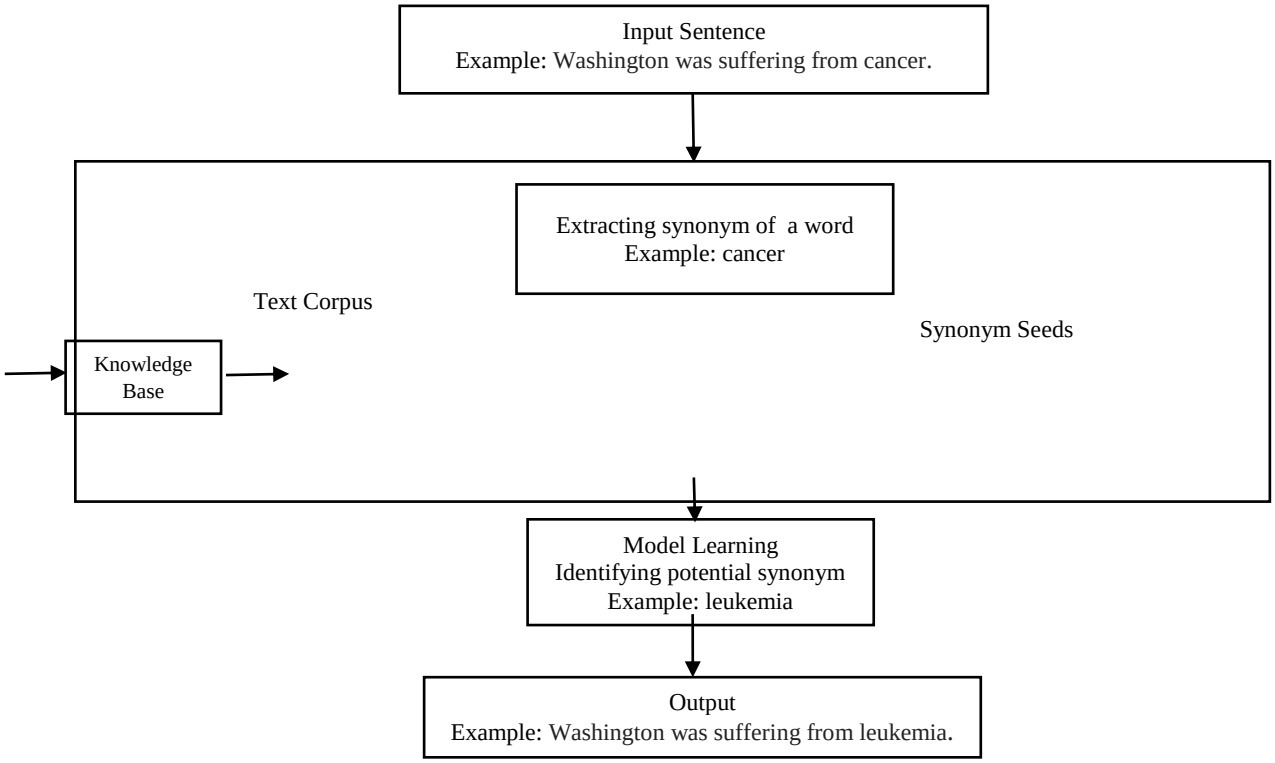
Synonyms represent the semantic relation between the words in a language. The synonyms can be a substitutable word in a given context without changing the meaning (absolute synonyms) and the sense of the context (sense synonyms). Identifying the synonym of a word in the context is a trivial task for human. But the same trivial task is difficult to achieve by the machines. Only with rigorous training and large appropriate knowledge base, machine can achieve this synonym identification task.

The task of identifying semantically similar terms and the semantic relation between the word pair has received the lot of attention and many methodologies have been proposed for Semantic Similarity Measurement (SSM). Methods of semantic similarity measurement can be categorized as knowledge base and distributional methods. Earlier literatures make use of manually constructed resources like Wikipedia or WordNet (Miller 1995) for SSM. While resources like WordNet provides limited information to the machine and, it is not available for all the languages. Distributional Semantic Models (DSMs) are the alternative for the knowledge base method. DSMs model work on Distributional hypothesis i.e. two words are considered similar if they share common context. For example, some words like "USA" and "United States" often mentioned in similar context and they are synonym of the country USA. Most communally used DSMs are Latent Semantic Analysis (LSA) and Random Indexing (RI) (Sahlgreen, 2005). Recent works combine knowledge base like Wikipedia with distributional method (Mihalcea and Hassan 2011). Advantage of DSM over knowledge base is DSM requires no etymological knowledge other than corpus. On contradictory part disadvantage of DSM over knowledge base is DSMs can't identify different type of synonyms and could not able to different sense of polygamy words.

Another often used SSM method is pattern matching. Semantic similarity measurement is done based on the observed pattern in context. For example, consider the sentence "United State of America is also called as America" by which we can identify the semantic relation between "United State of America" and "America". As DSMs uses distributional feature, pattern matching method uses identified patter of a sentence as feature to identify the synonyms. Some work (MendQU, XiangRen, JiaweiHan – 2017) combine DSMs and pattern matching. Resent works make use of machine learning concepts like supervised learning and Deep neural network for classifying the word pair as synonym and not.

ISSN: 2233-7857 IJFGCN

Below figure shows the general view of synonym identification task. The input may be a word, text or document. As processing, key word as to be extracted from the input (text or document) and appropriate methodology must be applied in order to identify synonym of the extracted key word. The identified synonyms must replace the key word as the output.



ID	Sentence
1	**Washington** is the first president of the **USA**.
2	**Washington** is a wealthiest country.
3	**Cancer** involve abnormal cell growth.
4	The exact cause of **leukemia** is unknow.

Entity	Synonym
USA	USA, America
cancer	Cancer, cancers, leukemia

Figure 1 : General View of Synonym Identification Task

In this paper, we try to identify the different techniques used to solve the synonyms identification task. The organization of this paper is as follows. Current and future scope are discussed in section 2. Objectives of this work are discussed in section 3. Detailed survey on the considered work is discussed in section 4. Benefits and conclusion are discussed in section 5 and 6 respectively.

2. Current and Future Scope

Synonyms identification has been used in several NLP application, one of the remarkable work is automatic synonym detection or extraction (Wang and Hirst, 2019; Wang et al., 2018; Castelli, 2018), In turn this application has a great advantage in tasks which includes information retrieval, machine translation, spelling correction, speech recognition, text categorization (Hairst, 2016). Based on word alignment of parallel corpora a multilingual approach has (Vana der Palas et al., 2019) higher

ISSN: 2233-7857 IJFGCN

performance scores for the task of synonym identification than the monolingual approach. Other work on semantic distance between words and concepts (Mohammad, 2017) emphasize on the benefits of multilingual over the monolingual treatment.

Benefits of synonyms acquition can be extended and used in various applications like sentence rephrasing, source code parsing, synonyms identification in medical term, cryptography (for encrypting the information) etc. Performance of the system for synonym identification can be improved by combining the advantages of different methodologies. The performance measures of synonym procurement can also be increased by improving the knowledge base that has been used for training the machine.

3. Objectives of the Study

Assessing and understanding the inner meaning of sentences is a trivial task for human. To achieve same task via machine is difficult. It requires enormous dataset or knowledge base and different methodologies to train the machine. Identification of synonym in a given context is similar kind of issue which require good amount of dataset and technique. In this paper we tried to give a brief description about the different words that has been done in automatic synonym identification.
Objectives of this work are:

- To identify the different methodology used to solve the automatic synonym identification problem.
- To identify the advantage and disadvantage of each methodology.
- To gain the insight of, how different literature or work combine the different methodology to get better accuracy.
- To compare different methodology with respect to their performance.

4. Related Work

This section provides brief description of different works and approaches used for synonym identification tasks. In this section we also try to analyze the performance and result of each work.

Table 1: Detailed Table of Synonym Identification Tasks

Sl.no	Author	Publication	Year	Title	Methodology	Description	Result	Remark
1	DainaInkpen and Graeme Hirst	Association for Computational Linguistics	2006	Building and Using a Lexical Knowledge Base of Near-Synonym Differences	Unsupervised decision-list algorithm	This work presents a new lexical knowledge (near synonym difference). Unsupervised decision-list algorithm is used to derive the patterns from the special dictionary of synonym difference.	The precision and recall for this work were estimated as 70-80%	They can consider more features for pattern extraction
2	Kaname Kashara and Christopher	Association for Computational Linguistics	2006	Synonym Retrieval Using Word Vectors from Text Data	Thesaurus based and Single value decomposition (SVD)	This work makes use of word vector concept. They build two-dimension word vector from dictionary definitions of words which can be used to calculate degree of	They build a word matrix of high dimension that improves t performance of synonym retrieval.	This work can be enhanced for other NLP applications like word sense disambiguation, information retrieval etc..

ISSN: 2233-7857 IJFGCN

						semantic similarity.		
3	Masato Hagiwara	Association for Computational Linguistics	2008	A Supervised Learning Approach to Automatic Synonym Identification based on Distributional Feature	Distributional Semantic Approach and Pattern based approach	In this paper the synonyms acquisition is viewed as a classification problem. The model will classify the word pair in to synonyms or non-synonyms. They build nearly 5 synonym classifiers. As a corpus New York Times section of English Giga word is considered	Distributional Feature(DFEAT) classifier has as greater performance 95.25% but when this classifier is combined with pattern-based feature i.e. (DFEAT-PAT) classifier the precision has been increased to 95.37%	This paper make use of the supervise learning technique for synonym identification.
4	MladenKaran, Jan Snajder, BojanaDalbeloBasi	Association for Computational Linguistics	2012	Distributional Semantic Approach to Detecting Synonyms in Croatian Language	Latent semantic (LSA)analysis and Random indexing(RI)(Basic models of DSM)	In this paper they build several models using LSA and RI. For knowledge base they make use of large hrWaC corpus . Model has been evaluated on dictionary-based similarity test.	LSA model has the great performance than the RI model. Best accuracy achieved were 68.7%, 68.2%, 61.6% on noun, adjectives and verb , respectively.	This paper took the great advantage of basic models like LSA and RI. The performance of the model can be improved by incorporating additional techniques like WSD
5	Glyn Caon, Mark Truran and Helen Ashman	Proceeding of the first Australasian Web Conference(AWC)	2013	Finding synonyms and other semantically-similar terms from coselection data	Clustering algorithm	This paper makes use of the Co selection concept(selection of the related URL\topic by the user from the result of surfing). They build weighted terms graph and identify the cluster overlap to calculate similarity between the co selected URL. This study also shows that both text and image search can be used to for synonym identification.	This work got good result even with weak parameters. The number of false positive is low especially for traditional text search.	This study can be enhanced further and could able to extract other kind of lexical knowledge.
	Ching-Yun Chang and	Association for Comp	2014	Practical Linguistic Steganography using Contextual	Vertex coding algorithm	This work makes use of synonym substitution as the major transforma-tion in linguistic	This work improves the data embedd-ing capacity	This work experiments the use of NLP concepts for

ISSN: 2233-7857 IJFGCN

6	Stephen Clark	utational Linguistics		Synonym Substitution and a Novel Vertex Coding Method		steganography. They address the two major issue of synonym substitution i.e. words with more than one sense and identifying the synonym of a word with respect the context. They constructed graph where words are represented as vertices, synonyms as edges and unique bits are assign to each word calculated by vertex coding algorithm.		linguistic steganography. This work can be extended to different language.
7	Suntae Kim .DongsunKim	Springer Science +Business Media New York	2015	Automatic identifier inconsistency detection using code dictionary	Based on code dictionary	This work is an attempt to solve the problem of inconsistent identifier in the source code using code dictionary. Code dictionary is build using the API document of popular Java projects by using Natural Language Processor (NLP) parser. They consider three type of inconsistent identifiers (semantic, syntactic, and POS)	This work could able to detect the inconsistent identifier in the software code with 85.4% precision and 83.9% recall.	This word is very useful for the developer to find the inconsistent identifier in their source code and improve the software quality.
8	Tugba YILDIZ ,Banu DIRI and Savas YILDIRM	Association for Computational Linguistics	2016	Turkish synonym identification from multiple resources: monolingual corpus, mono/bilingual online dictionaries, and WordNet	Distributional Semantic Approach	This paper extracted the features of the entity from different resources like monolingual online dictionaries, bilingual online dictionary , WordNet and monolingual Turkish corpus. Machine learning algorithm has been applied to those extracted features to identify the semantic relation between word pair.	Considering all the attributes as a feature set of training data , the success rate is 95.2% and the F-measure for synonym is 81.4% where the false positive rate is 24% and false negative rate is 1.6%.	This work uses variety of features obtained from multiple sources so the model could able to achieve 95.2% success rate. They could also make use of antonym relations as a filter to improve the performance of synonym identification.
	AnaSa	Assoc	2017	Hard Synonym	Distributio	This paper makes use	They calculate	This word is

ISSN: 2233-7857 IJFGCN

9	binaUban	iation for Computational Linguistics		and Applications in Automatic Detection of Synonyms and Machine Translation	nal Semantic Approach	of the concept called hard synonyms(semantic relation between two words that are synonyms in more than one language). They build their own database with four different language. Database has the information like word ,it's translation in other languages ,POS. With the help of the wordtovector and the database they constructed hard synonyms are extracted which are considered as true synonyms	the recall which gives the percentage of the hard synonyms which were conformed as synonyms in the dictionary. For English and French, the recall is 40.32%	potential investigation of the concept hard word and their usage. This work shows how hard synonyms are used for synonym extraction from corpora and to machine translation
10	Meng Qu, Xiang Ren ,Jiawei Han	arXiv: 1706.08186 vl[cs.CL]	2018	Automatic Synonym Discovery with Knowledge Bases	Distributional Semantic Approach and Pattern based approach	This paper present's a frame word called DPE. This frame work is the combination of distributional features based on corpus-level statistics and textual pattern based on local contexts.	This frame work has the better performance compared to PATTY(Pattern based approach)	This paper combines the advantages of both distributional model and pattern-based model.
11	Amir Hazem and Beatrice Daille	Association for Computational Linguistics	2018	Word Embedding Approach for Synonym Extraction of Multi-Word Terms	word-embedding-based approach	This work presents new word embedding approach for automatic synonym retravel of multi-word term(MWT)	Gives better performance compare to baseline algorithm	This work can be extended to synonym of various length
12	Kai Lei, Shangchun Si, Desi Wen and Ying Shen	Association for Computational Linguistics	2019	An Enhanced Computational Feature Selection Method for Medical Synonym Identification via Bilingualism and Multi-Corpus Training	Supervised learning model Support vector machine (SVM)	This work proposed a method to identify the synonym for medical terminology of chines language. They have considered 13 features from both chines and English language and identify those features that are more useful to identify the synonym of medical	This work has achieved 97.37% precision rate, 97.33% F1 score and 96.00% recall rate	This work has achieved adequate result which can be improved by concentrating more on fields like symptoms, drugs and diseases.

ISSN: 2233-7857 IJFGCN

						terminology in chines language.		

5. Benefits

Lexical knowledge from Synonym identification can be used in different applications like Word sense disambiguation (Soroa and Agirre,2009), Automatic thesaurus construction, Finding the similarity between documents (Saric,2012), WordNet acquisition (Broda, 2008), Text summarization (Inui 2003), Expansion of query (Pantel 2009), machine translation and goggle search engine.

6. Conclusion

In this paper, we have identified different methodologies and the way it has been used for automatic synonym identification. We identified methodologies like knowledge base, distributional base, pattern base, supervised learning methodologies etc. In this study we came to know that amalgamating different methodologies gives better result than using discrete techniques. Integrating different methodology help to combine the advantages of each integrated methodologies and helps to overcome the short comes of those methodologies. The study of automatic synonym identification has a greater scope in most of the NLP application. Our study is helpful to know the existing work and different methods to solve automatic synonym acquition.

References

[1] Diana Inkpen and Graeme Hirst, "Building and Using a Lexical Knowledge Base of Near Synonym Differences", *In ACL 2006*

[2] Kaname Kashara and Christopher, "Synonym Retrieval Using Word Vectors from Text Data", *Association for Computational Linguistics* 2006.

[3] Masato Hagiwara, "A Supervised Learning Approach to Automatic Synonym Identification Based on Distributional Features*", In Proceedings of the 14th conference on Computational Linguistics-Volume 2, pages 539–545. Association for Computational Linguistics* 2008.

[4] Mladen Karan and Jan Sanasjder, "Distributional Semantics Approach to detecting synonym in creation language, *In ACL Workshop on Automatic Summarization* 2012.

[5] Ching-Yun Chang and Stephen Clark, "Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method", *In Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)* 2013.

[6] Glyn Caon1, Mark Truran2 and Helen Ashman1 ,"Finding Synonyms And Other Semantically-Similar Terms from Co-Selection Data", *In ACL* 2014.

[7] Suntae Kim · Dongsun Kim,"Automatic Identifier Inconsistency Detection Using Code Dictionary", *In ACL Workshop on Automatic Summarization* 2015.

[8] Tugba YILDIZ1,Banu D_IR_I and Savas YILDIRIM, "Turkish Synonym Identification From Multiple Resources: Monolingual Corpus, Mono/Bilingual Online Dictionaries, And Wordnet", *In ACL-IJCNLP* 2016.

[9] Ana Sabina Uban , "Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation", *In ACL 2017.*

[10] Meng Qu, Xiang Ren and Jiawei Han, "Automatic Synonym Discovery with Knowledge Base", *In Proceedings of the 9th International Conference on Semantic Systems (Semantics)* 2018.

[11] Amir Hazem and Beatrice Daille ," Word Embedding Approach for Synonym Extraction of Multi-Word Terms", *In ACL* 2018.

[12] Kai Lei, Shangchun Si, Desi Wen and Ying Shen, "An Enhanced Computational Feature Selection Method For Medical Synonym Identification via Bilingualism and Multi-Corpus Training" ,*In ACL* 2019.

ISSN: 2233-7857 IJFGCN

ISSN: 2233-7857 IJFGCN