2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT) | 978-1-6654-7224-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/CCiCT56684.2022.00082

From linguistic features to their extractions: Understanding the semantics of a concept

Chandni Maggo[1], Puneet Garg[2]

[1]Assistant Professor, Manav Rachna University, Faridabad, Haryana, India

[2]Assistant Professor, ABES Engineering College, Ghaziabad, U.P. , India

chandnimagoo@yahoo.com[1], puneetgarg.er@gmail.com[2]

Abstract: **Understanding the meaning of a word has always been a challenging task for machines. There are circumstances when even an easy word becomes difficult to understand. This understanding is not limited to its meaning but other criteria like identifying similar words, solving ambiguity, co reference resolution, etc. Semantic features and semantic relations can be identified as the building blocks for the semantic illustration of a concept. Understanding a language by machine requires understanding of words both linguistically and computationally. This paper tries to critically review the semantic features and relations created with their extraction methods.**

Keywords: **Natural language Processing, Production norms, Semantic classes, Semantic features.**

I. INTRODUCTION

Semantics of a concept can be represented majorly through its denotative meaning and connotation [1]. Denotation or conceptual meaning represents the core sense of a concept which is true for everyone whereas connotation or associative meaning is beyond conceptual representation and vary from person to person. Thus, for representing the meaning or context of a concept [2] semantic features and semantic relations play a vital role. Semantic features are the focal point for describing the representation of the concept. Semantic features and relations have also been considered as an approach to many psychological and behavioral aspects like semantic memory, episodic memory [3] which model their use in various applications like conversational agents [4] and machine translations. To represent Semantic information, various norms are constructed which represent each concept with it features. These norms provide researchers with substantial amount of data that is consistently evaluated on significant dimensions. There have been several features based [5],[6] and distributional-based models [7],[8],[9]. These models have proved useful in various natural processing tasks (NLP). Formerly distributional based models have been criticized for feature based despite of representing certain accomplishments. A thorough comparison of both models has been done [9] for researchers' insight. Earlier the feature-based models [10] described a word in binary form like birds can fly while cats cannot whereas modern feature-based models of semantic representations are constructed by involving participants that describe concept's meaning by generating features. For any natural language processing-based tasks like question answering, machine translation, text summarization etc requires huge and reliable dataset for effective performance. Thus, a lot of work has been done in building a dataset of semantic feature norms for huge set of noun and verb concepts. The next section includes the classification of semantic features followed by the issues which can be resolved with semantic features. Further we discuss various semantic features extraction methods .Many researchers have come to definite consensus that Semantic representation can be based on features and relation identification. Both can be used to measure the similarity and relatedness among various concepts.[5],[11],[12] asked the participants to list down the features they consider relevant to describe the meaning of a concept. They concluded that the overlapping features (common features) resulted in determining the similarity between concepts. This approach was semantic feature-based approach. For e.g., crocodile and alligator can be considered similar because they share common features like both are reptiles, have big jaws and belong to same family Crocodylidae etc. Similarity measures through distributional view, instead, can be inferred by the distribution of lexical co-occurrence frequencies of each concept pair across words[13].This approach is based on the similar context and thus is majorly used for finding relatedness among concepts rather similarity, for e.g. crocodile and tears.Often the similarity and relatedness measures are represented through a perceptual and distributional view The perceptual view more often represents the feature based representation while distributional view is corpus based . [14] reviewed that the similarity or the relatedness can be based upon word's concreteness i.e., how the concept is causally related to sensorial experience. Concrete word's meaning can be obtained through features which can be directly experienced for e.g., dog has properties has tail, barks, has four legs etc. Meaning of an

978-1-6654-7224-1/22/$31.00 ©2022 IEEE
DOI 10.1109/CCiCT56684.2022.00082

Abstract word is grounded deep in internal sensory experience and are more valence [15] for e.g., hate. [16] has shown that abstract and concrete concepts have different structural representation. Concrete words provide greater contextual information as compared to abstract words.

II. CLASSIFICATION OF SEMANTIC FEATURES

Many researchers have identified different classes to categorize various features identified for a particular concept. [2] have identified eight classes for noun entities namely specificity, boundedness, animacy, sex and gender, kinship, social status, physical properties, and functional properties. These properties are sufficient to classify any feature of noun entity. For e.g., features of cow entity can be classified under animacy , physical properties and functional properties as shown in (Table I). As far as the verb classes are concerned the largest and the most widely deployed classification in English is Levin's taxonomy[17]. [18] has extended the Levin's verb classes by incorporating 57 novel verb classes which were not introduced by Levin. Some of them are listed as: urge (ask, persuade), allow(allow, permit), admit(include, welcome) etc. They evaluated that these novel classes can be used in many NLP tasks and has a fantastic coverage over the English Lexicon verbs.

TABLE I
EXAMPLE OF CLASSIFICATION OF NOUN ENTITY COW

Animacy	Physical Features	Functional features
Living thing Female	Has tail Has four legs Color white	Gives milk

III. REVIEW OF VARIOUS FEATURE PRODUCTION NORMS

Previous work on semantic feature production norms in English includes databases by [5]. The feature production norms focused on 541 nouns, specifically living and nonliving objects. The production of norms is critical to provide researchers with concepts that can be used in future research. In their experiments participants were asked to list down various features for a given concept based on different properties like physical, functional etc (Table II). [5] work was limited to noun concepts without focusing on ambiguous words. Their work did not manage to collect features of ambiguous words like apple, bank etc. Following his work [11] expanded the corpus set by contributing norms for 456 concepts that included both nouns and verbs. The classification of their norms into feature types showed that living things describe well with sensory features than for non- living concepts. For the latter functional features are more relevant. [13] broadened the concepts other than nouns and verbs with 1808 concepts, and [19] norms included a reproduction of [5] concepts by adding several hundred more concrete concepts (Table III). [5] norms have features that have been produced by taking into consideration of at least 5 participants, whereas [19] generated set has features that have been produced with a production frequency of two or more which enabled researchers

TABLE II
EXAMPLES OF FEATURES ALONG WITH PRODUCTION FREQUENCIES FOR CONCEPTS FROM Mc rae et al. [5] NORMS.

APPLE	SHELL	BEAR
is_red(26)	used_for_protection(12)	is_large(23)
a_fruit(24)	found_on_beaches(11)	has_fur(20)
grows_on_trees(20)	is_hard(11)	an_animal(19)
is_green (17)	a_house(10)	has_claws(15)

to choose their own cutoff point while excluding idiosyncrasies associated with individual participants. The primary aim of [19]. work was to provide computational linguistics with a useful resource for training and evaluating systems to automatically extract property-norm-like semantic feature representations from text corpora. Researchers have used these norms to discover many aspects of the semantic representation. The conceptual features generated by participants through property norms provide stimuli to test various claims about the representation of conceptual knowledge. Disambiguation has been resolved in his work. [20] expanded his work on three previous databases of concept for over 4400 words including nouns, adjectives, verbs, and other parts of speech (Table IV). The process of conducting the experiments were like that of [5] and Buchanan et al. The dataset has three sources majorly as m = McRae [5] and v = Vinson and Vigliocco[11] and b=Buchanan[20]. For the similarity measure between features the cosine similarity function was used.

TABLE IV
EXAMPLE OF CUE-FEATURE SET FROM Buchanan et al. [20] NORMS

W	Cue	Feature	Translated	F_f	F_t	N	N_f	N_t	Pos_cue	Pos_feature	Pos_translated	A1
b	abnormal	different	Differ	16	16	58	27.5862069	27.5862069	adjective	adjective	verb	characteristic
b	abnormal	normal	normal	16	18	58	27.5862069	31.03448276	adjective	adjective	adjective	0
b	abnormal	normalcy	normal	1	18	58	1.724137931	31.03448276	adjective	noun	adjective	characteristic
b	abnormal	normality	normal	1	18	58	1.724137931	31.03448276	adjective	noun	adjective	characteristic
b	abnormal	ordinary	ordinary	6	6	58	10.34482759	10.34482759	adjective	adjective	adjective	0
b	abnormal	Strange	strange	11	11	58	18.96551724	18.96551724	adjective	adjective	adjective	0
b	abnormal	Weird	Weird	20	20	58	34.48275862	34.48275862	adjective	adjective	adjective	0
b	above	Atop	Top	1	30	59	1.694915254	50.84745763	other	adjective	adjective	characteristic
b	above	Below	Below	12	12	59	20.33898305	20.33898305	other	other	other	0
b	above	High	High	9	28	59	15.25423729	47.45762712	other	adjective	adjective	0

W=Where; b=Buchanan et al.; F_f= frequency of features; F_t=frequency of translated features; N=input; N_f=normalized features; N_t=normalized trnslated features; Pos_cue=Parts of speech for cue; pos_translated=parts of speech for translated features.

III. ISSUES ADDRESSED USING FEATURE PRODUCTION NORMS.

A. Correctness of a sentence.

In humans the understanding of correctness of a sentence comes from underlying semantic knowledge [21]. There are situations when a sentence can be syntactically correct and semantically wrong or vice versa like apple eats jack is syntactically correct but semantically doesn't make sense. Semantic representation through features come into play for understanding. Human understands that apple is type of fruit which is meant to be eaten and the verb eat represents only those objects which are meant for eating. Interestingly a two-year-old cannot understand because of lack of semantic understanding. Same situation arises when it comes to machine understanding a text. Thus, its necessary to render an untrained person or a machine with semantics of a language for its understanding. Here semantic classes come into play.[2] has suggested that any noun entity existing can be classified into eight classes namely specificity, boundedness, animacy, gender, kinship, social status, physical properties, and functional property. He further added that any feature(s) representing a concept, or an entity should fall under a specific class. Though a lot of research has been done for past two decades on categorization of feature class and their properties by different researchers [22]. They mentioned that the meaning of any living object can be easily described using sensory features and for artifacts instead functional features would be more important.

B. Clear understanding of objects, subjects, and verbs

Features classification and representation can be used to differentiate between objects, verbs(events) and subjects. Objects differ from events through various dimensions. Objects are identified mainly through features whereas events possess relations. Researchers have argued that the features can be used to differentiate among different objects also. Some authors[11],[18] have argued that objects and events differ in featural representations. Objects have more feature than events. Moreover, semantically objects are more correlated as compared to events[23,29,30].

TABLE III
EXAMPLE OF PREPROCESSED FEATURES FOR THREE CONCEPTS FROM Devereux et al. [19] NORMS
(this is not a webpage feature representation)

CONCEPT	RELATION	FEATURES	PROD. FREQ
TOMATO	Is	Red	30
	has	Pip_seeds	22
	is	Fruit	9
	is	Circular_round	17
WALLET	Has	Money	26
	Is	For_Men	9
	Has	Bank notes	4
	Has	Cards	18
AEROPLANE	Has	Wings	28
	Is	Expensive	4
	Does	Land	3
	Is	Transport	6

TABLE V
*EXAMPLE OF CONCEPT RELATION FEATURE TRIPLES FROM Kelly et al.[28] DATASET

Feature	Recorded triple	Prod Freq
	CAR	
Has wheels	Car have wheel	19
Used for transportation	Car use transportation	19
Has 4 wheels	Car have 4 wheel	18
Has an engine	Car have engine	13
Require petrol	Car require petrol	13
	Penguin	
Is black	Penguin be black	24
A bird	Penguin be bird	22
Is black-and-white	Penguin be black-and-white	22
Has abeak	Penguin have beak	22
Beh-cannot fly	Penguin cannot fly	21

* the contents in the table has been copied from Kelly et al.[28]

C. *Quantitative measures of semantic similarity*

Semantic features produced in production norms can be used to find the similarity between concepts by using certain distance calculation techniques like cosine similarity [5],[19], Euclidean distance [24].

IV. EXTRACTION METHODS OF SEMANTIC FEATURES AND RELATIONS

From the literature review in the previous sections, it has been observed that cognitive psychologists [5],[6],[17],[19] have contributed majorly to linguistics development by producing feature norms. The mode of experiment used by all of them was similar in nature. The features in their production norms were produced by participants from various domains. Though data collected through such experiments was used by many researchers for their work, but it suffered from various anomalies. The participants sometimes do not report certain properties even though he is aware of facts concerned with those properties. For e.g., in [19] a feature for whale is ***does breathe*** whereas this feature was absent for tiger. Moreover, participants can write only those features which they are able to depict verbally. The mental representations are far more enriched then verbalize set. But the original data can be obtained from human beings only which can be manipulated by using various computations. So, there is need to overcome these anomalies by using techniques from NLP. There are various semantic features extracting methods used by various researchers.[25] aims to build a system capable of extracting feature properties. Their approach was based on distributional semantic models and was meant to reduce the labor-intensive tasks of manually generating new features. Many other researchers also contributed to this field.[26,32] proposed a different approach called Strudel, the aim of this approach was to capture semantically meaningful patterns rather than flat co-occurrence of words. It considered concepts and searched for those nouns, verbs, and adjectives which were linked to the target concepts by a finite set of templates. The next step ranked the concept-property pairs based on the number of distinct linking patterns. Strudel method had highest precision. Another work[27,31] used class-based approach for extracting semantic features. It was the first work which apart from extracting the features also performed prediction of relations between concepts and features. RASP parser was used to generate grammatical relations. This work became the baseline for another semantic feature extracting approach[28].Their work was based on seeking common sense properties like tomato is a vegetable from some common preexisting datasets, but their dataset says it as a fruit. They used syntactic, semantic, and encyclopedic information and designed rules to extract concept, relation and feature triples shown in (Table V).

V. DISCUSSION

The aim of this paper was to review the journey of production of the semantic features and their methods of extraction. It has been observed that there have been many useful datasets created by linguistics psychologists for extending the English lexicons. Though the experiments conducted by these experts were considered as baselines for many researchers, but they were computationally weak. So, a review of various feature extraction techniques have also been discussed.

VI. REFERENCES

[1] Czezowski. T, “Connotation and Denotation", *Semiotics in Poland,1979, pp. 73-80.*

[2] Frawley, William ," Linguistic Semantics", Hillsdale, NJ , Lawrence Erlbaum Associates 1992,pp. 10-20.

[3] Tulving, E," Episodic and semantic memory", In E. Tulving & W. Donaldson (Eds.), *Organization of Memory ,* New York, NY, Academic Press, Inc., pp. 382-402.

[4] Gregor Sieber, Brigitte Krenn ,"Episodic Memory for Companion Dialogue" , Proceedings of the 2010 *Workshop on Companionable Dialogue Systems, Association for Computational Linguistics*, Sweden, pp. 1-6.

[5] McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C.,"Semantic feature production norms for a large set of living and nonliving things", 2005, *Behavior Research Methods*, *37*(4), pp.547–559.

[6] Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. "Representing the meanings of object and action words: The featural and unitary semantic space hypothesis",2004, *Cognitive Psychology*,*48*(4), pp.422–488.

[7] Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. ," Topics in semantic representation",2007, *Psychological Review*, *114*(2), pp.211–244.

[8] Jones, M. N., & Mewhort, D. J. K., "Representing word meaning and order information in a composite holographic lexicon",2007,*Psychological Review*, *114*(1), pp.1–37.

[9] Riordan, B., & Jones, M. N. ,"Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation",2011,*Topics in Cognitive Science*,3(2),pp. 303-345.

[10] Smith E. E., Shoben E. J., Rips L. J.," Structure and process in semantic memory: A featural model for semantic decisions",1974, *Psychological Review 81*(3),pp.214-241.
[11] Vinson, D. P., & Vigliocco, G., "Semantic feature production norms for a large set of objects and events",2008, *Behavior Research Methods*, 40(1), pp.183–190.
[12] Buchanan, E. M., Holmes, J. L., Teasley, M. L., & Hutchison, K. A., "English semantic word-pair norms and a searchable Web portal for experimental stimulus creation",2013, *Behavior Research Methods*, *45*(3), 746–757.
[13] Landauer, T. K., & Dumais, S. T.,"A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge",1997, *Psychological Review,*205 *104*(2),pp. 211.
[14] Montefinese, M., Vinson, D., & Ambrosini, E., "Recognition memory and featural similarity between concepts: the pupil's point of view,2018, *Biological Psychology, 135*, pp.159-169.
[15] Kousta S.-T., Vigliocco G., Vinson D. P., Andrews M., & Del Campo E. "The representation 199 of abstract words: Why emotion matters",2011, *Journal of Experimental Psychology: General,* 200 *140*(1), 14–34.
[16] Sebastian J. Crutch1,2 and Elizabeth K. Warrington," Abstract and concrete concepts have structurally different representational frameworks", 2004,*Advance Access publication*, pp. 615-627.
[17] Levin B., "English Verb classes and Alternations: A Prelimnary Investigations", 1993,University of Chicago Press; New edition ,ISBN-13:978-0226475332.
[18] Korhonen A.,Briscoe T.,"Extended Lexical-Semantic Classification of English Verbs",2004, *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, Association for Computational Linguistics, pp. 38-45.
[19] Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, Billi Randall," The Centre for Speech, Language and the Brain (CSLB)concept property norms", *Behavior Research Methods*, 2014, Springer US, Volume 46, Issue 4, pp 1119–1127.
[20] Erin M. Buchanan,K. D. Valentine, Nicholas P. Maxwell, " English semantic feature production norms: An extended database of 4436 concepts*", Behavior Research Methods*,2019, Published online: 1 May 2019,© The Psychonomic Society, Inc. 2019.
[21] Chierchia, Gennaro and Sally McConnell-Ginet," Meaning and Grammar: An Introduction to Semantics",2000,The MIT Press; second edition, ISBN-13: 978-0262531641,
[22] Faraah M.J. , Mc Clelland J. L.," A computational model of semantic memory impairment: Modality specificity and emergent category specificity",1991, *Journal of Experimental Psychology: General*, **120**, pp. 339–357.
[23] Vinson D. P., Vigliocco G., Cappa S., Smi, S.,"The breakdown of semantic knowledge: Insights from a statistical model of meaning representation",2003, *Brain & Language*,pp. 347-365.
[24] Viglicco G., Vinson D.P., Lewis W., Garrett M. F. "Representing the meanings of object and action words: The featural and unitary semantic space hypothesis",2004,*Cognitive Psychology*,pp. 422-488.
[25] Baroni, M., & Lenci, A., "Distributional memory: A general framework for corpus-based semantics",2010,*Association of Computational Linguistics*, 36(4), pp. 673–721
[26] Baroni, M., Murphy, B., Barbu, E., & Poesio, M., "Strudel: A corpus-based semantic model based on properties and types",2010, *Cognitive Science,* 34, pp. 222–254.
[27] Devereux, B., Pilkington, N., Poibeau, T., & Korhonen, A., "Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data.",2009, *Research on Language & Computation,* , pp. 137–170.
[28] Kelly C., Devereux B., Korhonen A. "Automatic Extraction of Property Norm-Like Data From Large Text Corpora", 2013,Cognitive Science A Multidisciplinary Journal ,pp. 638-682
[29] Gupta, S., & Garg, P. (2021). An insight review on multimedia forensics technology. Cyber Crime and Forensic Computing: Modern Principles, Practices, and Algorithms, 11, 27.
[30] Pustokhina, I. V., Pustokhin, D. A., Lydia, E. L., Garg, P., Kadian, A., & Shankar, K. (2021). Hyperparameter search based convolution neural network with Bi-LSTM model for intrusion detection system in multimedia big data environment. Multimedia Tools and Applications, 1-18.
[31] Gupta, M., Garg, P., & Agarwal, P. (2021). Ant Colony Optimization Technique in Soft Computational Data Research for NP-Hard Problems. In Artificial Intelligence for a Sustainable Industry 4.0 (pp. 197-211). Springer, Cham.
[32] Beniwal, S., Saini, U., Garg, P., & Joon, R. K. (2021). Improving Performance During Camera Surveillance by Integration of Edge Detection in IoT System. International Journal of E-Health and Medical Communications (IJEHMC), 12(5), 84-96.