Handcrafted Semantic Hierarchy to Develop Urdu WordNet

Palwasha Jogezai, Ayesha Zafar
Kinnaird College for Women, Lahore
palwashajogezai@gmail.com, ayeshazafarsultan@gmail.com

Abstract

This research paper highlights the issues and challenges faced during the development of semantic hierarchy of Urdu WordNet. 118 developed hierarchies are analyzed by using database World Atlas of Language Structure alongside the resources used for previous researches on Urdu WordNet [7] [8]. Problematic hierarchies are discussed with solutions supplied.

1. Introduction

WordNet is an online lexical database which is developed on the basis of contemporary theories of psycholinguistics of human lexical memory, at Princeton University. Its basic concepts and purpose of development are shared in the Five Papers on WordNet [1]. The construction of Princeton WordNet (PWN) [2] inspired the development of lexical databases for other languages [3] [4] [5] [6]. Urdu WordNet [7] was also developed following its pattern in order to align with linguistic, cultural and other contexts in Pakistan. Further, the developed senses of Urdu WordNet (UWN) were aligned to PWN 2.1 [8].

This research focuses mainly on building hypernym-labeled noun hierarchies of Urdu WordNet because the detail about hypernyms and hyponyms semantic relation provides additional disambiguating information. The lexical resource was manually constructed and it is helpful for natural language processing tasks such as machine translation and information retrieval system.

The paper is organized in the following sections. Section 2 reviews the current literature regarding WordNet hierarchies based on semantic relations. Section 3 describes the approach of developing Urdu WordNet semantic hierarchies. Sections 4 presents the challenges and solutions faced during the process. Finally, Section 5 concludes the paper by reporting the future work required in this direction.

2. Literature Review

The lexicon division of WordNet has been done into five classes, i.e. nouns, verbs, adjective, adverbs and function words which makes it different from a standard dictionary. Similarly, it incorporates the details of semantic relations [9]. For example, hyponymy/hypernymy semantic relations are represented in the creation of WordNet. Semantic relation that occurs between word meanings is hypernymy/hyponymy: e.g., {tree} is hypernym of {plant}, {plant} is a hypernym of {maple}. It's not like the lexical relations that occur between word forms are synonymy and antonymy. It is variously named subordination/superordination, subset/superset, or the ISA relation [9]. Hyponymy generates a hierarchical semantic organization that is duplicated in the noun files by the use of labeled pointers between sets of synonyms (synsets) [25].

In conventional dictionaries the definition of tree, for example, doesn't include information that trees possess roots. There is no information regarding its coordinate term. Similarly, the information related to its kinds and features are also not available in common dictionaries. All such information is missing due to economic pressure to minimize redundancy [9]. The knowledge about such semantic relations is useful in NLP applications: e.g., reformulation of query [10] [27], addressing query [11], summarization of written text [12], classification [13]. There are two categories in which the task on concept hierarchy learning [14] is done: Harris [15] distributional hypothesis & Hearst's [16] lexical patterns. The basic order imposed on nouns semantic memory is a tree, not circular form, meaning that lexicographers give tree represented graphically. So the construction of lexical tree can be done by adopting superordinate terms: oak – tree – plant – organism, and according to Miller et al., this can be read as "is a" or "is a kind of." This shows that a hierarchy is going from upward to limited generic terms from huge amount of specific terms at the bottom [9]. Hierarchies are said to be providing the "conceptual skeletons" for nouns; say for example the knowledge about some specific nouns is found to

be hung onto this tree like structure e.g. Christmas tree [9].

In the field of computer science such hierarchies are known by the term inheritance systems [9]. Psycholinguists assume that that comparison of superordinates cannot be done with hyponyms [17]. However, Quillian claimed explicitly that inheritance system is formed through the lexical memory of individuals for nouns [18] [19]. Collins and Quillian reported experimental tests, assuming hierarchical levels number can be identified when the two meanings are taken apart [9] [20]. Cognitive scientists and linguists came to conclusion that Quillian was not right, inheritance system is not how semantic memory is organized [21] [22] [23]. Miller and Charles indicated that there is no difference in meaning of word, but it's in the usage of word or that the distance is established semantically, not pragmatically [24].

The hierarchical principle can be construed with assumption that the existence of all nouns is in one hierarchy [9]. Theoretically, {entity} is placed at the bottom and {object, thing} are placed as nearest subordinates and then continuing towards specific meanings. However, practically, it fails to convince. The solution lies in partitioning the nouns due to various advantages one being less size occupied and the possibility of having the different files assigned to lexicographers for writing and editing. But, again, the problem is to choose the primitive semantic components. This issue of having the possibility of combination of adjective-noun happening was resolved by Philip N. Johnson-Laird, being discussed in the revised edition of Miller et al. with rationale [9]. This is one of the major challenges being faced even while development of Urdu WordNet, where levels didn't go any deeper.

3. Research Methodology

The manual constructions of 118 hypernym-labeled noun hierarchies of Urdu WordNet were done by incorporating both expand and merge approaches. The hierarchies of hypernym-labeled nouns in PWN 2.1 were kept as a model [25]. Moreover, World Atlas Language Structures was also used as a supporting resource. Urdu WordNet 1.0 Wordlist [26] was used as corpus and high frequency nouns from Urdu news websites were selected. The following steps were followed during the process.

1. Urdu Nouns were chosen from the corpus.
2. The selected nouns were mapped with PWN 2.1 Sense ID.
3. Urdu noun hierarchies were constructed following the PWN 2.1 hypernyms.
4. Missing Urdu senses were added to complete the hierarchies.
5. Shallow Senses (of the hierarchies) are translated from PWN 2.1 to complete the Urdu hierarchies.
6. Urdu language resources Qamos-e-Mutradifat and Urdu Lughat were used in order to check the accuracy of the developed hierarchies.
7. Typological information (to confirm the hypernyms hierarchy interruption) was done from World Atlas Language Structures dataset.

The example of a complete hierarchy of Urdu word کام has been shown in **Figure 1**.

Eng ID	Eng Word	Category	Concept	Example	Synset
00708623	job	noun.act	روزمرہ یا مقررہ وقت کا کام	وہ ساداون گھر پر کام کرتی رہتی ہے	کام
00708412	duty	noun.act	کام جو اتفاقاً یا اسوقاً کرنا ضروری ہو	اس کی ڈیوٹی آج سے پانچ بجے تک ہے	ڈیوٹی، فرض، منصبی، خدمت، کام، نوکری
00570312	work	noun.act	کچھ کرنے کا عمل یا کام	تمام لوگوں نے کار خیر میں بڑھ چڑھ کے حصہ لیا	کار، کام، عمل
00403481	activity	noun.act	قوت عمل، سرگرمی، حرکیت	ہمیشہ تبت سرگرمیوں میں حصہ لینا چاہیے	فعالیت
00029085	human action/ act	noun.Tops	انسان کا کسی کام کو کرنے کا عمل	اس سارے عرصے میں کوئی انسانی سرگرمی دیکھنے میں نہیں آئی	انسانی سرگرمی
00028105	event	noun.Tops	واقعہ جو کسی وقت کسی جگہ وقوع پذیر ہو	اس واقعہ کے بعد وہ کبھی نظر نہیں آیا	واقعہ، ماجرا، ایونٹ
00022007	psychological feature	noun.Tops	کسی جاندار کی ذہنی تبدیلی کا ایک پہلو	اس کے اس برتاؤ کے پیچھے کون سی نفسیاتی خاصیت کارفرما ہے	نفسیاتی خاصیت
00002236	abstraction	noun.Tops	نظریہ یا خیال جو خاص مثالوں کے مشترکہ عوامل پر مشتمل ہو	تجرید کا عمل انسانی ترقی کے لیے ضروری ہے	تجرید
00002119	abstract entity	noun.Tops	چیز جس کا وجود صرف خیالوں میں ہو، وہ چیز جو صرف غیر مرئی شکل میں موجود ہو	غیر مرئی اشیا نظر نہیں آتیں	غیر مرئی شے
00001740	entity	noun.Tops	چیز جو اپنا وجود رکھتی ہو چاہے مادی یا خیالی	کسی شے کا وجود ہی اور خیالی دونوں صورتوں میں ہو سکتا ہے	وجود

Figure 1: Hierarchy of Urdu Word کام

For example hierarchy of moon in PWN 2.1 is as follows:

- Moon-- Star-- Celestial body-- Natural object-- Object-- Physical entity—Entity

Similarly, the hierarchy of چاند was developed by following the above hierarchy as:

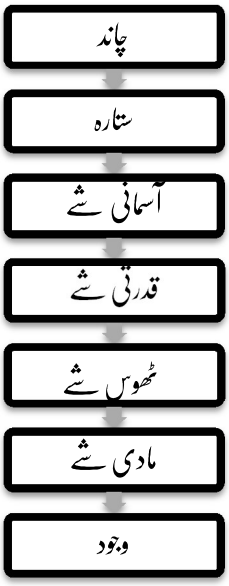

Figure 2: Hierarchy of Urdu Word چاند

4. Challenges and Solutions

In this section the Interrupted Hypernym-labeled Noun Hierarchies are being shared with explanation and detailed analysis on how and why were they being interrupted.

• 4.1. Development and Mapping of Urdu WordNet Hierarchies with PWN 2.1

The biggest challenge was to align and map Urdu WordNet hierarchies with PWN 2.1. In order to complete the hierarchies many new Urdu senses were developed manually. The example of newly added senses to complete the hierarchy of Urdu word تجربہ has been shown in the following **table 1** below:

English ID	English Word	Category	Urdu Word
00633318	**experiment**	noun.act	**تجربہ**
00635582	**research project**	noun.act	**تحقیقاتی منصوبہ**
00630686	**research**	noun.act	**تحقیق**
00627860	**investigation**	noun.act	**دریافت، استفسار، تفتیش، انویسٹی گیشن، کھوج، تلاش**
00570312	**work**	noun.act	**کار، کام، عمل**
00403481	**activity**	noun.Tops	**فعالیت**
00029085	**human action**	noun.Tops	**انسانی سرگرمی**
00028105	**event**	noun.Tops	**واقعہ، ماجرا، ایونٹ**
00022007	**psychological feature**	noun.Tops	**نفسیاتی خاصیت**
00002236	**abstraction**	noun.Tops	**تجرید**
00002119	**abstract entity**	noun.Tops	**غیرمرئی شے**
00001740	**entity**	noun.Tops	**وجود**

Table 1: Example of Newly Added Senses of Urdu Word تجربہ

In order to complete this noun hierarchy, two new senses تحقیقاتی منصوبہ and تحقیق were added. The missing senses were translated from PWN 2.1 with their concepts.

• 4.2. Translation Issues

Many of the Urdu Noun hierarchies couldn't complete because of the translation issues. All the synsets of the hypernym are not present in English language because of the difference and unavailability of concepts. For example, اڈا is not found in English culture. Similarly, اردو بازار، صدر بازار are places which are loosely translated in English as market but that's a very weak translation. Although the word "Bazar" is found in English dictionaries but it gives a different sense. Moreover, its concept doesn't match with the

Urdu Bazaar and Sadar Bazar because these are limited to Urdu culture.

Moreover, it was difficult to the complete Noun hypernymy hierarchies of آئل "Oil" (ID: 7568129). Only shallow levels of its hierarchy were developed. It was problematic to find out the exact translation of the words: lipid – macromolecule – organic compound. These are scientific terms and are not found in OUD.

For translation of "lipid" words that were looked up were: حیاتی کیمیا؛ لحمیات؛ روغنیات؛ شحمیات؛ چربی but none of them were mapped with the PWN 2.1 sense for lipid. Likewise, for "macromolecule" no Urdu translation exists. There is translation for "molecule" but macromolecule is also a scientific term that needs to be added to Urdu language. As for "organic compound" It's formed by the combination of two words, no such phrase exists in OUD.

Another example is English Word "terminology" (ID 6220865) which was translated as اصطلاح. It's evident that this difference between English and Urdu causes the term language unit not remains in the state of noun when it is translated. Hypernymy hierarchy of two words: "newspaper advertisement" (ID 7150204) and "advertisement" (ID 7149579) get interrupted due to word "promotion". Both "newspaper advertisement" and "advertisement" are translated as اشتہار so there is one noun for both of these nouns. Another issue is the word "promotion" which has no concept in Urdu.

For the word "discipline", انضباط, تادیب, مضمون, قاعدہ, لِٹریچَر were looked up in OUP but none of them gave English sense.

It was observed that noun hierarchy as lexical inheritance system is limited in depth and it seldom goes more than ten levels deep. The deepest examples usually contain technical levels that are not part of everyday vocabulary which have been discussed above with reference to translation issues. Moreover, shallowest levels are also considered as too vague

• 4.3. Mismatch of Part of Speech Category

Mismatch of part of speech category was another challenge. English word "abstract entity" is Noun in PWN 2.1 where as its translated word غیرمرئی is an Adjective. However, the same translated word was used in the hierarchies because it was used with the word شے so a compound word غیرمرئی شئے (Adj+N) was used.

• 4.4. Addition of translated Words

For the alignment and completion of Urdu hierarchies translated words were adopted and used in order to complete the hierarchies. However, these translated words are not available in OUD. For example, it was difficult to find out the proper word for "humanistic discipline" because humanistic is an adjective in Urdu and no coined word as "humanistic discipline" exists. Therefore, انسان شناس ادب as a translated word is used. However, such words are made considering compound translations of the English words which further lead towards the mismatch of part of speech category.

Also, definition for "Indo-European" and "Indo-Iranian" were not found in Urdu Lughaat so they were self created by taking definition from Wikipedia.

• 4.6. Issues of Compound Words

Another important issue was of compound words. English word "written communication" in PWN 2.1 is a compound word. It couldn't be found in Urdu language in compound form. So it needs to be added to have a complete set of hierarchy formed. Another example of interruption caused by "written communication" is that of ادب، لٹریچر "literature" (ID 6279556). Yet another hierarchy being interrupted by "writing communication" is اساطیر، This hierarchy for "myth" (ID 6287133) couldn't go to deep level due to absence of coined word "writing communication". Another reason was the multiple senses for the term "myth" as اساطیر،افسانے، کہانیاں،دیوکتھا were giving the same sense for the above mentioned single word.

Moreover, the "auditory communication" wasn't in OUD and its translation as سمعی مواصلات was used in order to complete the noun hierarchy and its semantic relations. Not only that but also the word "nonstandard speech" is not available in Urdu and it would need to have the word ہوا added which would turn it into non-Noun form. Just like "auditory

communication" and "written communication" nouns, another form "visual communication" causes interruption in the construction of hypernymy hierarchy of "name" with sense ID 6248892.

• 4.5. Confusing Hyponyms with Hypernyms

Another major challenge was to discover the actual relation of hypernym-hyponym. A noun "X" is a hyponym of a noun "Y" if "X" is a subtype or instance of "Y". Thus "Ashfaq Ahmed" is a hyponym of author and conversely "author" is a hypernym of "Ashfaq Ahmed" and so on.

5. Conclusion

There are issues and challenges constructing noun hypernymy hierarchy for Urdu WordNet by aligning and mapping it with PWN 2.1. Lexical relations, specifically hyponymy-hypernymy, are important in the development of information retrieval systems. There is rapid alteration in the theme of lexical semantics, computational lexicography, and computational semantics, but due to the availability of online lexical resources the construction of noun hypernymy hierarchies of Urdu WordNet is feasible. However, for Urdu language there is still need to develop more hierarchies in order to make Urdu WordNet's coverage better. Although, manually developed Urdu hypernymy hierarchy of nouns are highly accurate It is concluded that with handcrafted hierarchies there's a need for an automatic construction of hypernymy hierarchies of Urdu WordNet. Further research needs to be conducted in partitioning of noun hierarchy into separate hierarchies with unique top hypernyms. Moreover, it can further lead towards Parts and Meronymy: part whole relations.

6. References

[1] G. A. Miller, *Five Papers on WordNet,* 1993.

[2] C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, 1998.

[3] U. N. Singh. *Proceedings of the First Global WordNet Conference.* Central Institute for Indian Languages, Mysore, India, 2002.

[4] P. Sojka, K. Pala, P. Smrz, C. Fellbaum, and P. Vossen. *Proceedings of the Second International WordNet Conference*. Masaryk University, Brno, Czech Republic, 2004.

[5] P. Sojka, K. Choi, C. Fellbaum, and P. Vossen. *Proceedings of the Third Global WordNet Conference.* Masaryk University, Brno, Czech Republic, 2006.

[6] P. Vossen. EuroWordNet. Dordrecht, Holland: Kluwer, 1998.

[7] A. Zafar, A. Mahmood, F. Abdullah, S. Zahid, S. Hussain, and A. Mustafa, "Developing Urdu WordNet Using
the Merge Approach *", in the Proceedings of Conference on
Language and Technology 2012 (CLT12)*, Lahore, Pakistan.

[8] A. Zafar, A. Mahmood, S. Shams, and S. Hussain, "Structural Analysis of Linking Urdu WordNet to PWN 2.1", *in the Proceedings of Conference on Language and Technology 2014 (CLT14)*, Karachi, Pakistan.

[9] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. Introduction to WordNet: An On-line Lexical Database. Princeton University, New Jersey, 1993.

[10] R. Jones, B. Rey, O. Madani, and W. Greiner. "Generating query substitutions." In WWW '06: *Proceedings of the 15th international conference on World Wide Web*, New York, NY, USA. ACM, 2006, pp.387–396.

[11] V. Lopez, V. Uren, E. Motta, and M. Pasin. Aqualog: An ontology-driven question answering system for organizational semantic intranets. Web Semant., 2007, pp. 72–105.

[12] C. Dang, X. Luo, X., and H. Zhang. WordNet-based summarization of unstructured document. W. Trans. on Comp., 2008, pp. 1467–1472.

[13] J. Li, Y. Zhao, and B. Liu. "Fully automatic text categorization by exploiting wordnet." In *AIRS '09: Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, Berlin, Heidelberg. Springer-Verlag, 2009, pp. 1-12.

[14] S. Afonin. "On Automated Hypernym Hierarchy Construction Using and Internet Search Engine." Moscow, Russian Foundation, 2010.

[15] Z. S. Harris. Mathematical Structures of Language. Wiley, New York, 1968.

[16] M. A. Hearst. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the 14th International Conference on Computational Linguistics,* 1992, pp. 539–545.

[17] T. G. Bever and P. S. Rosenbaum. "Some Lexical Structures and Their Empirical Validity" in Jacobs, R. A., & Rosenbaum, P. S. (eds.). Readings in English Transformational Grammar. Waltham, Mass.: Ginn, 1970.

[18] M. R. Quillian. “Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities.” Behavioral Science, 1967, pp. 410-430.

[19] M. R. Quillian. “Semantic Memory.” In *Minsky, M. (ed.). Semantic Information Processing.* Cambridge, Mass.: MIT Press, 1968.

[20] A. M. Collins and M. R. Quillian, M. R. ”Retrieval Time From Semantic Memory.” *Journal of Verbal Behavior and Verbal Learning,* 1969, pp. 240-247.

[21] A. J. Wilkins. “Conjoint Frequency, Category Size, and Categorization Time.” *Journal of Verbal Learning and Verbal Behavior,* 1971, pp. 382-385.

[22] E. E. Smith. “Theories of Semantic Memory.” In Estes, W. K. (ed.). Handbook of Learning and Cognitive Processes, vol. 5. Hillsdale, NJ: Erlbaum. The Synonym Finder. 1978. Emmaus, Pa.: Rodale Press, 1978.

[23] C. Conrad. “Cognitive Economy in Semantic Memory.” *Journal of Experimental Psychology*, 1972, pp. 75-84.

[24] G. A. Miller, and W. Charles. “Contextual Correlates of Semantic Similarity.” Language and Cognitive Processes, 1991.

[25] A. Zafar, "Development of Urdu WordNet Noun Hierarchies and their Hypernymy Relations", Presented at 3-Day International Workshop on Corpus Linguistics, 2015.

[26] Center for Language Engineering, *Urdu WordNet 1.0 Wordlist,* 2013. Available at: http://www.cle.org.pk/Downloads/ling_resources/wordlists/Urdu%20WordNet%201.0%20Wordlist.pdf

[27] P. A. Chirita, S. C. Firan, and W. Nejdl. « Personalized query expansion for the web.” In *SIGIR '07:Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA. ACM, 2007, pp. 7-14