Aligning the IndoWordNet with the Princeton WordNet

Nandu Chandran Nair
DISI
University of Trento
Trento,Italy
nandu.chandrannair@unitn.it

Rajendran S Velayuthan
CEN
Amrita Vishwa Vidyapeetham
Coimbatore,India
rajushush@gmail.com

Khuyagbaatar Batsuren
DISI
University of Trento
Trento,Italy
k.batsuren@unitn.it

Abstract

The IndoWordNet is an Indian language lexical resource. The project started with Hindi WordNet, which was manually built from various resources with the preference for culture-specific synsets. Other languages were added later. The development approach used in IndoWordNet is very similar to that used in Princeton WordNet (PWN). PWN is a semantic network where English synsets are nodes, and semantic relations are edges connecting them. Due to the popularity of PWN, IndoWordNet also connected Hindi and English languages through direct (synonymy) and hypernymy linkages between their synsets. Due to the diversity of the languages, these linkages generate three types of mappings between IndoWordNet and PWN which generate the misalignment. This paper proposes to align the IndoWordNet with PWN using a large scale lexical-semantic resource called Universal Knowledge Core (UKC), which forms a semantic network where nodes are language-independent concepts. In the UKC semantic relations connect concepts and not synsets.

1 Introduction

Studies are in progress to make language resource development process cheap and quick, but even now, the process demands considerable resources and expert support. The generation of a language resource is influenced by many factors such as large global speaker population, high economic power, or high political interests (Stüker, 2009). As a result the majority of languages are under-resourced (Besacier et al., 2014). Even in 2019, if we use google translator for one of the official Indian languages, Malayalam, we can notice how a few words remain unrecognized (Figure 1). Consider the sample Malayalam sentence: "രാമു ചമ്മന്തി കഴിക്കില്ല(*Ramu chammanthy kazhikkilla*), translated as "Ramu will not eat". Here, "Chammanthy" is an Indian dish, and the translator has

Figure 1: Missing term in translator

failed to find an appropriate translation for this word. A language resource that allows culture-specific words should have the missing term in the target translation.

Back in 2006, the joint efforts of different universities and research groups across India made it possible to develop the IndoWordNet(Bhattacharyya, 2010) - the first wordnet for Indian languages. IndoWordNet was developed to capture the cultures of India in length and breadth by including 18 languages out of 22 official languages(Bhattacharyya et al., 2010). Hindi WordNet(Narayan et al., 2002)developed at IIT Bombay, India was used as the source wordnet for IndoWordNet. Other WordNets in IndoWordNet were extended from Hindi WordNet with culture-specific and language-specific synsets. In this paper, we use the notation "IndoWordNet" to refer to the project and notation "IWN" to refer to on the Hindi WordNet. The IndoWordNet team followed Princeton WordNet(PWN)(Fellbaum, 2012) principles at a minimum level during the development.

The IndoWordNet team also focused on the translation(Chakrabarti and Bhattacharyya, 2004) across Indian languages and English and they identified the challenges for linking Hindi with English(Saraswati et al., 2010). Based on this, IndoWordNet team proposed direct (synonymy) and hypernymy linkages. These types of linkages eventually cause different types of associations between the synsets of IWN and PWN. Our challenge is to align IWN with PWN. This could allow to generate automatic dictionary across terms and also highlight

the diversity among languages (Giunchiglia et al., 2017).

Our approach involves the usage of a large scale lexical-semantic resource called Universal Knowledge Core (UKC)(Tawfik et al., 2014). UKC forms a semantic network of language-independent concepts, which are linked with semantic relations. In our approach, we group the IWN synsets into three groups. We process each group of synsets in such a way to make them in a single group where one IWN synset has a concept in UKC. We have aligned IWN with PWN and find around 20K new concepts for PWN. Also, we identified around 3K synsets from IWN, which have no hypernym relations with other synsets.

The structure of the paper is as follows. Section II briefly describes IWN, PWN, and other multilingual resources like EuroWordNet, Global WordNet Grid and UKC. Section III describes the issues in the mapping of IWN with PWN. The detailed description of our approach is provided in section IV. In section V, the results obtained from the project are given. Finally, our conclusions and directions for future work are presented in section VI.

2 Background

Many multilingual wordnets such as EuroWordNet (Vossen, 1998), MultiWordNet (Pianta et al., 2002), and Global WordNet Grid (Pease et al., 2008) have been built based on PWN. EuroWordNet (EWN) languages are linked to a list of unstructured English word meaning. EWN has wordnets with the same structure as PWN. By translating words from PWN, MultiWordNet is adapted to the hierarchical structure of PWN and concepts of western culture. Global WordNet Grid combines wordnets and connects them to an ontology that contains core concepts of PWN like "person". Hence, concepts from many languages are defined using English in Global WordNet Grid aligned with the ontology of PWN, and in this paper, we focus on wordnet from India generated based on Hindi.

India is very diverse in many ways: religion, cultures, languages, etc. As many as 880 languages are spoken in India, and 22 official languages are adopted by different states and union territories. Hindi is one of the official languages of India. Hindi belongs to the Indo-Aryan language family, a sub group of Indo-European language family. Hindi, like any language, is enriched with concepts that are cultural manifestations. These concepts are available as lexical items in Hindi but may not be available in other languages. For example, the case of kinship terms in English. Figure 2 shows the eight words used for "cousin" based on maternal and paternal relationships.

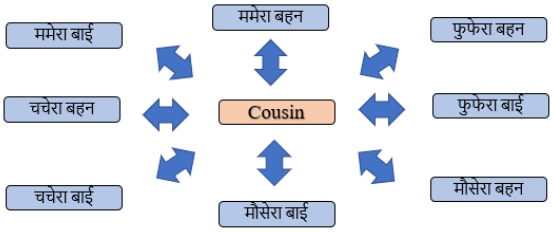

Figure 2: "Cousin" in English and Hindi

A project to develop a linked lexical knowledge base of Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan language families is known as IndoWordNet. It was coordinated by IIT Bombay, India with the assistance of research groups from different parts of India. Universities in various parts of India were responsible for the development of each language wordnet. Other languages were translated from Hindi WordNet to generate the IndoWordNet's respective wordnets. Synsets are linked by relations such as hypernymy or meronymy or troponymy. The same synset identifier maintained across the languages. IndoWordNet were used in the following projects conducted at India: Indian Language to Indian Language Machine Translation (ILILMT), Cross-Lingual Information Access (CLIA) and Indian language sentiment analysis (Dash et al., 2017).

One of the challenges of IndoWordNet team was the term translation from the Indian languages to English (Chakrabarti and Bhattacharyya, 2004). The study (Saraswati et al., 2010) lists the challenges faced when linking IWN and English synsets. The work proposed two types of linkages for connecting IndoWordNet synsets with English synsets: direct and hypernymy. The direct linkage occurs if synsets from IWN have synonyms in English and hypernymy linkages occur if synsets from IWN have no equivalents in English WordNet but only are general synsets. Possible areas of hypernymy linkages can be: kinship relations, musical instruments, kitchen utensils, tools, species and grains(Saraswati et al., 2010). Hence we can argue that PWN and IWN have different hierarchy between synsets. Figure 3 shows that in the PWN, the word "chair" has parent "seat" and "seat"

has parent “furniture”. In IWN, “chair” has four parents,“artifact”,“thing”,“being” and “seat”. And “seat” does not have “furniture” as a parent but “artifact”,“thing”, and “being” as parents.

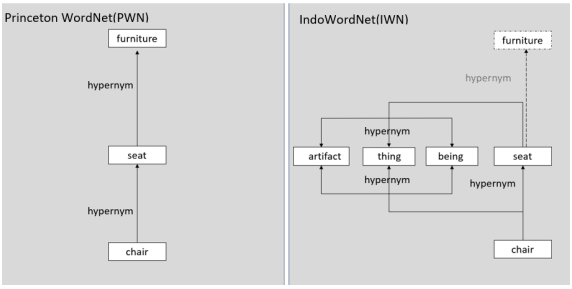

Figure 3: Ontology in PWN and IWN

Two methods are usually used to develop wordnets: merge (Snow et al., 2007) and expansion (Balkova et al., 2004). The merge approach uses the available language resources such as corpora, dictionary, or wordnet to create unique language-dependent wordnet. The merge approach relies entirely on language experts, and the resources being available. Also, the resultant wordnet from merge approach will have concepts that do not exist in PWN. For example, Dutch WordNet from EWN. The expansion approach translates a set of synsets from wordnet into a target language. The expansion has the advantage of extending semantic relations of the source wordnet and the disadvantage of being biased towards the source wordnet with less consideration towards finding the target wordnet's novel concepts. This means that the wordnet resulting from expansion approach has extensive coverage of concepts from the PWN, if PWN is used for translation. One such example is the Spanish WordNet from EWN.

Here we follow a third, somewhat different approach. We take two available wordnets, namely PWN and IWN, and we align them using the UKC so that the synsets in IWN and PWN which have the same meaning are put in correspondence. Hence our approach avoids the biasing towards any language, especially English, and hence finding the missing concepts is less hard than EWN. Also, our approach belongs on top of the previous approaches since we use existing wordnets, and saves time by not to focus on generating wordnet.

The UKC is also a multilingual lexical database based on the WordNet principles, but in the UKC the meaning is represented using lexical concepts. The UKC considers a concept as a mental representation of what is perceive. As such it is language independent(Giunchiglia et al., 2018). The UKC has been designed in such a way that there is no bias towards any language and culture which makes the UKC extendable and open. UKC contains the lexicons and lexico-semantic relations for 338 languages, containing 1,717,735 words and 2,512,704 language-specific word meanings along with 107,196 lexical concepts excluding named entities (Batsuren et al., 2019).

UKC has two components: Language Core (LC) and Concept Core (CC). In LC, each synset is associated with one language and at least one word within that language. The synsets are linked with concepts, satisfying the condition that each synset is linked with only one concept. CC is a semantic network where nodes are language-independent concepts. Each concept has a unique id which differentiates it from any other concept. The CC has a set of semantic relations between the nodes that relate the meanings of the concepts.

In addition to this, UKC also handles the lexicalized missing concept known as lexical gaps for a language by adding a new concept for that language along with a gloss. This gloss considers a local language description of the missing synset. UKC handles the languages independently and is capable of performing language similarity and diversity studies(Giunchiglia et al., 2017). UKC was used as the core source for finding cross-lingual evidence in a multilingual task (Batsuren et al., 2019). The studies (Bella et al., 2017) and (Bella et al., 2016) explain some applications of UKC. Figure 4 shows how the synsets of English and Italian are concepts aligned in UKC. LC has the vocabularies for the concepts “chair”, “seat” and “furniture” in English and Italian languages.

3 Problem Definition

Indian languages and English derive from different cultures and show language specific phenomena such as complex predicate structure(Chakrabarti et al., 2007). The linkages between IWN and English mentioned above cause three types of mappings between the IWN and PWN synsets: one to one mapping, many to one mapping, and one to zero mapping.

In this paper, we take mapping in the sense of “adding an equivalence relation for each synset in IWN to the closest synset in PWN”. Such types of mappings vary upon the languages. For exam-

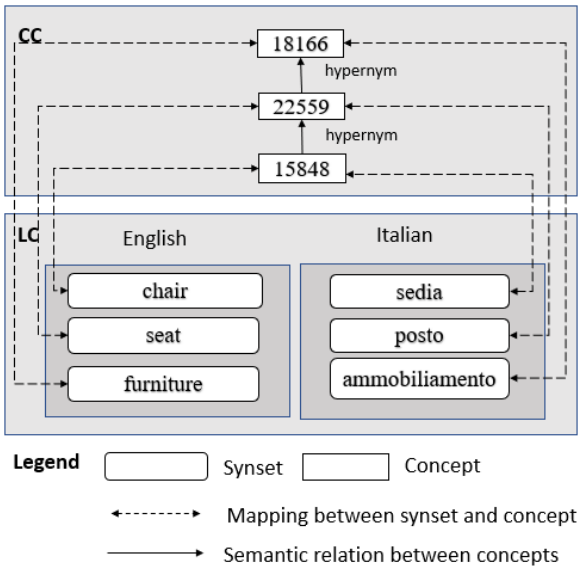

Figure 4: UKC conceptual mappings between English and Italian

ple, the study in (Cristea et al., 2004) highlights the alignment problems between PWN and the Romanian wordnet. Let us consider the three groups of mapping we have identified,

- One to one mapping:

  A synset from IWN has a corresponding synset in PWN and these synsets has one meaning. In Figure 5, the gloss from IWN "जिसने जन्म न लिया हो " (*jisne janm na liya ho*) which means "Who didn't born yet" has one corresponding synset "[unborn]" in PWN. Such type of synsets are those common in both cultures, like "chair".

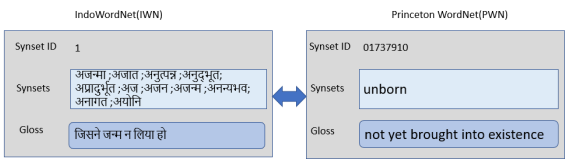

Figure 5: Example for one to one mapping

- Many to one mapping:

  Many synsets from IWN has a corresponding single synset in PWN that has the same meaning. In Figure 6, the glosses "वह स्थान जो पवित्र माना जाता हो " (*vah sthan joh pavitrh mana jatha ho*) and "देव स्थान या पुण्य स्थान " (*dev sthan ya puny sthan*) which mean "A place which is sacred " and "A place which is holy or divine " respectively, have only one corresponding synset "[holy place; sanctum; holy]" in PWN. It means that the two specific concepts in one language are mapped to a general concept in another language.

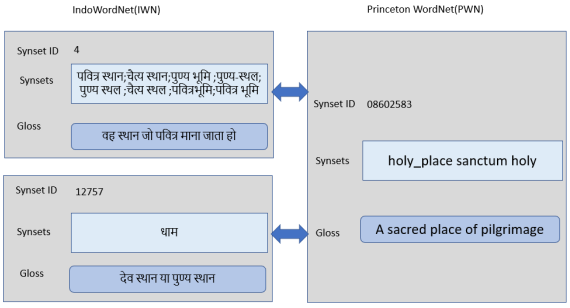

Figure 6: Example for many to one mapping

- One to zero mapping:

  One synset from IWN does not has a corresponding synset in PWN that has the same meaning. In Figure 7 the gloss "मनुष्य के जीवन में अलग–अलग ग्रहों के निश्चित भोगकाल " (*manushy ke jeevan mem alagu-alagu grahom ke nishchith fogkaal*) has no corresponding synset in PWN. The meaning of the gloss is "The period of definite companionship in many planets in human life". This word use when someone having a bad time period in their life and is related to planets in Indian astrology.

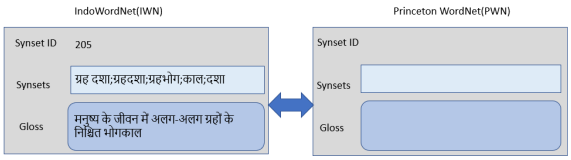

Figure 7: Example for one to zero mapping

The mappings limit IWN to be part of multilingual wordnets. We propose an approach that focuses on concepts that allows to link the languages independently which forms a single resource.

4 Aligning IWN with PWN

Our solution described below can be applied to wordnets of any language. We use the UKC to map the synsets between IWN and PWN that correspond to a single concept. While doing this we define three types of associations between the synsets of IWN and UKC. They are:

- Group A
  One synset from IWN has a corresponding single concept in UKC. These are the IWN synsets that have one to one mapping with PWN.

- Group B
  Many synsets from IWN have a corresponding single concept in UKC. These are the IWN synsets that have many to one mapping with PWN.

- Group C
  One synset from IWN does not have a concept in UKC. These are the IWN synsets that have one to zero mapping with PWN.

Our proposed approach for aligning IWN with PWN is explained below,

1. Set up the UKC
   This step focuses on preparing the UKC for the alignment of IWN with PWN. To take advantage of the PWN hierarchy, the UKC uses synsets from PWN as the concepts. This in turn makes sure the IWN synset aligned with the UKC concepts will associate the corresponding PWN synset. Also, it helps the UKC generating new UKC ids for those IWN synsets which do not correspond to UKC (and therefore) to PWN.

2. Classify the IWN synsets
   Classify the total synsets of IWN based on the association types (A, B and C) mentioned above. This step allows us to know the nature of concepts between IWN and PWN.

3. Process group A synsets
   The group A synsets of IWN are aligned with PWN. Hence it can be imported into the UKC. So the rest of the synsets from IWN could be new concepts for PWN.

4. Process group B synsets
   We analyzed the group B synsets and we found that it is a collection of 454 sub trees. The root element of each sub tree has a corresponding concept in the UKC. An interesting observation is that width and depth of the sub trees could be used to study the nature of lexical gaps between Indian languages and English.

5. Process group C synsets
   We checked to find any synset from group C can be the child to group A synsets. Hence, we found 9,174 synsets are new synsets for PWN and 3021 synsets have no connection with other synsets of IWN.

5 Results

Table I presents the conceptual mappings between IWN and PWN using UKC based on the groups A, B and C. The final alignment between the IWN and the PWN are validated by the linguists. Let us consider the results in detail below,

- Group A
  There are 11,212 group A synsets in IWN and the UKC has corresponding 11,212 concepts. So IWN is imported into the UKC as a new language, Hindi. Figure 8 shows the alignment of the concept “unborn” in UKC. Here, there is a one to one mapping between synset and concept. The concept is linked with synsets of each languages.

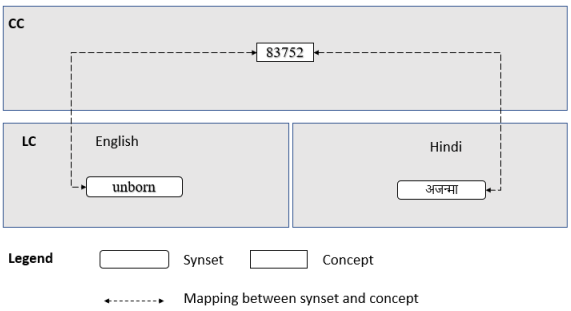

Figure 8: Group A synsets alignment

- Group B
  There are 12,048 group B synsets in IWN. The UKC has corresponding 454 concepts. The remaining 11,594 concepts are new concepts for the UKC. And also these 11,594 synsets are new synsets in the PWN. The research question here is to investigate whether the new identified synsets are lexical gaps or not. We are hoping to study the 454 sub trees and identify the areas resulting the lexical gaps. Figure 9 shows the alignment of the concept “holy place” in the UKC, one concept in CC is linked with one synset from each language. The UKC solves the many to one mapping by adding a new concept which has id -11111.

Table 1: Conceptual mappings between IWN and PWN using UKC

	IWN	UKC		PWN
Groups	#synsets	#concepts	#new concepts	#new synsets
A	11,212	11,212	0	0
B	12,048	454	11,594	11,594
C	12,195	0	9,174	9,174
total	35,455	11,666	20,768	24,290

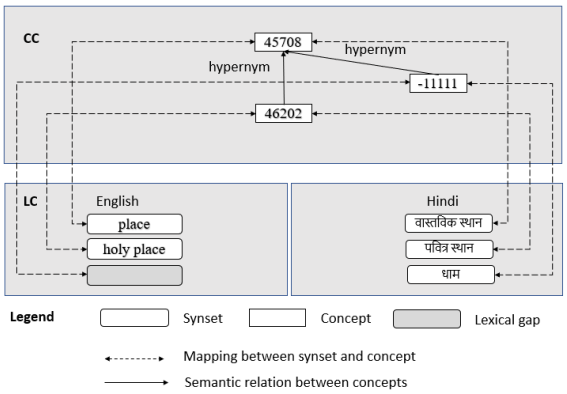

Figure 9: Group B synsets alignment

- Group C
  There are 12,195 group C synsets in IWN. The UKC has no corresponding concepts. So the 9,174 concepts are new for the UKC. Out of these concepts 3021 concepts have no hypernym relations with other 32,434 IWN concepts. Hence, 9,174 synsets are new for the PWN and need to investigate whether they are lexical gaps for the PWN. Figure 10 shows the alignment of a culture specific concept in the UKC. The UKC added a new concept in CC without the hypernymy relation and linked with the languages.

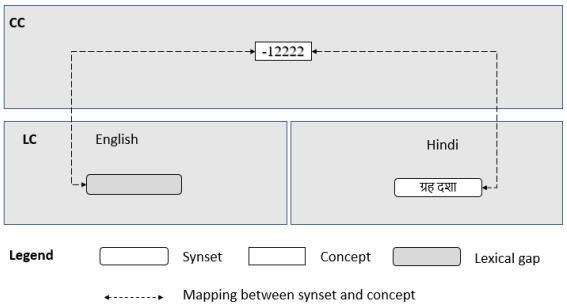

Figure 10: Group C synsets alignment

Like PWN, also in the IWN, various cases of polysemy have been found out(Peters and Peters, 2000). The polysemous 4906 synsets can be either homonymy, specialization polysemy, metonymy, metaphoric polysemy, or compound polysemy (Freihat et al., 2016). However, since this was out of the scope of the project we did not work on this further.

6 Conclusion and Future Work

This paper describes the initial stage of the generation of multilingual resources in a cheaper and faster way. We proposed an approach to align the IndoWordNet, which is the first lexical resource in Indian languages, with the PWN by taking advantage of existing linkages between the IWN and the PWN synsets.However, rather than focusing on the lexicalization problems and polysemy in IWN, we gave full attention to map one synset from IWN to one concept in UKC. The alignment of IWN with the PWN helps to connect more languages. We could integrate as many languages since the UKC forms a semantic network between concepts rather than between synsets of a language. We plan to integrate more Indian languages from IndoWordNet. Fig. 11 sample diagram of expected alignment. In Figure 11, concepts are linked with synsets from languages English and two Indian languages, Malayalam and Hindi.

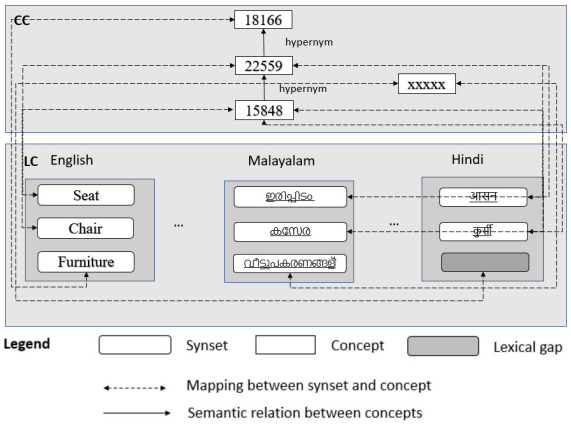

Figure 11: Alignment of the IndoWordNet with the PWN using UKC

Acknowledgments

We thank the University of Trento, Italy for allowing us to be involved in this project and for providing the facilities and full support for the successful completion.

Our heartfelt thanks to Professor Fausto Giunchiglia who gave us his continuous guidance and valuable comments during the project.

The team from Amrita Vishwa Vidyapeetham, India showed full dedication towards participation in the project.

This work has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 810105 (CyCAT)

Our project is a sample work which we hope will motivate many people to take part in the development of under-resourced languages.

References

Valentina Balkova, Andrey Sukhonogov, and Sergey Yablonsky. 2004. Russian wordnet. In *Proceedings of the Second Global Wordnet Conference*.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019. Cognet: a large-scale cognate database. In *Proceedings of ACL 2019, Florence, Italy*.

Gabor Bella, Fausto Giunchiglia, and Fiona McNeill. 2017. Language and domain aware lightweight ontology matching. *Journal of Web Semantics*, 43:1–17.

Gábor Bella, Alessio Zamboni, and Fausto Giunchiglia. 2016. Domain-based sense disambiguation in multilingual structured data. In *The Diversity Workshop at the European Conference on Artificial Intelligence (ECAI)*.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Pushpak Bhattacharyya. 2010. Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Pushpak Bhattacharyya, Christiane Fellbaum, and Piek Vossen. 2010. Principles, construction and application of multilingual wordnets. In *Proceedings of the 5th Global Word Net Conference (Mumbai-India)*.

Debasri Chakrabarti and Pushpak Bhattacharyya. 2004. Creation of english and hindi verb hierarchies and their application to hindi wordnet building and english-hindi mt. In *Proceedings of the Second Global Wordnet Conference, Brno, Czech Republic*. Citeseer.

Debasri Chakrabarti, Vaijayanthi Sarma, and Pushpak Bhattacharyya. 2007. Complex predicates in indian language wordnets. *Lexical Resources and Evaluation Journal*, 40(3-4).

Dan Cristea, Catalin Mihaila, Corina Forascu, Diana Trandabat, Maria Husarciuc, Gabriela Haja, and Oana Postolache. 2004. Mapping princeton wordnet synsets onto romanian wordnet synsets. *Romanian Journal of Information Science and Technology*, 7(1-2):125–145.

Niladri Sekhar Dash, Pushpak Bhattacharyya, and Jyoti D Pawar. 2017. *The WordNet in Indian Languages*. Springer.

Christiane Fellbaum. 2012. Wordnet. *The Encyclopedia of Applied Linguistics*.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2016. A taxonomic classification of wordnet polysemy types. In *8th Global WordNet conference*, pages 105–113.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *IJCAI*, pages 4009–4017.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. 2018. One world–seven thousand languages. In *Proceedings 19th International Conference on Computational Linguistics and Intelligent Text Processing, CiCling2018, 18-24 March 2018*.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India*.

Adam Pease, Christiane Fellbaum, and Piek Vossen. 2008. Building the global wordnet grid. *CIL18*.

Wim Peters and Ivonne Peters. 2000. Lexicalised systematic polysemy in wordnet. In *LREC*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Jaya Saraswati, Rajita Shukla, Ripple P Goyal, and Pushpak Bhattacharyya. 2010. Hindi to english wordnet linkage: Challenges and solutions. In *Proceedings of 3rd IndoWordNet Workshop, International Conference on Natural Language Processing 2010 (ICON 2010)*.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y Ng. 2007. Learning to merge word senses. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)*, pages 1005–1014.

Sebastian Stüker. 2009. *Acoustic modelling for under-resourced languages.* Ph.D. thesis, Karlsruhe Institute of Technology.

Ahmed Tawfik, Fausto Giunchiglia, and Vincenzo Maltese. 2014. A collaborative platform for multilingual ontology development. *World Academy of Science, Engineering and Technology*, 8(12):1.

Piek Vossen. 1998. Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer.