The Dimensions of Lexical Semantic Resource Quality

Hadi Khalilia
University of Trento / Italy
hadi.khalilia@unitn.it

Abed Alhakim Freihat
University of Trento / Italy
abdel.fraihat@gmail.com

Fausto Giunchiglia
University of Trento / Italy
fausto@disi.unitn.it

Abstract

Measuring the quality of lexical-semantic resources is a challenging problem. In this paper, we describe a general approach for quality evaluation in lexical-semantic resources in terms of the quality of their synsets. We also introduce a complete definition for the quality of lexical-semantic resources as a set of synset incorrectness, incompleteness, and connectivity measures that evaluate all synset components. This study demonstrates that synset quality is a summation process that integrates the quality measures of synset components. Furthermore, we then address the main challenges that affect the optimal quality achievement of lexical-semantic resources. Our work, thus, serves to evaluate the quality of monolingual and multilingual lexical-semantic resources and achieves accurate results in natural language processing (NLP) applications.

1 Introduction

A lexical-semantic resource is an organized database of the vocabulary of a language, stores information about the morphemes – the smallest possible unit of a language, such as words and meanings. NLP experts consider lexical-semantic resource the central repository for NLP applications. These resources are categorized into monolingual, which holds mappings between words in a specific language, and multilingual, which has relations across lexical entries in different languages.

In these lexical-semantic resources, synsets operate as foundational elements which follow the principle of relational semantics. Each synset has a unique number and consists of lemmas – a set of synonymous words, a gloss which is a natural language text that describes a synset, and optional examples, which are usually used to clarify the sense of lemmas. For example, the following is a synset:
`#02961779 car, auto, automobile, machine, motorcar: a motor vehicle with four wheels, usually propelled by an internal combustion engine; "he needs a car to get to work".`
The lemmas are `"car, auto, automobile, machine, motorcar"`, the gloss is `"a motor vehicle with four wheels, usually propelled by an internal combustion engine"`, and the synset example is `"he needs a car to get to work"` (Miller et al., 1990). One of its semantic relations is `"a motor vehicle is` **`a hypernym of`** `a car"` whereas `a motor vehicle` is a lemma in this synset:
`#03796768 motor vehicle, automotive vehicle: a self-propelled wheeled vehicle that does not run-on rails.`
This example shows that the construction of synsets needs significant effort and substantial linguistic expertise to establish a correct, coherent, and complete synset and have accurate linguistic relations. Together, the lemmas, gloss, and example qualities form the basics of the synset quality, ensuring the usability that allows NLP

applications to access information stored in PWN without barriers. To achieve high usability for lexical-semantic resources, researchers have developed automatic approaches for measuring a synset quality. While (Jarrar, 2006) implemented a method to ensure synset correctness and define a correct gloss, (Fierdaus et al., 2020) presented an unsupervised learning approach that automatically validates synset lemmas in Indonesian. With reference to synset connectivity, (Freihat et al., 2015) introduced a model to discover and reduce sense-enumeration polysemy, which are wrong relations founded among synsets in PWN (Miller et al., 1990). Using this model, they improved the quality of contents in PWN. However, the quality of lexical-semantic resources remains an understudied subject. There has not yet been a general automatic approach or comprehensive efforts to evaluate synset parts to increase confidence and reliability with a validated resource as well as decrease the consumed time by linguistic experts during manual evaluation.

This paper introduces a notion with eight instructions that measure the quality of a synset by validating its constituent elements and semantic relations together. This study approaches the dimensions of a synset quality and introduces a description of the challenges of overload and underload components in monolingual or multilingual resources.

This article is organized as follows: Section 2 provides background information on lexical-semantic resources and their quality, which are the core of this work; Section 3 discusses related work. In Section 4, we describe our approach for evaluating synset quality and we introduce the main challenges of lexicon quality in Section 5. Finally, our conclusions are outlined in Section 6

2 Lexical Semantic Resources

This section presents a brief background on lexical-semantic resources and their types. We also offer an overview of the necessary notations that researchers use to define the quality of lexical-semantic resources, such as synsets and relations. Furthermore, we show the terms that we utilized to explain synset quality, such as lemmas, gloss, genus, differentia, semantic relations, directed acyclic graph, and others.

Lexical-semantic resource organizes relations between its items based on psycholinguistic principles to present knowledge for linguists and the users of NLP applications (Giunchiglia et al., 2018). Development teams have developed lexical-semantic resources in many ways, which gives each resource a precise interior structure to accommodate a native speaker's needs about the language. A lexical-semantic resource should store at least the following information: words and phrases, parts of speech (noun, verb, adjective, or adverb), the meaning of words with usage examples, and relations between words and phrases (Moustafa, 2014). In general, NLP experts classify lexical-semantic resources into two categories:

1. A monolingual lexical resource is a lexicon that holds mappings between lexemes in a specific language, such as synonymy, polysemy, derivational relatedness, and other mappings. Some Well-known WordNets are PWN ((Miller et al., 1990); (Fellbaum, 1998)), a famous electronic lexical database; linguists and psycholinguists have constructed PWN as a conceptual dictionary based on the principles of the English language. (Mititelu et al., 2016) in Dutch, (Abderrahim et al., 2016) in Arabic, and other monolingual resources.
2. A multilingual lexical resource is a lexicon that contains lexico-semantic relations across lexical entries in different languages. Some widely available multilingual lexical resources are UKC (a high-quality and large-scale lexical resource developed based on psycholinguistic principles for different languages (Moustafa, 2014)), EuroWordNet (Vossen, 1999), BabelNet (Navigli and Ponzetto, 2010), and other multilingual resources.

Both categories for lexical resources include different vocabularies such as nouns, verbs, adjectives, and adverbs. NLP experts and linguistics have grouped synonyms under each type of vocabulary into a set called **synset**. The structure of a synset is organized as follows:

- **Lemmas** synonyms are written as the canonical form of a set of word forms. For example, write is the lemma of the words write , writes , and wrote .
- **Synset gloss** is a natural language text that defines the corresponding lexical concept of the synset, consisting of a genus that corresponds to the classifying property and differentia that corresponds to the distinguishing characteristics of the synset.
- **Synset Examples**: a lexical-semantic resource sometimes, development team enriches synset gloss with sentences as examples to clarify the

shared meaning and show that synonyms are exchangeable in some context.

Synsets connect with other items in a lexical-semantic resource through lexical or semantic relations; forming a network is a directed acyclic graph. In this graph, each node corresponds to a synset, and links represent relations. Lexical links are organized between words, such as the `antonym` that expresses those two senses are opposite in meaning. Semantic relations are used to create mappings between synsets; for example, `the red value of color` , which denotes the source `red` is the value of attribute name `color`.

The quality of a lexical-semantic network is highly dependent on the quality of synset parts and relations among synset pairs. The following section introduces the state-of-art of lexical-semantic quality.

3 Literature Review

Lexical-semantic resources are the basis of natural language processing (NLP) functions, such as disambiguation of word sense, semantic labeling, and question answering. These functions are, in fact, necessary to process and store human semantic knowledge across many languages. Lexical-semantic resources help merge words with their semantic sense to easily and efficiently make the task performance of many applications of NLP, such as machine translation, data integration, and word sense disambiguation.

With the increased efficiency of NLP models developed over time, lexical-semantic resource quality has become a challenging research problem. Content quality is investigated in the literature, and there have been no comprehensive works evaluating lexical-semantic resources completely. For example, (Ramanand and Bhattacharyya, 2007) introduce an automatic validator of WordNet to validate synset synonyms. The system has three phases organized as follows:

1. **Input:** the system reads synset lemmas.
2. **Validation:** applies a set of instructions on inputs using the online dictionary (dictionary.com).
3. **Output:** prints a decision about each lemma by checking whether it fits a synset.

They carried out an experiment on a set of nouns from WordNet, and the results showed that their system was efficient and achieved a good accuracy for tested synsets.

(Purnama et al., 2015) presented a supervised learning approach that automatically validates synset glosses in Indonesian. The strategy utilized a backpropagation feedforward neural network model and decision tree to predict the correctness state of a gloss: accept or reject. Experimental results show that their strategy is effective and achieve an accuracy average near 0.75.

Many researchers have proposed approaches to measure synset relatedness. For example, (Nadig et al., 2008) proposed an approach for hypernymy validation. It is a three-step algorithm that uses Hearst's patterns described in (Hearst, 1992). These patterns are easily recognizable in a text and indicate the lexical relation of interest. The algorithm receives two synsets and then decides whether they have a hypernym-hyponym relationship. As a case study, they carried out an experiment on the synset relations of PWN, and they were able to validate (0.71) of noun synsets in PWN.

Sense enumeration polysemy is inaccurate relation founded between terms and synsets through senses in WordNet. (Freihat et al., 2015) described an approach that discovered this type of semantic relation. They introduced a solution consisting of three stages to solve wrong semantic connections and reduce the high polysemy in compound nouns. As a result, the approach removed the sense enumerations in WordNet and then improved WordNet's quality.

A universal knowledge core is a multilingual lexical resource developed and described by (Moustafa, 2014). This work presented a model to evaluate a concept's incompleteness, which computed how many times a concept existed in a specific language in the resource. They used the model to assess synsets and classify ambiguous words in them.

The literature introduces approaches categorized into three groups: the first focuses on synset correctness by validating lemmas and glosses. The second measures how much lemmas and glosses within synsets in different languages are complete. The last group discusses semantic relatedness to check whether synset connections are correct and complete. These approaches are interpreted to analyze the quality of the components individually. In this paper, we define a general approach that evaluates the quality of the synset parts comprehensively and automatically. Also, we describe the main challenges of lexical-semantic resource quality, such as polysemy and missing lemmas.

4 Defining Synset Quality

Synsets are the foundations of lexical-semantic resources, each expressing a distinct concept. The resources organize the relations between synsets via semantic relations, as mentioned above. A gloss and an example sentence are enclosed in a synset, and semantic linkages with other synsets determine a sense. NLP researchers present the shared meaning of synset lemmas as the most precise meaning for the synset. The accuracy of meaning represents the optimizing value of lexical-semantic resource quality.

In general, each synset inserted in WordNet has a unique ID called SynsetID and is defined in terms of its synonyms, gloss, or semantic relations, as shown in Section 2. For instance, consider a definition of a synset whose `SynsetID: 08283156`.

`#08283156 Table, Tabular Array: a set of data arranged in rows and columns; see table 1.`

`"Table, Tabular Array"` are the lemmas of the given synset, `"a set of data arranged in rows and columns"` is the gloss, and `"see table 1"` is the synset example (Miller et al., 1990). Some semantic relations of the above synset are described in the list below.

1) `Table is a hyponym of table of contents .`
`#06501650` contents, table of contents: a list of divisions (chapters or articles) and the pages on which they start.

2) `Table is a holonym of row ,` and `Row is a meronym of table.`
`#08450457` row: a linear array of numbers, letters, or symbols side by side.

3) `Array is a hypernym of table.`
`#07955622` array: an orderly arrangement; "an array of troops in battle order".

4) `tabular is related to table.`
`#03134301` tabular: of or pertaining to or arranged in table form.

This example suggests that lexical-semantic resource definitions may provide helpful clues as to the gloss `"a set of data arranged in rows and columns"` for validating the synonymy `"Table"` and the example in the synset like `"see table 1"` for verifying the gloss. At the same time, we can use a thesaurus or a dictionary to prove the correctness of the inserted example. Therefore, verifying synset parts indicates that the synset is correct and holds the first dimension for a synset quality. So, we can infer that the **correct synset** is a synset that includes a set of correct elements, correct gloss, and correct examples as follows:

- **Correct lemmas**: synonyms are written as the canonical form of a set of word forms. For example, `go` is the lemma of the words `go` , `goes` , and `went` (Giunchiglia et al., 2017).
- **Correct gloss**: a natural language text that describes the property (genus) of concept and distinguishing characteristics (differentia) of the concept.
- **Correct examples**: contain one or more examples that clarify the exact meaning of the described concept. The synset examples make clear that the concept in `(a)` is about the school as `a building` while the example is about the school as an `institution` in `(b)`.

`(a) school, schoolhouse: a building where young people receive education; the school was built in 1932, he walked to school every morning.`
`(b) school: an educational institution; the school was founded in 1900.`

Furthermore, we introduce that the **complete synset** is a synset with complete lemmas, complete gloss, and complete examples. A definition for each part is described in the following:

- **Complete lemmas**: all expected lemmas of a specific synset should have existed in the synset. There are no missing synonyms from the synset in a specific language.
- **Complete gloss**: a natural language text that includes both parts, genus, and differentia together without a missing. Genus corresponds to the common-key knowledge, both the parent and the child concept express. The differentia is the specific part of the child concept.
- **Complete examples**: this part contain one or more examples that describe the usage of each lemma in the same synset. It can be a phrase or a sentence in a language, e.g., English. The synset examples are complete: if the number of synset lemmas is less than or equal to the number of examples.

In addition, **synset connections** with other items in a lexical-semantic resource should be complete and correct to achieve high quality. Connections can be described as complete if they include at least one instance of the expected semantic relations. With reference to the previous example, we find the synset whose `SynsetID: 08283156` relates to other synsets in WordNet via five relations: `"a table is` **`a hyponym of`** `a table of contents"`, `"a table is` **`a holonym of`** `a row"`, `"a row is` **`a meronym of`** `a table"`, `"an array is` **`a hypernym of`** `a table"`, and `"a tabular is` **`related to`** `a table"`. The given synset is fully connected because it has at least one sample of the expected semantic relations such as `hypernymy(is-a)`, `meronymy(part-of)`, and `related-to`. On the other hand, to confirm the correctness of the synset relations, we can use well-known dictionaries to prove the correctness of the relations.

Our work has adopted the principles of evaluating the synset quality dimensions: correctness, completeness, and connectivity, using the PWN synset as an example. We generalize the expanded approach to other WordNets to consider the interoperability and adoption of all resources. The dimensions of a synset quality are shown in Figure 1.

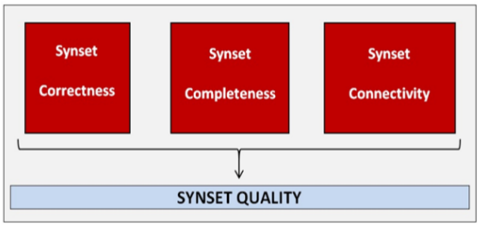

Figure 1: The Dimensions of Synset Quality

5 Lexicon Quality Challenges

Lexical-semantic resource quality has several challenges categorized into two categories: OVERLOAD work or UNDERLOADS, such as inappropriate senses, incorrect lemmas, and faulty connections among synsets, which need extra work. Therefore, they produce OVERLOAD components. On the other hand, missing senses, lemmas, and connections cause an UNDERLOADS problem. The significant challenges of lexicon quality are as the following:

5.1 Polysemy

A lexical-Semantic resource, e.g., WordNet, organizes the relation between terms and synsets through senses. A term may have many meanings, which is called a polysemous term. Polysemy is the ambiguity of a term used in different contexts to express two or more different meanings. Probably, a wrong semantic connection can occur in WordNet. A misconstruction that results in the incorrect assignment of a synset to a term is called `Sense Enumeration` (Freihat et al., 2015). A compound noun contains modifier and modified parts which cause `a compound-noun polysemy`. It generates the incorrect assignment of a semantic relation in a lexical-semantic resource because the modified noun or the modifier is synonymous to its corresponding noun compound and belongs to more than one synset (Freihat, 2014; Kim and Baldwin, 2013). `Specialization polysemy` causes inappropriate relations. For example, a hierarchical relation between the meanings of a polysemous term, when `meaning A` is a more general meaning of a `meaning B` . We should also say that `meaning B` is a more specific meaning of `meaning A` (Freihat et al., 2013b).

5.2 Missing Senses

Despite the highpolysemous nature of WordNet, there is a substantial number of missing senses in WordNet. For example, newly added words in languages cause missing senses for some terms in lexical resources (e.g., WordNet). Such as `crypto mining` sense is missing from the synsets of `mining` term in WordNet (Ciaramita and Johnson, 2003).

5.3 Missing Lemmas

WordNet contains synsets with missing lemmas. For example, the term `brocket` denotes two synsets in WordNet. The lemmas of two synsets are incomplete because they don't include the term `brocket deer` , which is a synonym of the lemmas in `(a)` and `(b)` (Verdezoto and Vieu, 2011).

`(a) brocket: small South American deer with unbranched antlers.`
`(b) brocket: male red deer in its second year.`

5.4 Missing Relations

WordNet organizes relations between synsets, while the substantial number of relationships between synsets remain implicit or sometimes missing, as in the case of synset glosses relations. For example, the relation between `correctness` and `conformity` is implicit and missing, making two synonyms incorrect (Freihat et al., 2013a).

Conclusion

We introduced the notion and the dimensions of synset quality; discussed how much the significance of synset quality affects the quality of the lexical-semantic resource. This paper addressed the main challenges that affect the optimal quality achievement of lexical-semantic resources.

We recommend formalizing the principles of synset quality notion, investigating how much the process of synset quality evaluation can be (semi-) automated. For example, given the formal parts of a synset, such as lemmas, a gloss, examples, and semantic relations can be parsed to know whether a synset has a good quality.

References

Mohammed Alaeddine Abderrahim, Mohammed Dib, Mohammed El-Amine Abderrahim, and Mohammed Amine Chikh. 2016. Semantic indexing of arabic texts for information retrieval system. *International Journal of Speech Technology*, 19(2):229–236.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 168–175.

Christiane Fellbaum. 1998. A semantic network of english verbs. *WordNet: An electronic lexical database*, 3:153–178.

Valentino Rossi Fierdaus, Moch Arif Bijaksana, and Widi Astuti. 2020. Building synonym set for indonesian wordnet using commutative method and hierarchical clustering. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 4(3):778–784.

Abed Alhakim Freihat. 2014. *An organizational approach to the polysemy problem in wordnet*. Ph.D. thesis, University of Trento.

Abed Alhakim Freihat, Fausto Giunchiglia, and Biswanath Dutta. 2013a. Regular polysemy in wordnet and pattern based approach. *International Journal On Advances in Intelligent Systems*, 6.

ABED ALHAKIM Freihat, FAUSTO Giunchiglia, and BISWANATH Dutta. 2013b. Solving specialization polysemy in wordnet. *International Journal of Computational Linguistics and Applications*, 4(1):29.

Abed Alhkaim Freihat, Biswanath Dutta, and Fausto Giunchiglia. 2015. Compound noun polysemy and sense enumeration in wordnet. In *Proceedings of the 7th International Conference on Information, Process, and Knowledge Management (eKNOW)*, pages 166–171.

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*.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Coling 1992 volume 2: The 15th international conference on computational linguistics*.

Mustafa Jarrar. 2006. Position paper: towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering. In *Proceedings of the 15th international conference on World Wide Web*, pages 497–503.

Su Nam Kim and Timothy Baldwin. 2013. Word sense and semantic relations in noun compounds. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):1–17.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Verginica Barbu Mititelu, Corina Forăscu, Christiane Fellbaum, and Piek Vossen. 2016. Global wordnet conference.

Ahmed Maher Ahmed Tawfik Moustafa. 2014. *A collaborative Platform for multilingual Ontology Development*. Ph.D. thesis, University of Trento.

Raghuvar Nadig, J Ramanand, and Pushpak Bhattacharyya. 2008. Automatic evaluation of wordnet synonyms and hypernyms. In *Proceedings of ICON-2008: 6th International Conference on Natural Language Processing*, volume 831. Citeseer.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.

I Purnama, Mochamad Hariadi, et al. 2015. Supervised learning indonesian gloss acquisition. *IAENG International Journal of Computer Science*, 42(4).

J Ramanand and Pushpak Bhattacharyya. 2007. Towards automatic evaluation of wordnet synsets. *GWC 2008*, page 360.

Nervo Verdezoto and Laure Vieu. 2011. Towards semi-automatic methods for improving wordnet. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

PJTM Vossen. 1999. Eurowordnet.