Enhancing Lexical Resources: Synset Expansion and Cross-Linking Between ItalWordNet and MariTerm

Lucia Galiero[1], **Federico Boschetti**[2], **Riccardo Del Gratta**[2], **Angelo Mario Del Grosso**[2], **Monica Monachini**[2]
[1]Università di Bologna, Forlì, Italy
[2]Istituto di Linguistica Computazionale "Antonio Zampolli", Consiglio Nazionale delle Ricerche (ILC-CNR), Pisa, Italy
lucia.galiero@studio.unibo.it
{federico.boschetti, riccardo.delgratta, angelomario.delgrosso, monica.monachini}@ilc.cnr.it

Abstract

This paper outlines the first operation towards a full update of MariTerm, a WordNet-like resource on maritime terminology developed and maintained by CNR-ILC, in preparation for future compliance with FAIR principles (Wilkinson et al., 2016). The project focused on expanding and linking synsets between ItalWordNet (IWN), a general lexical database for Italian, and MariTerm to enrich IWN with maritime concepts. A semi-automatic pipeline was developed to facilitate this process, prioritizing critical semantic relations and automatic evaluation. Key outcomes include an enriched ItalWordNet with links to MariTerm concepts and a revised MariTerm with connections to IWN synsets. While further refinement is needed, this work marks a significant step toward integrating maritime terminology into ItalWordNet.

1 Introduction

In the last twenty years, research best practices in Digital Humanities have integrated the effective organization and representation of knowledge with the need of making data available for long-term preservation and re-use. In this regard, leveraging legacy resources can contribute to save them from being consigned to oblivion, but also enriches modern resources (Frontini et al., 2016).

The project hereafter described takes measures in this direction with a work involving two key lexical resources for the Italian language: ItalWordNet[1], a comprehensive and generalized lexical database, and MariTerm (Marinelli and Spadoni, 2007), a specialized resource for maritime terminology.

As for the task, the main objectives of the project were the expansion of ItalWordNet, and the linking of relevant synsets from ItalWordNet to MariTerm and vice versa. For further clarification, the concept of "expansion" here refers to updating ItalWordNet synsets with missing semantic information (mostly critical semantic relations and synsets) from MariTerm. The term of "linking" should instead be intended the creation of systematic links between related synsets across the two resources.

The remainder of this contribution is organized as follows: Section 2 illustrates previous work, Section 3 provides a brief overview of the lexical resources involved, Section 4 will delve into the major setbacks encountered, Section 5 will discuss the implemented approach, and finally Section 6 will discuss and summarize results, as well as possible advancements of the present work and specifications on distribution of produced data.

2 Related Work

Previous work in linking a specialized resource to a general WordNet has been provided with the case of the GeoNames ontology (Frontini et al., 2016). The resource was originally issued in English and was already available as Linked Open Data (LOD) in RDF. Its content was later integrated to Princeton WordNet 3.0 (PWN) (Fellbaum, 1998) and to IWN. The most recent LOD WordNet resource for Italian provided by the ILC-CNR is represented by

[1] See for more: https://www.ilc.cnr.it/progetti/italwordnet-2/

IWN 2.0[2], which was created in compliance with the WordNet 2.0. specifications[3].

Another case for a cross-linked lexical resource is provided by Ancient Greek WordNet, (Bizzoni et al., 2014), whose content has also been released in LOD and matched English-Greek concepts have been mapped to PWN. Within the creation of AGWN, PWN has also been implemented as a pivot to link Greek concepts to several other languages covered by the project, mainly Croatian and Latin. This cross-lingual linking model that implements PWN as pivot network is also known as MultiWordNet (MWN) (Pianta et al., 2002). However, as we will illustrate in the upcoming paragraph, ItalWordNet builds relations with other WordNets through another model.

3 The Resources

3.1 ItalWordNet

ItalWordNet (Roventini et al., 1998), was initially developed the late 1990s and the early 2000s as part of the EuroWordNet project (Vossen, 1998)[4], which created WordNets for several European languages, and the Italian national project SI-TAL[5]. IWN organizes Italian words into synsets, i.e., sets of groups of synonyms that share a common meaning and are interchangeable in certain contexts, capturing internal lexical-semantic relationships. In all its versions, IWN is designed to align with Princeton WordNet, and has been upgraded and enlarged over time (Niero, 2006; Bocco et al., 2003). Nevertheless, IWN has always been developed independently form other WordNets, resulting in different sets of semantic relations and in potential loss of information when converting the resource to new formats (Quochi et al., 2017).

3.2 MariTerm

Developed in the mid-2000s in collaboration with the Port of Livorno, MariTerm is a lexical-semantic database focusing on maritime terminology, based on the Word-Net model. While it perfectly mirrors IWN's WordNet-like structure, the MariTerm ontology maps maritime terms to their specific concepts of nautical science and maritime transport. Although lacking the extensive coverage of generalized lexicon seen in IWN, it represents an invaluable resource for specialized terminology.

3.3 Resource architecture

Both ItalWordNet and MariTerm can be marshalled in eXtensible Markup Language, and their structures are both based on the EuroWordNet model (Vossen, 1998). Specifically, alignment to EWN for multilingual application is achieved via the InterLingual Index (ILI). Thus, these two resources for Italian are not directly aligned to the PWN, as it was in the case of AGWN (Bizzoni et al., 2014) and GeoNamesWordNet (Frontini et.al., 2016). Nonetheless, the shared overlapping representation of synsets in both IWN and MariTerm ensures rich semantic connectivity, facilitating the cross-resource linking. **Error! Reference source not found.** summarizes the content and the extension of all resources involved[6]. It should also be noted that, at the time of the writing, both original resources were available in XML format but did not comply to any of the Global Wordnet standards[7]. Moreover, since MariTerm does not have an updated version in LOD to date (see Section 5 for more), it was decided to implement a version of IWN that was not converted to a LOD representation for the sake of suitability. Regarding the scope of this work, neither the resulting resources have been updated to the aforementioned standards in order to reflect the structure of the originals.

[2] Available at: http://hdl.handle.net/20.500.11752/ILC-66.

[3] WN 2.0. Specifications available at https://www.w3.org/2006/03/wn/wn20/

[4] More info also at: https://archive.illc.uva.nl/EuroWordNet/

[5] More at: https://www.ilc.cnr.it/progetti/tal-2/

[6] Both original resources contained duplicate synsets. Specifically, 809 were found in IWN and 27 in Mariterm, and were identified by means of the same numerical ID and first word form listed. Numbers remained consistent through all the phases of the work presented in this contribution, and no significant information was lost. Although these duplicates were retained in all resources, the numbers in Table 1 and Table 3 reflect only the unique items, excluding duplicates.

[7] See for more: https://globalwordnet.github.io/schemas/

4 Major Setbacks

The overarching structure of both IWN and MariTerm showed total overlap. However, this shared feature did not suffice from the earliest stages of developing a pipeline, and other potential issues gradually accumulated.

One key issue was that synsets in each resource had different unique identifiers, making a direct alignment impossible. Furthermore, even in cases where lemma (word form) and sense attributes matched for a couple of synsets, semantic relations often diverged, leading to inconsistencies in meaning. This other type of mismatch made it even more impractical to rely solely on these attributes for updating and linking the two resources. For instance, it was observed that each of the matched synsets for the word “navigare” (EN: “*to sail*”) corresponds to different sense numbers. This might be a possible result of IWN and MariTerm being developed as independent resources, but also a result of the word “navigare” (EN: “*to sail*”) having multiple sense variants in both lexical databases.

Further ambiguity was introduced by the presence of multiple lemmas within a single synset, a common feature in both resources but that could lead to redundant or wrong alignments altogether.

Most importantly, neither resource seemed to feature a suitable attribute or section for cross-resource linking. Notably, the MariTerm introductory paper (Marinelli and Spadoni, 2007) stated the presence of a section for cross-resource linking, which was still nowhere to be found upon further inspection. Specifically, the links provided in such sections were represented by plug-in relations, a type of semantic relations that connects synsets across specialized and general WordNets (Niero, 2006). In their turn, plug-in relations can be identified as upward (if linking a specific term to a general one), downward (if vice versa) and horizontal (if connecting synsets by any other type of relation). The two former kinds of links were the most used for this work due to the semantic relations that were chosen as object of focus.

On a minor note, discrepancies in definitions between corresponding synsets increased the challenges. In fact, some synset couples displayed different definitions, which posed a problem as to which definition should be used in the final output files. In other cases, definitions were missing from at least one of the two resources, with the risk of losing crucial information for synset comparison and evaluation. One example for actual definition discrepancy comes from the synsets for the word “pagaia” (EN: *“paddle”*):

- IWN definition: “remo a pala larga con il quale si voga senza appoggiarlo alla falchetta.”

  (EN: “wide-bladed oar with which one rows without resting it on the sickle.”)

- MariTerm definition: “remo con le due estremità a pala che si maneggia tenendolo al centro con entrambe le mani si usa sulle canoe e su altri natanti di tipo fluviale o balneare”.

  (EN: “paddle-like ends that is handled by holding it in the middle with both hands – it is used on canoes and other watercraft of the river or bathing type”)

Following is instead a case for missing definition from a resource, taken from the synset for the word “azimut” (EN: “*azimuth*”):

- IWN definition: *missing*

- MariTerm definition: “nella navigazione astronomica indica l’arco di cerchio compreso tra il nord e la verticale dell astro stesso.”

  (EN: ”in astronomical navigation, the arc of a circle between north and the vertical of the star itself.”)

5. Expansion and Linking stages

This work focused on critical semantic relations such as near-synonymy, hyponymy, hyperonymy, and their variants for synsets belonging to different parts of speech (i.e., “xpos_near_synonym” etc.).

5.1 Preliminary candidate extraction

The expansion and linking process we developed consisted of two main phases, each implemented in a dedicated Jupyter Notebooks using Python[8]. The

[8] Notebooks and data available at: https://drive.google.com/drive/folders/1mJLIS16qRkAp8UobGEkoxtSfsq8-p1o6

first one focused on extracting shared synsets as follows: extraction of all synsets from both XML files, identification of shared ones, and saving the output in a CSV file for further analysis.

Within this context, the first problem that was solved was the presence of multiple lemmas inside synsets across resources. Specifically, the Python script was designed to match synsets only if the first lemma listed was the same, and such a logic would be implemented for this and all subsequent portions of the pipeline[9].

5.2 Scoring, matching synsets and expanding IWN

5.2.1. The similarity score framework

The second phase centred on the expansion and linking process. Once the transitory resource with preliminary candidates was refined, it was necessary to build a scoring metric between synsets across resources to ensure that only the best match of synsets was picked up, thus identifying and preventing incorrect updates and ambiguities from early on.

The first step towards calculating similarity for synset pairs was vectorization, which was applied to all definitions inside the datasets, including both definitions for a given main synset and its possible target synsets. Albeit mainly used on large collections of texts rather than lexical databases, the TF-IDF approach still yielded consistent results in a time-efficient way, due to it being applied only to definitions. Without looking at the data and assuming that all synsets and target words linked to a semantic relation featured a definition in both resources, the amount of sentences reaches up to a potential 200.000 total sentences of different lengths. Definition redundancy, in a way, provides another major advantage for the use of TF-IDF. As a matter of fact, definitions for target words tend generally to be the same as the one for the synset they point to. Such consistency allows TF-IDF to better capture uniqueness of words across the definition pool, leading to more nuanced results.

The resulting vectors were then compared via cosine similarity. The rest of the scoring metric is based on two main calculations: similarity of synset definitions and weighted relation similarity. The former compares the definitions (also known as “glosses”) of synsets with the same lemma in both resources. In its own regard, relation similarity compares the definitions of target synsets in both databases, multiplying scores obtained by target synset definitions by a weight that reflects the importance of the semantic relation (e.g., hyperonymy, hyponymy). Weights for hyperonymy and hyponymy amount to 0.82 and 0.7 respectively, with both values being based on an existing work by Tülü et al. (2019) which assigned weights for semantic relations inside WordNet 3.0. On the other hand, weights for all other relations at the core of this project were not presented in the work described by Tülü et al. (2019), probably due to IWN inheriting all its semantic relations from EuroWordNet, thus presenting a notable structural difference with WordNet 3.0. Therefore, weights for relations that were not mentioned by Tülü et al. (2019) were manually assigned, with near synonymy being awarded 0.6. All others were given a baseline of 0.5. The abovementioned structural differences with WordNet 3.0 present an argument as to why advanced pipelines for automated weight assignment for WordNet semantic relations like SemSpace (Ohran and Tülü, 2021) were deemed unsuitable for the extent of the present work.

The scores for definition similarity and weighted relation similarity are then summed together. Shared relations between synsets were awarded bonuses, while relations that were present in MariTerm, but missing in IWN, were penalized. This tailored method allowed for more accurate and precise alignment between the two lexical resources.

5.2.2. Synset update pipeline

As soon as the score computation was complete and its detailed breakdown saved in another CSV file, the expansion process began by selecting lemmas inside the spreadsheet that met certain similarity criteria. Subsequently, synsets are expanded with missing semantic relations.

To account for missing definition in synsets, a fallback mechanism was implemented, where a fallback definition is retrieved by examining in a predefined order the semantic relations associated with the given synset without gloss.

In case of discrepancies between glosses were detected (i.e., where glosses across resources vary

[9] A more refined approach for future work could involve checking for intersections between resources, e.g. using more shared lemmas to retrieve similar synset pairs, thus improving the accuracy and reliability of synset linkages.

for a synset or a target word) any definition from MariTerm inside IWN was replaced with the gloss from IWN where applicable. In the special case where the synset from IWN was the only one not having a definition, it inherited the gloss from MariTerm. In all other cases, no gloss substitution was carried out. Moreover, if a missing semantic relation pointed to a target word that did not feature its own synset inside of IWN, the automatic process ensures the creation of a new synset with a new identifier.

Finally, a key feature of all WordNet-like resources is consistency, where connections among WordNet synsets are reciprocal (i.e., if a synset shows an hyponymy relation to a certain target synset, the latter will contain, in its dedicated entry, a hyperonymy relation pointing to the first synset). These connections were also considered while designing the Python script and are consequently processed to maintain consistency within IWN.

After all expansions were applied, the system re-orders synsets for better navigation.

5.3 Post-expansion modifications

5.3.1 Manual edits and further gloss inconsistencies

Before proceeding with the linking, a manual check was carried out to validate the newest updates. Out of the approximately 400 synsets involved in the update, several modifications addressed issues in 20 synsets that were either incorrectly aligned or missed during the expansion process. This stage implied manually inserting or creating new synsets and resolving inconsistencies in definitions and semantic relations. Roughly 100 of the new synsets also had a numerical ID that was already taken from other synsets, and new numeric identifiers were checked and inserted manually. Moreover, 241 of the newly added synsets still presented gloss inconsistencies, where a synset displayed a definition when listed as a target synset, but not in its main entry. A small, dedicated script in Python was designed to resolve such conflict.

5.4. Linking stage

The linking step constituted the final stage of the whole pipeline. Since the original resources did not feature a suitable section or attribute to insert cross-resource links, the most viable solution seemed to be the creation of a custom node. This would be designed to store cross-resource connections (or plug-in relations) and would have the same structure of the section connected to PWN via the ILI index. Such an arrangement seemed to be the most suitable since the section connected to the ILI is defined inside of both resources and shows the same structure in each database.

The MariTerm introductory paper (Marinelli and Spadoni, 2007) described the node as being situated right between the internal links among resource synsets and the external links between the WordNet database and the ILI. Eventually, it was decided to mirror the description provided by Marinelli and Spadoni (2007) and place the node accordingly.

As for the linking process per se, an automatic mechanism for updating the plug-in nodes within the structure of the files was crucial. It started by clearing any existing plug-in nodes for each synset to ensure a clean state. For each synset, it matched corresponding synsets from MariTerm and IWN, creating new nodes for plug-in relations. Specifically, the nodes were populated with the newly added relations, which were slightly edited by the addition of a “plug-”prefix to differentiate them from ordinary semantic relations. Another kind of relation that populated the nodes is represented by “plug-synonym” relations. As described by Niero (2006), these are used for “overlapping synsets, i.e., synsets that have a similar meaning albeit belonging to different databases” (Niero, 2006).

During the linking process, all plug-relations were tracked to avoid duplicates, and any synset without a match was given an empty node to maintain structural consistency. Finally, the output files were saved with properly formatted entries, ensuring the newly integrated data was well organized and ready for future use.

6. Discussion

6.1 Results

Table 2 illustrates breakdown of results yielded by the semi-automated pipeline, both for the IWN synset expansion and cross-resource linking. In its turn, Table 3 provides insight into the content of the updated IWN versions, before and after manual post-hoc modifications. The final IWN version produced by the pipeline is intended to be the one that underwent manual changes and was enriched with the new synsets and the links to MariTerm concepts.

Data involved	Count
MariTerm > IWN expansion	
Preliminary candidate synsets	1157
Identified synset matches	747
IWN Word meanings updated	397
New synsets created in IWN	363
New Relations in IWN	1160
IWN <> MariTerm Linking	
IWN synsets linked to MariTerm	742
New IWN relations linked to MariTerm	843
MariTerm synsets linked to IWN	751
MariTerm relations linked to IWN	0

Table 1 - Results of the whole pipeline, divided by section

Resource	Synsets	Lemmas	Internal Relations
Automated, unedited IWN	49482	46425	133004
Post-hoc edited IWN	49477	46423	132983

Table 2 - Expanded versions of IWN (before and after manual edits)

The pipeline successfully matched 750 synsets from MariTerm into IWN. Of these, 397 ItalWordNet entries were successfully updated, resulting in the addition of 1160 missing relations and 363 new synsets. The lower number of updated entries should not be seen as discouraging, since the automatic process purposedly skips all matches that had no penalties, therefore needing no update whatsoever. As for the linking part, 742 IWN synsets were mapped as plug-synonyms to MariTerm, and 848 out of the 1160 new relations were linked back to the corresponding MariTerm synsets.

On the other hand, the mapping of IWN matched synsets and relations inside MariTerm was not as successful. While a total of 751 of synsets contained a very simple link to ItalWordNet, the updated file contains 1085 relations for plug-synonymy. The consequent implication is that some of the plug-synonyms are duplicates that refer to homonyms. Moreover, out of the 843 relations that were linked from ItalWordNet to MariTerm, the reversed linking did not work for any of the relations added to ItalWordNet. Such an outcome calls for urgent improvements in revising and automating the linking pipeline.

Theoretically, as described by Niero (2006), once the links are established between two overlapping synsets by means of plug-synonymy or plug-near synonymy, the following step is the creation of a new synset inheriting the hypernymy relations from the general lexical resource, with the hyponymy relations and synonyms being passed down from the specialized database. Since the creation of dedicated nodes provided a base for adding plug-ins, the creation of new synsets with these features represents a possible future advancement in enhancing the two resources.

Moreover, the mapping of IWN concepts inside MariTerm faced challenges, such as duplicate plug-synonym relations, suggesting that the automated pipeline needs refinement to better handle term overlap and ambiguity. Additionally, the manual evaluation of the new synsets inside IWN highlights the need for further improvement of the semi-automated pipeline.

6.2 Future advancements

Applications based on the proposed method are developable provided that two (or more) WordNet-based resources share the same representation, sets of semantic relations and same format standards, like in the described case.

Once the issues with the cross linking in the produced resources are resolved, the updated IWN and MariTerm could largely benefit from being fully converted in LOD format (e.g., RDF) or even by following the OnotoLex lemon format (McCrae et al., 2017). Given how Italian WordNet-like resources present potential issues for cross-lingual mapping of concepts due to their alignment to the relatively old EWN model, a future parallel advancement for long term interoperability with PWN could even consider a full conversion of LOD MariTerm to Open Multilingual WordNet, following the steps described in the work done by Quochi et al. (2017).

All in all, legacy resources like MariTerm have the potential to be integrated into broader frameworks like ItalWordNet, ensuring both the preservation and active use of specialized terminology. However, the degree of complexity of these kinds of resources raises important questions as to how much automation should be implemented to process, create and innovate WordNets.

6.3 Distribution

The data contained in the updated MariTerm with the plug-in extensions is currently available on the CLARIN4ILC repository at: `http://hdl.handle.net/20.500.11752/OPEN-1034`, where it will be stored for the long term. It is advisable to check the page regularly as license specifications and other metadata will be updated soon.

References

Yuri Bizzoni, Federico Boschetti, Harry Diakoff, Riccardo Del Gratta, Monica Monachini, and Gregory Crane. 2014. The Making of Ancient Greek WordNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), pages 1140–1147, Reykjavik, Iceland. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1071_Paper.pdf

Andrea Bocco, Luisa Bentivogli, and Emanuele Pianta, 2003. ArchiWordNet: Integrating WordNet with Domain-Specific Knowledge. In *Proceedings of the Second International WordNet Conference*. Pages 39-46, Brno, Czech Republic https://hdl.handle.net/11583/1917825

Christiane Fellbaum. 1998. *WordNet, an electronic lexical database.* MIT Press, Cambridge, Massachussets

Francesca Frontini, Riccardo Del Gratta and Monica Monachini 2016. GeoDomainWordNet: Linking the Geonames ontology to WordNet. In *Lecture notes in computer science*, pages 299-233. https://doi.org/10.1007/978-3-319-43808-5_18

Rita Marinelli and Giovanni Spadoni, 2007. Modeling a Maritime Domain Ontology, In *Tenth International Symposium on Social Communication,* pages 511–515, Santiago de Cuba, Cuba

John P. McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference*, pages 19-21.

Federica Niero. 2006. WordNet e sue applicazioni. Revisione e implementazione di un database di termini matematici [BSc. Thesis, Università degli Studi di Padova] pages 73-77. https://www.math.unipd.it/~laurap/grupponlp/TesiNieroFederica.pdf,

Umut Orhan, and Cagatay Neftali Tülü. 2021. A novel embedding approach to learn word vectors by weighting semantic relations: SemSpace. *In Expert Systems With Applications, 180*. Pages 1-3, 5-7 https://doi.org/10.1016/j.eswa.2021.115146

Emanuele Pianta, Luisa Bentivogli and Christian Girardi. 2002. *MultiWordNet: developing an aligned multilingual database*. In *Proceedings of the First International WordNet Conference* , pages 293-302. Mysore, India https://hdl.handle.net/11582/499

Valeria Quochi, Roberto Bartolini, and Monica Monachini. 2017. ‘ItalwordNet goes open´. LiLT, Vol. 10, Issue 4l

Adriana Roventini, Antonietta Alonge, Nicoletta Cazlolari, Bernardo Mangini, and Francesca Bretagna. 1998. ItalWordNet: Building a large semantic database for the automatic treatment of Italian. *In Linguistica Computazionale*, (XVIII-XIX), pages 745-791 https://hdl.handle.net/11582/2033

Cagatay Neftali Tülü, Umut Orhan, and Ehran Turan. 2019. Semantic Relation’s Weight Determination on a Graph Based WordNet. In *Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi.* pages 67, 75-76 https://doi.org/10.17714/gumusfenbil.432582

Piek Vossen. 1998. *EuroWordNet: A multilingual database with lexical semantic networks.* Kluwer Academic

Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Wiillem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18