Inferences for Lexical Semantic Resource Building with Less Supervision

Nadia Bebeshina-Clairet, Mathieu Lafourcade

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2300–2305
Marseille, 11–16 May 2020

Inferences for Lexical Semantic Resource Building with Less Supervision

Nadia Bebeshina[1,2], **Mathieu Lafourcade**[2]
[1]Praxiling, [2]LIRMM
Université Paul Valéry, Université de Montpellier, France
nadia.clairet@gmail.com, lafourcade@lirmm.fr

Abstract

Lexical semantic resources may be built using various approaches such as extraction from corpora, integration of the relevant pieces of knowledge from the pre-existing knowledge resources, and endogenous inference. Each of these techniques needs human supervision in order to deal with the potential errors, mapping difficulties or inferred candidate validation. We detail how various inference processes can be employed for lexical semantic resource building with less supervision. Our experience is based on the combination of different inference techniques for multilingual resource building and evaluation.

Keywords: inference, multilingual knowledge resource, semantic relations, validation

1. Introduction

The lexical and semantic resource building based on inference represents an appealing area of research. Inference consists in proposing new elements automatically based on the existing ones. In the context of a lexical semantic resource, inference is a process of calculating new pieces of knowledge i.e. semantic relations without using any external structured knowledge repository. Thus, an inference engine can be seen as a system that, given a set of rules, provides the action of creating the new elements in the resource. Thus, such this mechanism allows building lexical semantic resource incrementally and, presumably, minimise the human effort necessary for the resource building. Unfortunately, the knowledge parts obtained by inference have to be semi-automatically double checked to avoid the propagation of errors. The part of human effort in the inference process may be very important to perform rule design, validation, resource improvement.

In the present paper, we describe a resource building pipeline based on monolingual inference and cross-lingual inference. Its main contribution is proposing a method where the resource building process appears as a self learning process as the feedback of the evaluation by inference impacts the resource that is being built.

The paper is organised as follows. First, we introduce the state of the art techniques for lexical semantic resource building. Second, we detail the resources we use for the inference based resource building. Third, we provide the details about the resource building and evaluation experience.

2. Related Work

Cross-lingual relationship inference benefits from active research efforts. In the framework of the large knowledge bases (KBs) such as NELL (Mitchell et al., 2015), several approaches focused on the equivalence between entities and relationships have been introduced. For instance, authors in (Hernández-González et al., 2017) describe the experience of merging several monolingual editions of NELL. Authors in (Nickel et al., 2015) detail the statistical relational learning on knowledge graphs (KGs) and point out the importance of type constraints and transitivity. Similar to (Wang et al., 2015), they base their method mainly on large scale KBs such as Nell (Carlson et al., 2010), KnowItAll (Etzioni et al., 2005), YAGO (Rebele et al., 2016) or DeepDive (Shin et al., 2015).

Consistent research efforts have been done to extend finer grained lexical semantic resource models such as WordNet to build resources in other languages including resource poor languages such as Basque as proposed by (Agirre et al., 2002) and many others. These authors stress the importance of concept-to-concept and word-to-word mapping while building a lexical semantic resource. The BabelNet project (Navigli and et al., 2012) has been the first large-scale experiment of combining different manually crafted resources and models with unsupervised resource building techniques.

The endogenous rule-based inference process has been studied by (Zarrouk, 2015) and (Ramadier, 2016) in the framework of the RezoJDM, the LSN for French. Their methods rely on the relationships and relationship meta-information that are already present in this LSN in order to propose the new ones according to one of the following schemes: deduction and induction based on taxonomy, abduction based on semantic similarity, and inference by refinement. (Gelbukh, 2018) introduced a comparable inference mechanism to enrich a collocationnal knowledge base by suggesting new collocations through the inference by abduction where semantic similarity is calculated on the basis of WordNet (Fellbaum, 1998).

3. Resources

3.1. RezoJDM

The **RezoJDM** (Lafourcade, 2007) is a lexical semantic network (LSN) for French built using crowd-sourcing methods and, in particular, games with a purpose (GWAPS)

such as JeuxDeMots[1] and additional games [2]. This general purpose commons sense network has been built since 2007. This resource is a directed, typed and weighted graph. At the time of our writing, RezoJDM contains 3.7 millions of terms that are modelled as nodes of the graph and 290 millions of relations (arcs).

3.2. Multilingual Lexical Semantic Network (MLSN)

The **MLSN** (Bebeshina-Clairet, 2019) is a multilingual LSN with an interlingual pivot which contains French, English, Spanish, and Russian sub-parts. It was built for the cuisine and nutrition domain but also includes pieces of general knowledge as per the non-separability between the general and the domain specific knowledge verified by (Ramadier, 2016). The MLSN is a directed, typed, and valuated graph. It contains k sub-graphs corresponding to each of the k languages it covers and a specific sub-graph, the *interlingual pivot*. The MLSN relies on a term (node) set T and a relation (arc) set R. MLSN relations are typed, weighed, and directed arcs. The MLSN nodes may correspond to one of the following types :

1. lexical items (i.e. *garlic*) ;
2. interlingual items (pertaining to the interlingual pivot, they are also called *covering terms*) that are not necessarily labelled in a human readable way;
3. relational items (i.e. relationship reifications such as *salad[r_has_part]garlic*);
4. category items modelling categories, parts of speech or other morpho-syntactic features (i.e. *Verb:Present*, *Noun:AccusativeCase*).

4. Experiment

4.1. Overview

Our experiment is targeted at the MLSN building. It has been guided by the following observations :

- in some languages some semantic information can be captured through the syntactic and grammatical features whereas in some others such information lays "deeper" and needs more complex semantic analysis to be discovered;
- in the framework of a multilingual lexical semantic resource, one language part of the resource may provide missing semantic information to its other language parts;
- given the interoperability of two LSNs, one resource can be validated against another one in terms of presence or absence of real or inferrable semantic knowledge.

According to those observations, we hypothesise that the semantic relations extracted from corpora in a syntactic feature rich language such as Russian can be used to enrich a semantic resource in other language given a link between these resources (i.e. interlingual or natural pivot). The interoperability of the MLSN with a richer and stable resource RezoJDM provided us with the automatic evaluation procedure. The feedback of this procedure drives the overall experiment close to the self learning process.
The experiment involves three basic steps:

1. Semantic relation identification and extraction from a POS-tagged Russian corpus using a set of rules;
2. cross-lingual inference of the extracted relations that creates new relations in the French sub-graph of the MLSN based on the relations now available in the Russian sub-graph;
3. Validation of the relations inferred cross-lingually against the RezoJDM LSN. Self-adjustment of the MLSN: negative valuation of the relations (inference chains) that contributed to the wrong or undecidable inference result.

4.2. Extraction

The extraction has been performed on a corpus of cooking instructions (2 473 654 words) collected on the Web[3]. The procedural language of cooking recipes gave us the opportunity to spotlight a set of relation types that may be difficult to infer monolingually in other language MLSN subgraphs than the Rusian subgraph as well in the pivot. This corpus has been pre-processed using the most recent version of the Russian Malt parser (Sharoff and Nivre, 2012).

The Russian language structure allows defining rules to extract predicate-argument information from procedural texts such as cooking instructions. We targeted the extraction part of the experiment on the following semantic relation types:

- *characteristic* with some distinction between composition-based characteristics (**raspberry** jam), characteristics describing a state or a transformation (**ground** cashew), and qualitative characteristics (**juicy** orange) as, to some extent, such distinction is observable in Russian;
- *manner* with some distinctions observed in the corpus between the instrument-based adverbial phrases (decorate *(how?)* **with sliced fruits**, part-whole manner (cut **into cubes**) and other realisations;
- *place* and, in particular, place of action with the distinction between **surface** and **container** places.

Additionally, typical successor relations have been extracted.
We designed a set of 33 rules to perform the relation extraction from text. These rules are not exhaustive. The premises of the rules are the syntactic and grammatical features such as noun cases coupled with functor

[1] http://www.jeuxdemots.org
[2] http://imaginat.name/JDM/Page_Liens_JDMv2.html
[3] Mainly from https://www.gotovim.ru/

words, the conclusion is the creation of a candidate relation.

`if X Y` (description of the context)
`and X Verb` (first premise)
`and Y Nom:CaseAccusative` (second premise)
`X r_object Y` (conclusion)

`if X`"na" `Y` (description of the context)
`and X Verb` (first premise)
`and Y Noun:CaseLocative` (second premise)
`X r_has_part::essential component Y` (conclusion with annotated relation)

`if X` "na" `Y` (description of the context)
`and X Verb` (first premise)
`and Y Noun:CaseAccusative` (second premise)
`X r_place_action::surface Y` (conclusion with annotated relation, Y is expected to be a surface)

`if X` "v" `Y` (description of the context)
`and X Verb` (first premise)
`and Y Noun:CaseAccusative` (second premise)
`X r_manner::modality Y` (conclusion with annotated relation, i.e. "roll into a ball")

The extraction rule examples show the variety and the kind of semantic information that can be captured through the observation of syntactic and grammatical features of Russian.
The extraction step of our experience remains dependent on the output of the parser. In our case, due to the specificity of the cooking domain and its language, some lemmas have been unknown. In many cases, the parser tagged past participle forms as adjectives which made more difficult the extraction of sequences of actions.
In addition to the main set of rules, we defined a small set of morphology based rules to extract *part-of* and *state-of* relations from the adjective forms. To cite an example, the adjectives having the suffix `[-yann]` allow creating extraction rules such as

`if X r_carac Y`
(semantic context i.e."spoon r_carac wooden")
`and X Noun`
`and Y Adj [STEM][yann][FLEXION]`
(morphological structure of the typical characteristic of X *derevyannyi*), "wooden"
`Z Noun [STEM][o]`
(inference of the Noun Z, the material X is composed of *derevo*, "wood")
`X r_has_part::substance Z`
(conclusion)
The results of the extraction process are given in the next sections.

4.3. Inference

The inference scheme is based on the use of an interlingual pivot in the MLSN. It has been started as a natural pivot and incrementally evolves towards a fully interlingual one. DBnary (Sérasset, 2014) has been exploited to yield translation links (English and Russian edition) as this resource is oriented towards a sense based alignment as proposed by (Tchechmedjiev, 2016). The pivot has been enriched through multiple processes such as exogenous acquisition from external monolingual and multilingual knowledge resources or extractions from domain specific texts as detailed in (Bebeshina-Clairet, 2019).
Interlingual nodes are linked to the language specific nodes through the relations typed *r_covers* (figure 1). One term may have multiple covering terms. One interlingual term may cover multiple "language" terms. Every cross-lingual link is a path traversing the pivot. Nevertheless, the pivot cannot be considered as interlingual as the alignments between the available senses of the sub-graphs are being completed.

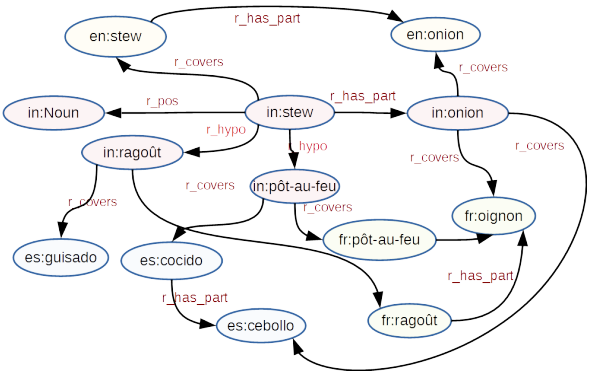

Figure 1: MLSN architecture

The actual inference process combines ascending inference step (source language ⇒ pivot) and the descending inference step (pivot ⇒ target language).

During the ascending step, the semantic relations of the source sub-graph are inferred in the pivot with a set of constraints that apply depending on the type of the relations to be inferred :

- Part-of-Speech constraints i.e. we basically need a noun (source term) and an adjective (target term) to form a r_carac (typical characteristic) candidate;
- semantic relatedness i.e. triangulation test: "if strawberry r_isa berry and berry $r\_location$ dish, then strawberry $r\_location$ dish". Thus, we infer the new relation if its extremities are connected through a typed path (semantically related);
- semantic similarity i.e.the neighbourhood of the interlingual term covers a part of the neighbourhood of the source term. Similar to the abduction inference scheme, to propose the relations of a term T to the term T' we need T and T' to be similar enough ("share" a certain number of semantic relations).

During the descending phase, the inference process is completed by the evaluation process based on lookup and inference in the RezoJDM LSN. First, the inferred relations are added into the French sub-graph of the MLSN. Second, the evaluation is done using the tool Helix[4] designed to check

[4] `http://www.jeuxdemots.org/rezo-ask.php`

the presence of the relations and complete the RezoJDM. Third, the feedback of the evaluation affects the weight and the status of the relations in the French sub-graph of the MLSN as well as in the interlingual and the original (Russian sub-graph) parts.

The results of the ascending inference process reflect how good the coverage of the pivot over the source language is in the MLSN. The descending inference phase shows the pivot coverage of the target sub-graph within some type of knowledge (semantic relation type). It also yields (trough the evaluation) the information on the wrong relations. These wrong relations are kept in the MLSN with a negative weight to prevent from their acquisition by other means (exogenous process, inference schemes) The results

type	**#extr**	**#asc**	**#desc**
r_carac	12 488	22 631	40 762
r_manner	11 084	6 000	8 929
r_place_action	5 679	2 548	2 969
type-place (place feat.)	2 298	5 551	6 267
part-of (charac. feat.)	543	3 371	2 983
state-of (charac. feat.)	756	1 318	440
Overall	32 848	41 419	62 350

Table 1: number of extracted relations (**#extr**), ascending (**#asc**), and descending (**#desc**) inference of the relations extracted from text.

of the inference process listed in the table 1 show the two-step acquisition of semantic information.

First, the relations typed *r_carac*, *r_manner*, and *r_place_action* (typical place where an action can take place) have been extracted. Second, based on these sets of relations, it has been possible to use the morphological (suffixation) and grammatical (use of plural, use of the Accusative case as opposed to the Locative case) in order to extract supplementary sets of semantic information i.e. *part-of* (composition), *state-of*, and *type-place*. These pieces of semantic knowledge are used for the relation annotation (attaching meta-information to the relations of the (M)LSN).

We notice that, compared to the extraction, the inference processes yield numerous candidate relations. This is due to the presence of sense refinements in the MLSN (i.e. sense refinements obtained from the pre-existing resources (such as WordNet (Fellbaum, 1998), RezoJDM, ConceptNet ((Speer and Havasi, 2012). When a term in the source language (inference entry point) is unrefined (not disambiguated), it is linked to the potential senses possibly present in the pivot in the target sub-graph. Term disambiguation is crucial for the lexical semantic resource building. When automatically validated, inferences allow disambiguation of the polysemous terms present in the MLSN.

4.4. Evaluation, Adjustment, and Self Adjustment

The evaluation process is an inference based process run on the RezoJDM LSN. It is used to automatically enhance the MLSN as well as the semantic relation extraction. The evaluation with the Helix tool returns the following response regarding a relationship to be tested :

- "true" (the relation is present in the RezoJDM);
- "true by inference" (the relation is inferrable in the RezoJDM);
- "do not know" (the relation is absent from the RezoJDM and the inference processes are unable to validate the relation, this response is quasi-equivalent to "false");
- "false" (the relation has a negative weight);
- "false by inference" (the relation is inferrable as false in the RezoJDM);
- "unknown term" (source or target term of the relation to be tested is absent from RezoJDM).

type	**#desc**	**#true_rel**	**#true_inf**	**#undec**
r_carac	40 762	815	2 454	37 493
r_manner	8 929	35	2 419	6 475
r_place_act	2 969	74	453	2 442
type-pl	6 267	49	3 181	2 777
part-of	2 983	64	283	2 636
state-of	440	44	264	132
Overall	62 350	1081	9 054	51 955

Table 2: Evaluation of the inferred relations. For simplicity, only the relations accepted by lookup (**#true_rel**) or inference (**#true_inf**) and the rejected relations **#undec** ("do not know", term is absent, inference) are listed.

The salient aspect of the evaluation results is the importance of the monolingual inference run on the RezoJDM LSN in order to validate or invalidate the relations proposed through the cross-lingual inference. The percentage of automatically validated (true) relations complies with the rate usually observed in the context of manual evaluation (around 6% to 10%) of the automatically inferred relations (Bebeshina-Clairet, 2019), (Zarrouk, 2015). The inference-related results also show that many of the relations were not present in the RezoJDM. This resource could be enhanced using the proposed approach as the cross-lingual relation acquisition may "confirm" the inferrable RezoJDM relations.

The inference output analysis and the evaluation result are used to adjust the MLSN in terms of two central criteria:

- **coverage**: how many inferences are due to the silences of the MLSN? how many relevant relations are exploited during the inference process? Negative relations (relations with a negative weight) are not silences, they represent the set of true negatives and thus help guiding the inference process;
- **relevance**: the relations proposed by inference, are they semantically true? do they correspond to the shared knowledge? To be ultimately used for semantic analysis or information retrieval tasks, the MLSN

needs to be relevant and reflect what humans do know about the subjects it covers.

5. Discussion

The impact of the syntax and morphology based extraction from texts is limited as many of the observed features are polysemous. Indeed, the same "preposition + noun case" combination may correspond to different semantic information. Multiple patterns can match the same pair of source and target terms. Sometimes, the ambiguity of the extracted relations may be solved by inference. For instance, in the cases similar to that of *refrigerator* that appears in both, the constructions that suggest a "surface" place and the constructions that suggest "container, volume, room", we could attempt inferring that *refrigerator* (as well as *cake* or *cheese*) are solid containers with sides or bottom part that can be used as a surface. We do not observe such ambiguity during the extraction process from Russian procedural texts when we analyse the syntactic behaviour of the terms corresponding to the other types of "places" such as *jar*, *tandoor*, *steak* or *supermarket*.

Solving such ambiguities may be costly in terms of human effort (rule design, manual validation of extracted relations) and may need introducing some additional semantic knowledge. Thus, monolingual and cross-lingual inferences appear as a way to reduce human supervision effort and to automatically select the reliable pieces of knowledge for the lexical semantic resource construction.

Building the LSNs that can support inference-based evaluation procedures requires a lot of effort. However, our experiences show that a reliable resource in only one language is sufficient for building and enhancing other language subgraphs in a significantly less supervised manner.

6. Conclusion

In this paper, we first introduced the state of the art techniques for lexical semantic resource building. It seems clear that one strong bottleneck is the question of validation which is done really confidently only manually. Then, we detailed the resources we made use of for our inference based resource building approach. Finally, we provided the details about the resource building and the figures of some evaluation experiment. The undertaken evaluation showed clearly that our approach is meaningful and promising. Indeed, using a large lexical KB to ensure as automatically as possible that inferences are correct seems to be a proper way toward an unsupervised approach.

7. Bibliographical References

Agirre, E., Ansa, O., Arregi, X., Arriola, J. M., de Ilarraza, A. D., Pociello, E., and Uria, L. (2002). Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. In *Proceedings of First International WordNet Conference. pp. 32-40. Mysore (India).*

Bebeshina-Clairet, N. (2019). *Construction d'une ressource termino-ontologique multilingue pour les domaines de la cuisine et de la nutrition*. Theses, Université Paris 13, January.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. R. H., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91 – 134.

Fellbaum, C. (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Gelbukh, A. F. (2018). Inferences for enrichment of collocation databases by means of semantic relations. *Computación y Sistemas*, 22(1).

Hernández-González, J., Hruschka Jr., E. R., and Mitchell, T. M. (2017). Merging knowledge bases in different languages. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 21–29, Vancouver, Canada, August. Association for Computational Linguistics.

Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. In *SNLP'07: 7th International Symposium on Natural Language Processing*, page 7, Pattaya, Chonburi, Thailand, December.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., and Welling, J. (2015). Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Navigli, R. and et al. (2012). Babelnet: The automatic construction, evaluation and application of a . . .

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2015). A review of relational machine learning for knowledge graphs: From multi-relational link prediction to automated knowledge graph construction. *CoRR*, abs/1503.00759.

Ramadier, L. (2016). *Indexation and learning of terms and relations from reports of radiology*. Theses, Université de Montpellier, November.

Rebele, T., Suchanek, F. M., Hoffart, J., Biega, J., Kuzey, E., and Weikum, G. (2016). Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In Paul T. Groth, et al., editors, *International Semantic Web Conference (2)*, volume 9982 of *Lecture Notes in Computer Science*, pages 177–185.

Sérasset, G. (2014). DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web – Interoperability, Usability, Applicability*, pages –. To appear.

Sharoff, S. and Nivre, J. (2012). The proper place of men and machines in language technology processing russian without any linguistic knowledge. Dialogue 2011, Rus-

sian Conference on Computational Linguistics.

Shin, J., Wu, S., Wang, F., De Sa, C., Zhang, C., and Ré, C. (2015). Incremental knowledge base construction using deepdive. *Proc. VLDB Endow.*, 8(11):1310–1321, July.

Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *LREC Proceedings*.

Tchechmedjiev, A. (2016). *Semantic Interoperability of Multilingual Lexical Resources in Lexical Linked Data*. Theses, Université Grenoble Alpes, October.

Wang, Q., Wang, B., and Guo, L. (2015). Knowledge base completion using embeddings and rules. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1859–1865. AAAI Press.

Zarrouk, M. (2015). *Endogeneous Consolidation of Lexical Semantic Networks*. Theses, Université de Montpellier, November.