Extracting semantic relations via the combination of inferences, schemas and cooccurrences

Mathieu Lafourcade[1] Nathalie Le Brun[2]
(1) LIRMM, 860 rue de St Priest, 34095 Montpellier cedex 5, France
(2) Imagin@t, 34400 Lunel, France
lafourcade@lirmm.fr, imaginat@imaginat.name

Abstract

Extracting semantic relations from texts is a good way to build and supply a knowledge base, an indispensable resource for text analysis. We propose and evaluate the combination of three ways of producing lexical-semantic relations.

1 Introduction

The semantic relations, whether ontological (*hyperonymous*, *hyponyms*, *parts* / *whole*), lexical (synonyms), or semantic roles (agent, patient, instrument, way, place, etc.) are of a major interest for almost all of the applications of NLP where the system has to "understand" what a text means. That is the case, for instance, in automatic translation, indexing, summary, detection of similar texts, etc. The creation of procedures to produce semantic relations therefore meets multiple needs in the field of NLP.

There are several ways to extract semantic relations. Some methods are manual, as for WorldNet (Miller, 1995), while others are more or less automatic (BabelNet, (Navigli and Ponzetto, 2010)) or contributory (Lafourcade, *et al*., 2015). Among the many methods of extracting semantic relations from texts, the performances are very unequal. Some are highly accurate, but this precision requires a thorough semantic analysis, which is costly. Moreover, the need to analyze texts with great precision considerably slows down the process of extracting semantic relations. Conversely, some statistical methods include virtually no language processing of the input text.

In this paper, we assess the interest of combining three different strategies to extract semantic relations. This approach is implemented in the context of never ended learning, within the lexical network resulting from the JeuxDeMots project (Lafourcade, 2007). The idea is to implement extraction / exploitation loops in which an automatic extraction system plays the role of contributor within the network. Players / contributors validate or invalidate relations, either through games (GWAPs) or through direct contributions. Thus, we can assess in a holistic way the performance of our SIC (Schema-Inferences-Cooccurrences) system, which feeds the network and uses it to carry out its task.

In the following, we mention, among the previous works in automatic extraction of semantic relations, those whose methodology is similar to ours. Then, we detail three extraction strategies and how we combine them. Finally, we detail and discuss the results.

2 Previous works

Using lexical-semantic schemes to extract synonymy and hyperonymy relations from texts has been proposed by (Hearst, 1992). The schemes may be of "A is a B" type. Herbelot and Copestake (2006) used such diagrams to extract relations in biology from Wikipedia pages with excellent precision (88%) but a fairly low recall (20%). Ruiz-Caasado (2005) and (2007) evoke the automatic learning of such schemas to extract relations from Wikipedia, and insert them into Wordnet. Here again, it is noted that the performances regarding the recall are rather weak. The approach of using automatic learning of schemas from texts has also been exploited by Snow et al., (2004), also to identify hyperonymy or hyponymy relations. In (Girju, *et al.*, 2003), a supervised approach aims to determine the semantic constraints to extract meronymic relations. The constraints are defined by the *part-of* relation of Wordnet and serve as training data. Ramadier and Lafourcade (2016) propose a similar approach, with the difference that they determine the constraints manually and identify many semantic relations.

Many authors try to extract relations from Wikipedia by exploiting the structure information of the pages. For example, Sumida and Torisawa, (2008) used this strategy on Wikipedia in Japanese to extract 1.4 million hyponymy relations with an precision of 0.75. In the same way, Ponzetto and Strub (2007) exploit the Wikipedia category links to identify hypernymy relations. Pachenko (2013) presents an in-depth analysis of functions of evaluation of semantic relations between terms. One of the conclusions is that none of the evaluation measures of the semantic relations is better than the others. The proposition of such or such semantic relations, mostly ontological (hypernyms, co-hyponyms, etc.) is conditioned by these different measures of similarity. It should be noted that little work in this field is based on the use of knowledge bases to extract new semantic relations in a continuous loop learning approach. However, such bases are often used for training in automatic learning. Moreover, most approaches are limited to ontological relations, such as hypernymy (*is a*), synonymy (*syn*), and meronymy or holonymy (*has parts* / *is part of*). Relations like *cause* / *consequence*, *characteristics*, *location*, *agent*, *patient* and *instrument* (for verbs) are rarely extracted. In the approach we describe here, we use pure text, exclusively in French, from Wikipedia or otherwise, and we do not exploit the structure of the source document. We also want to extract information from non-encyclopedic texts (such as novels, for example). For each extraction method, we use, at various degrees, the lexical-semantic network JeuxDeMots (JDM).

3 Combining three Relations Extraction Methods

We present three quite simple methods for extracting semantic relations between pairs of terms. We then outline their combination (the first and second ones are new methods).

3.1 Cooccurrences and Relations

At first, we need a cooccurrences network. The method to get it takes into account compound terms and includes a pre-processing on the text, which consists of several steps:

First, a term is replaced by its lemma, but only when the term is a conjugated verb. For example, the segment: *les poules dorment* will become *les poules dormir*. On the other hand, the segment *les poules couvent* remains unchanged since *couvent* can be a conjugate form of the verb *couver* but also the substantive *couvent* (convent).

Secondly, we identify occurrences of compound terms by confronting with the JDM network. The spaces are replaced by underscores, which avoids their segmentation.

The punctuation marks are preserved, but detached from the terms that precede them: *chat, =>chat ,*. Caps are not changed; with the exception of what is mentioned above, neither pos tagging nor parsing is performed.

Compound terms are identified by comparison with those existing in the JDM network. In case of conflict (for example, a segment A B C with two compound words A_B and B_C), a priority is applied to the right (A B_C). Then, segmentation is made using the spaces characters. A *k-word* window is used to establish the co-occurrence relations, with a decreasing weight from *k* (adjacent word) to *1* (word at a distance of *k* terms). We used a window of 10 words, in order to maximize the recall, which is the objective of this method.

We use the JDM knowledge base as a support for the determination of lemmas and morphosyntactic categories, but also for the approximate identification of the types of semantic relations. More precisely, we have rules of this type, which exploit the parts of the speech:

- If X *r_pos* verb & Y *r_pos* adv → X *r_manner* Y
- If X *r_pos* noun & Y *r_pos* adj → X *r_carac* Y
- Default settings : if X is in co-occurrence with Y, → X *r_assoc* Y

The rules are strict and must be understood as: if X is a verb and only a verb and if Y is only an adverb then X will be linked to Y by a *manner* relation. For example, the following sentence: "the cat quickly caught the black rat". The pretreatment phase provides us with the text: "the cat quickly catch the black rat" (we do not indicate weights). The following relations are weighted by the weight of the cooccurrence between the two terms:

cat *r-assoc* catch	rat *r_assoc* black
catch *r_manner* quickly	quicly *r_assoc* rat
catch *r_assoc* rat	cat *r_assoc* rat
catch *r_assoc* black	cat *r_assoc* black
…	quickly *r_assoc* cat

3.2 Lexical-Semantic Schemes with Constraints

Ramadier and Lafourcade (2016) modified the method of Hearst (1992) as well as Herbelot and Copestake (2006), which exploits the semantic schemas, so that the terms satisfy semantic relations coming from the JDM lexical network. For example:

- X of Y with X *r_isa* artefact & Y *r_isa* person → Y *r_own* X (soldier's rifle)
- X of Y with X *r_isa* part of body & Y *r_isa* person → Y *r_part* X (soldier's arm)
- X of Y with X *r_isa* person & Y *r_isa* human place → Y *r_place* X (the girl of the coron)

Of course, some schemas are not associated with constraints, for example:

- X is located in the/my/a/some/ Y → X *r_place* Y
- X is a type of Y → X *r_isa* Y
- X is part of Y → X *r_holo* Y
- X consists of Y → X *r_has_parts* Y

A relation between two terms will be weighted by the number of times it was discovered using different schemes in separate text segments.

3.3 Induction and Abduction within a Lexical-Semantic Network

Zarrouk, *et al.* (2014) and Zarrouk and Lafourcade (2015) proposed an inference-based method to produce new semantic relations. This strictly endogenous approach relies on the JDM network: no text is used. It is based on deduction and various forms of abduction.

3.4 How to Combine these Approaches?

The combination of two methods consists in retaining only the semantic relations found jointly by each of the two methods. Although the co-occurrence method produces non-specific relations (i.e. *associated ideas* relations between terms), which have no equivalences in the other two methods, these neutral relations are used as follows:

X *r_t* Y + X *r_assoc* Y → X *r_t* Y
A neutral relation (type *r_assoc*) validates a typed one (type *r_t*)

We combine approaches in pairs because a combination of the three approaches, while increasing accuracy, would reduce too much the number of retained relations. We will therefore retain the relations produced by at least two of the three methods. The weight of the combination is the geometric mean (square root of product) of the relations.

4 Experimentation and Discussion

The inference approach was tested on the lexical network. For the two others, which require texts, we used a corpus consisting of Wikipedia in French (for schemas and cooccurrences) and the work of Emile Zola (for cooccurrences). This choice of corpus reflects the desire not to limit ourselves to encyclopaedic texts, and to enrich the collection of semantic relations by exploiting the advantages of novelistic literature: to offer (1) a greater diversity of relations between terms, and (2) more common sense information and relating to everyday life.

The extracted relations are: *synonymy* (for verbs, names, adjectives, adverbs), *agent*, *patient*, *instrument*, *manner*, *take place* (for verbs), *hypernymy*, *hyponymy*, *instance*, *characteristics*, *is located in*, *is a place for*, *parts of*, *whole*, *cause*, *consequence* (for nouns).

The *productivity* is the ability of a method to produce relations. We use this measure in place of the traditional recall, which we are not able to evaluate. Indeed, we do not have linguists / lexicographers to determine the complete set of relations that should be extracted from our corpus. Such a work would be very cumbersome in that it is necessary to go through each text by hand. In addition, the inter-annotator agreement is generally not very high (less than 50% on average). In order to evaluate productivity, we take the inference method as a reference and assign it a productivity value of 1. In practice, this method yielded about 60 million relations (which are potential until they have been validated) between November 1, 2016 and March 30, 2017.

The *precision* is the ratio between relations assessed as fair and all proposed relations. We are of course seeking to maximize this critical criterion, while maintaining good productivity.

Finally, *relevance* is the ratio between the relevant relations and the right relations. Deciding if a relation is relevant remains relatively subjective, but respondents (by crowdsourcing and GWAP) generally agree. Relevance is related to the specificity of a relation; in general, the more a relation is specific to a class of terms, the more it is relevant.

4.1 Methodology

The assessment was carried out jointly by two methods: 1) manual validation of a random sample, and 2) matching of player responses via JeuxDeMots. This evaluation is carried out continuously (the data below are those of the period from November 2016 to March 2017), the results shown below are those at the end of March 2017. For manual validation, the relation to be evaluated is submitted to a player (through the *Askit* game, http://jeuxdemots.org/askit.php), who must decide on its validity and relevance. The matching method involves the classic game of the JeuxDeMots project: a player is offered a game with the first term of the relation to be evaluated, and the type of the relation in question. For example, if the relation *rat r_carac black* has been extracted, then games are proposed with the term *rat* and the instruction to give terms relevant to the relation *r_carac* ; then, the player must give some characteristics of *rat*. If *black* is among its answers, then, the relation *rat r_carac black* is validated. This method is equivalent to questioning people to see if the extracted relation emerges or not. Players can move on if they do not know. We have selected as a priority the relations whose weight corresponds to the 2nd quartile (*i.e.* the 50% with the highest weights).

4.2 General Results

To evaluate the methods independently of each other, we use method I (Inferences) as a reference for the number of relations produced in 5 months, i.e. 60 million. The evaluation of the 3 methods considered individually for the 3 criteria of *productivity*, *precision* and *relevance* are presented in Table 1.

	Schemas (S)	**Infer. (I)**	**Cooc. (C)**
Productivity	0.37 (22 M)	1 (60 M)	3,16 (190 M)
Precision	93 %	65 %	12 %
Relevance	88 %	75 %	47 %

Table 1: Evaluation of the 3 methods individually

The method of extracting through *lexical-semantic schemes* (S) with constraints produces very few false relations but is relatively slow. The 7% error corresponds to the impossibility of applying constraints, when at least one of the two terms of the relation is not sufficiently provided with information. The extracted relations are relevant, which is normal since they are taken directly from the texts.

The *cooccurrence method* (C), as expected, is very productive, fast, and very imprecise (a lot of waste). Correct relations are relevant once in two.

Finally, the *inference method* (I) (which is based only on the JeuxDeMots knowledge base) shows quite good performances. Errors come mainly from the impact of polysemy, which disrupts deductive and abductive inferences. The difference in productivity between S and C is essentially explained by the speed difference of the two methods (S slow and precise, C fast and fuzzy).

The combination of two-by-two methods consists in retaining a relation only if it is proposed by both methods. It is therefore an intersection between the proposals of the two methods.

	S+I	**S+C**	**I+C**
Productivity	0.22	0.35	0.78
Precision	96 %	94 %	87 %
Relevance	93 %	84 %	88 %

Table 2: 2-by-2 methods evaluation. The productivity is the highest when *Inference* and *Cooccurrences* are combined. Highest precision is achieved when *Schemas* is combined with *Inferences*.

We find that for each pair productivity decreases with each method taken in isolation, which is an expected result. The S + I combination has very low productivity, indicating that few relations are produced by both S and I methods.

	(S+I) ∪ (S+C) ∪ (I+C)
Productivity	1.28
Precision	99.4 %
Relevance	96 %

Table 3: Evaluation of the approach retaining relations proposed by at least 2 methods (union). Clearly, this method combination tends to maximize the three evaluation criteria.

The approach through combination of the three methods not only produces a 28% increase in productivity compared to method I (taken as a reference), but also increases precision and relevance. If we use method I as a reference, the combination of the approaches allows to reinforce the precision with the method S and the relevance with S and C. Combining S + C allows (again with reference to I) to add relations which could not have been inferred. We recall that C is applied to a corpus of text larger and more general than the method S. This approach tends to increase the common sense relations we are able to capture, without corrupting precision and relevance. Still, relevance seems to be quite difficult to get even

though the usage of a large corpus helps focusing on mostly relevant relations.

4.3 Evaluation per Relation Type

We evaluated our approach through relation types.

Relation type	Precision	Relevance
For nouns		
R_isa	98	97
R_carac	99.8	96
R_has-part	98.6	97
R_holo	98.9	98
R_own	97.6	94
R_place_n	99.9	97
R_member_of	98.6	96
R_produce	94.2	98
For verbs		
R_agent	99.5	96
R_patient	99.8	95
R_manner	99.9	97
R_place_v	97.2	98
For both Nouns and Verbs		
R_consequence	97;1	93
R_cause	97.6	93

Table 4: Evaluation detailed by relation type.

There are some variations of precision and relevance amongst the different types of relations. Some relations quite explicit in texts of Wikipedia, for instance the *r_isa, r_carac, r_place* are extracted quite faithfully with a quite high precision and relevance.

The *place* relation (for nouns and verbs) is more difficult to detect for verbs than for nouns. For verbs, some wording about the *manner* may be similar to wording related to the *place*, which can lead to wrong relation type identification.

There is often confusion between *r_holo* and *r_member_of* because these two relations are often expressed in a similar way if not identical. The main distinctive criterion is the use of the singular or plural, or of an entity representing a set.

Le chat fait partie des félins The form *félins* being in the plural, it can be considered as a set, hence leading properly to the *r_member-of* relation. Similarly, for the sentences:

Le soldat fait partie de l'armée. L'abeille fait partie de la ruche

The terms *armée (army)* and *ruche (hive)* have a set aspect, leading also to *the r_member-of* relation. But if we consider:

(a) Les fibres de ce bois sont longues.

(b) Les animaux de ce bois sont craintifs.

It is much more tricky to properly identify the proper relation types between *r_holo* or *r_member* or even *r_place.* Even human validators may hesitate to identify the appropriate relation. For sentence (a), the most appropriate relation is *r_holo* (fibers are part of the wood (matter)). For sentence (b) animals are both part of the woods (forest) (relation *r_member*) and are in the woods (*r_place*). The *cause* and *consequence* relations are much quite difficult to spot, because they are expressed in different ways.

Difficult Cases

We encountered some difficult cases. Consider this definition from Wikipedia:

> « La Frégate du Pacifique (Fregata minor) est une espèce d'oiseaux marins appartenant à la famille des Fregatidae. » (Eng : The Frégate du Pacifique (Fregata minor) is a species of sea bird belonging the Fregatidae family.)

The *S* method leads to: *frégate du Pacifique r_isa oiseaux marins*, which stumbles into the problem of the number, as a singular noun (*frégate*) cannot be a plural noun (*oiseaux marins*). We have to deal with some special handling when the lemma of the right part should be considered. In that case, we should obtain: *frégate du Pacifique r_isa oiseau marin*. Note that in this typical case, the cooccurrence and inference mechanisms are very useful. Another relation extracted from this example is *frégate du Pacifique r_member-of Fregatidae.*

Some Typical Failure Examples

Consider the following sentence: *Sa beauté créait bien des tourments (Eng. Her beauty was the origin of many torments.)*

Our approach deduces the following relation: *beauté r_produce tourments*, but in fact the wording is quite metaphorical and misleading, and the proper relation would be *r_consequence*. Such sentences are more frequent in literary texts than in encyclopedic ones (like Wikipedia), nevertheless they are quite common and unless being able to undertake a deep semantic analysis, such relations would remain difficult to identify properly. We also have problem with anaphoric chain, like in this sentence (Wikipedia): *Les chiens de prairie (Cynomys) forment un genre de rongeurs qui comprend cinq espèces.*

Our system identifies wrongly the relation *rongeur r_has part espèces*. In fact, we should

have had: *chiens de prairie r_member_of rongeurs.* But such identification is beyond the reach of our methodology, as it requires some reconstruction of the surface expression.

Even if we try to be as precise as possible (at the expense of some recall), there are still some cases in which our approach wrongly identifies the relation types, as some deep understanding seems to be mandatory. Such an automatic understanding could be very costly to obtain in terms of computing on the one hand, and on the other hand requires a quantity of knowledge (at least knowledge of common sense) and … that is precisely what we are trying to capture.

Hence, increasing the size of the corpus can lead to increase the chance to capture a given relation by several different schemas, and thus increasing both precision and relevance. Since common sense relations are the most appropriate for identifying other candidate relations, it would seem that novels are the best source for relation extraction.

4.3 Evaluation with Semantic Class

We evaluated our approach on the basis of the semantic class of the term for which the relations have been discovered (the left-hand side term of discovered relations).

Semantic class	Precision	Relevance
animal	98.3	98
plant	99.7	96
actor/actress	98.3	97
vehicle	98.8	98
disease	98.5	96
medicament	99.2	98
food	99.6	98
movie	97.2	98
city	98.5	97

Table 4: Evaluation detailed by semantic class.

A term belongs to a given semantic class if the word that designates that class is a possible hypernym for it. We should remind that a term could belong at the same time to several semantic classes, as it can be polysemous. For example, the term *frégate* (frigate) is both a bird and a boat/vehicle. The semantic classes listed above account for around 68 % of all terms. The 32% other terms belong to other semantic classes (process, book, geographical place, other types of persons, natural phenomena, etc.). In Table 4, we can see that for the main semantic classes there is not a strong variation amongst precision and relevance values. Perhaps the *movie* semantic class is an exception for precision, the reason being it is difficult to infer precise relation for a given movie unless analyzing elements of the story. The *disease* semantic class, on the other hand has some good precision but low relevance, the relations found although being correct are quite general.

5 Conclusion

We have presented three methods for the identification of semantic relations between terms. One of them is strictly endogenous and relies on a knowledge base (the JeuxDeMots network), the other two are based on texts, but also use semantic information external to the texts. Overall, these methods are light. Individually, they have defects (productive but imprecise / precise but not very productive). Their union, by offsetting their respective defects, significantly improves productivity, relevance, and precision.

The performance of our method seems not to be very sensible toward the semantic class of the term for which a candidate relation is extracted. We were not able to spot a particular semantic class that would really be underperformed. These results are some good news, as its support the idea that there is no need for some specific treatment for some semantic class of words.

What about extracting automatically some semantic schemes from corpora? We are in the process of undertaking such a task, however there are at least two pitfalls: first, we do extract a very large number of relations that have to be manually validated before being exploited (supervised approach). This is a long and delicate task. Second, it is particularly difficult to automatically determine the semantic constraints to be associated with a semantic scheme. Indeed, on what criteria can we choose in the network the relations that must be verified by the elements of the scheme? To avoid producing schemes with too general constraints, which would lead to erroneous relations, the system tends to select very specific constraints. This has the effect of over-multiplying the schemes (several tens of thousands) and thus considerably increasing the computing time.

Since strictly semantic (and not lexical) relations are independent of languages, a way of further improving the process would be to adapt this approach to the extraction of relations from texts of different languages, using either a translation process, or the multilingual JDM network currently under study.

References

GIRJU, R., BADULESCU, A., and MOLDOVAN, D. (2003) *Learning semantic constraints for the automatic discovery of part-whole relations.* In Proc. Conf. North American Chapter of the Association for Computational Linguistics on Human Language Technology , NAACL '03, pages 1–8. Association for Computational Linguistics, 2003.

HEARST, M. A. (1992) *Automatic acquisition of hyponyms from large text corpora.* In Proc . 14th Conf. on Computational Linguistics, COLING '92, pages 539–545. Association for Computational Linguistics, 1992.

HERBELOT, A. and COPESTAKE, A. (2006) *Acquiring ontological relationships from Wikipedia using RMRS.* In Proc. ISWC 2006 Workshop on Web Content Mining with Human Language Technologies , 2006.

LAFOURCADE, M. (2007) *Making people play for Lexical Acquisition.* In Proc. SNLP 2007, 7th Symposium on Natural Language Processing. Pattaya, Thailande, 13-15 December 2007, 8 p.

LAFOURCADE, M., LE BRUN N., and JOUBERT A. (2015) *Games with a Purpose (GWAPS)*, ISBN: 978-1-84821-803-1 July 2015, Wiley-ISTE, 158 p.

MILLER, G. A. (1995) *Wordnet: A lexical database for English.* Communications of the ACM , 38(11):39–41, November 1995.

NAVIGLI, R.and PONZETTO, S. P. (2010*) Babelnet: Building a very large multilingual semantic network.* In Proc. 48th Annual Meeting of the Association for Computational Linguistics , ACL'10, pages 216–225, 2010.

PANCHENKO, A. (2013) *Similarity Measures for Semantic Relation Extraction*. PhD Dissertation, Université catholique de Louvain & Bauman Moscow State Technical University, 193 p.

RAMADIER, L. ET LAFOURCADE, M. (2016) *Patrons sémantiques pour l'extraction de relations entre termes - Application aux comptes rendus radiologiques.* In 23rd French Conference on Natural Language Processing (JEP-TALN-RECITAL 2016), Paris, France, 4-8 July 2016, 6 p.

RUIZ-CASADO M., ALFONSECA E., AND CASTELLS P. (2005) *Automatic extraction of semantic relationships for Wordnet by means of pattern learning from wikipedia.* In Proc. 10th Int. Conf. Natural Language Processing and Information Systems, NLDB'05, pages 67–79. Springer, 2005

RUIZ-CASADO M., ALFONSECA E., AND CASTELLS P. (2007) *Automatising the Learning of Lexical Patterns: an Application to the Enrichment of Wordnet by Extracting Semantic Relationships from Wikipedia.* In Data & Knowledge Engineering , Issue 3 (June 2007) 25p.

SNOW R., JURAFSKY D., AND ANDREW Y. NG. (2004) *Learning syntactic patterns for automatic hypernym discovery.* In Advances in Neural Information Processing Systems (NIPS), 8 p. 2004.

SUMIDA, A. AND TORISAWA, K. (2008) *Hacking Wikipedia for hyponymy relation acquisition.* In Proc. of IJCNLP 2008 , pages 883–888, 2008.

ZARROUK, M., LAFOURCADE, M., and JOUBERT A. (2014). About Inferences in a Crowdsourced Lexical-Semantic Network, *EACL 2014 (14th Conference of the European Chapter of the Association for Computational Linguistics)*, Gothenburg (Sweden), April 2014

ZARROUK, M. and LAFOURCADE, M. (2014) *Inferring Knowledge with Word Refinements in a Crowdsourced Lexical-Semantic Network.* In proc of the the 25th International Conference on Computational Linguistics (COLING 2014), Dublin, Irlande, 9 p.