Semantic Relation Extraction with Semantic Patterns: Experiment on Radiology Report
Mathieu Lafourcade, Lionel Ramadier

To cite this version:
Mathieu Lafourcade, Lionel Ramadier. Semantic Relation Extraction with Semantic Patterns: Experiment on Radiology Report. LREC: Language Resources and Evaluation Conference, May 2016, Portorož, Slovenia. hal-01382320

HAL Id: hal-01382320
https://hal.archives-ouvertes.fr/hal-01382320
Submitted on 16 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Semantic Relation Extraction with Semantic Patterns: Experiment on Radiology Report

Mathieu Lafourcade, Lionel Ramadier
LIRMM, University of Montpellier
860, rue de St Priest, 34095 Montpellier
E-mail: Mathieu.lafourcade@lirmm.fr, lionel.ramadier@lirmm.fr

Abstract

This work presents a practical system for indexing terms and relations from French radiology reports, called IMAIOS. In this paper, we present how semantic relations (causes, consequences, symptoms, locations, parts…) between medical terms can be extracted. For this purpose, we handcrafted some linguistic patterns from on a subset of our radiology report corpora. As semantic patterns (de (of)) may be too general or ambiguous, semantic constraints have been added. For instance, in the sentence néoplasie du sein (neoplasm of breast) the system knowing neoplasm as a disease and breast as an anatomical location, identify the relation as being a location: neoplasm r-lieu breast. An evaluation of the effect of semantic constraints is proposed.

Keywords: Relation Extraction, Semantic constraints, NLP

1. Introduction

In the domain of radiology, the amount of pictures and textual reports is growing at an unprecedented rate. This quantity of medical information exceeds the ability of the radiologist to manage it effectively. One challenge is then to figure how to make computer help to effectively use these findings and to improve the management of patients. NLP methods allow for enhanced indexing and searching of medical reports with the linking of relevant terms through semantic relations and enabling on the fly inference of new knowledge.

This work presents a practical system for indexing terms and relations from French radiology reports, called IMAIOS. The main task of this system is to extract semantic relations from unstructured text that can be integrated in a lexical-semantic network in order to improve the quality of the latter. The IMAIOS system differs from other techniques by the use of the French lexical-semantic JeuxDeMots (JDM) network as supporting resource. Although the JDM network is primarily a knowledge base of common sense, it contains also many specialized facts, including medicine/radiology, which have been added within the framework of IMAIOS system. With such a large lexico-semantic network, we can devise quite crude algorithms that are very efficient.

In this paper, we present how semantic relations (causes, consequences, symptoms, locations, parts…) between medical terms can be extracted. For this purpose, we handcrafted some linguistic patterns from on a subset of our radiology report corpora. As semantic patterns (de (of)) may be too general or ambiguous, semantic constraints have been added. For instance, in the sentence néoplasie du sein (neoplasm of breast) the system knowing neoplasm as a disease and breast as an anatomical location, identify the relation as being a location: neoplasm r-lieu breast. An evaluation of the effect of semantic constraints is proposed.

2. Background

Most work concerning the extraction of semantic relation focus on domain-independent relations (Snow et al., 2006; Chkolovski et al., 2004). In the general domain, the extraction of semantic relations between entities uses either statistical approaches (Hindle et al., 1990; Nazar et al. 2012) either techniques of machine learning as well as approaches based on the use of linguistic patterns (Hearst et al., 1992) and even approaches combining these two techniques.

Seen the difficulty, according the linguistic pattern selected to determine the kind of relations between two terms (because of the ambiguity of the linguistic pattern) Girju (Girju et al., 2003) suggested adding semantic constraints in order to discover meronymy relation. They determine 20 constraint thanks to a machine learning algorithm and obtain 83% of precision. Other lexical and syntactical constraints were applied to relations expressed by verbs (Ferber et al., 2011).

Concerning relation extraction (RE) in biomedical domain, there are four main techniques. finding co-occurrence (Jelier et al., 2005), using pattern or rule (Auger et al., 2008; Song et al., 2015; Rindflesch et al., 2000), supervised learning-based approaches (Song et al., 2015; Rink et al., 2011) and hybrid approach (Suchanek et al., 2006; Chowdhury et al., 2012). In the medical domain, RE systems often use, as knowledge base, UMLS (Unified Medical Language System) (Bodenreider, 2004). For instance, (Lee et al.) relied on UMLS to identify semantic relations between medical entities. Emberek, (2008) proposed a system to extract only four kinds of relations (Detect, Treat, Sign and Cure) between medical entities. Abacha et al., (2011) propose a method that allows extracting and annotating medical entities and relationships. They have used a rule based approach. Abacha uses a semi-automatic method for linguistic pattern generation while Lee uses a manual one. SemRep (Srinivasan et al., 2002) is a tool which allows to identify the semantic relations in the biomedical texts thanks to an approach based on a set of rules (Liu et al., 2012). F. Meng
et al., 2015 propose an MSA (multiple sequence alignment) based framework for generating automatically lexical patterns (but not semantic ones).

3. Our Approach: Semantic Patterns

The principle of our approach is quite simple (if not simplistic). As a first step, we identify compound terms. In a second step, we identify semantic relations between terms using semantic patterns. In the sentence fracture of orbital floor passing by the infra-orbital canal, the first step identify fracture_of_orbital_floor and infra-orbital_canal as an anatomical location. Thanks to the linguistic patterns passing by the, we are be able to validate the presence of a semantic relation between the two terms.

Our technique of entity extraction can be categorized in a dictionary-based approach and our method of relation extraction uses patterns or rules. The knowledge base on which our radiological reports entity extraction relies is the French lexical network JeuxDeMots (Lafourcade 2007). Although this network is general, it contains many specialty data, including medicine/radiology, which we have added within the framework IMAIOS project. The JDM network is a lexical-semantic graph for the French language whose lexical relations are generated both through GWAP (Games with a purpose) and via a contributory tool called Diko (manual insertion and automatic inferences with validation). At the time of this writing the JDM network contains over 23 millions relations between around 500,000 terms (many with inflected forms like plurals).

The identification of compound terms is made upstream compared to the content of JeuxDeMots network. We use the underscore to aggregate the two parts of a compound word so that it is considered as an entity at the time of the extraction (tibia_fracture). In our work, we are able to identify from JDM concepts like disease, symptom, anatomical location, characteristic (hyperintensity or hyperdensity), associated to terms. Moreover, to identify POS (part of speech) we use also JDM network. This information is available in the network.

3.1 Semantic Relation Extraction

Our relation extraction approach is based on the use of linguistic patterns, similar to (Embarek al, 2008). For each relation type (tableau 1), we build patterns and match them with the sentences to identify the correct relation.

<table>
<thead>
<tr>
<th>Types of relations</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>r_synonym</td>
<td>Synonyms or quasi synonyms</td>
</tr>
<tr>
<td>r_syn_strict</td>
<td>Strict synonyms (direct substitution is possible)</td>
</tr>
<tr>
<td>r_isa</td>
<td>Generic terms (hypernym)</td>
</tr>
<tr>
<td>r_charac</td>
<td>Typical characteristics</td>
</tr>
<tr>
<td>r_target</td>
<td>Target of disease (people, organ, etc)</td>
</tr>
<tr>
<td>r_symptom</td>
<td>Symptoms of disease</td>
</tr>
<tr>
<td>r_location</td>
<td>Typical locations</td>
</tr>
</tbody>
</table>

Table 1: list of relations using for extraction

We have chosen these 15 semantic relations for radiology report indexation following the advice of radiologists, but they can be of any general purpose. Some authors have already noted that the use of patterns is an effective method for automatic information extraction from corpora if they are efficiently designed (Embarek et al., 2008; Cimino et al., 1993).

3.2 Linguistic Patterns + Contraint = Semantic Patterns

For many relation types of the JDM network, we designed a set of linguistic patterns (tableau 2). These patterns are (for now) manually built through partial analysis of our corpus. In our experiment, we restricted ourselves to 42 semantic patterns, 12 of which are specific to medicine.

<table>
<thead>
<tr>
<th>Relations</th>
<th>Examples of patterns in English</th>
<th>Examples of patterns in French</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>E1 on the E2</td>
<td>E1 au niveau de E2</td>
</tr>
<tr>
<td>location</td>
<td>E1 in E2</td>
<td>E1 dans E2</td>
</tr>
<tr>
<td>location</td>
<td>E1 is on the E2</td>
<td>E1 se trouve dans E2</td>
</tr>
<tr>
<td>location</td>
<td>E1 passing by E2</td>
<td>E1 passant par E2</td>
</tr>
<tr>
<td>causes</td>
<td>E1 may trigger E2</td>
<td>E1 déclenchant E2</td>
</tr>
<tr>
<td>characteristic</td>
<td>E1 is characterized by E2</td>
<td>E1 est caractérisé par E2</td>
</tr>
<tr>
<td>synonym</td>
<td>E1 also called E2</td>
<td>E1 encore appelé</td>
</tr>
<tr>
<td>causes</td>
<td>E1 can produce E2</td>
<td>E1 peut produire E2</td>
</tr>
<tr>
<td>consequence</td>
<td>E1 cause E2</td>
<td>E1 provoque E2</td>
</tr>
<tr>
<td>hypernym</td>
<td>E1 is a E2</td>
<td>E1 est un E2</td>
</tr>
<tr>
<td>consequence</td>
<td>E1 leading to a E2</td>
<td>E1 menant à E2</td>
</tr>
</tbody>
</table>

Table 2: Examples of relations patterns. The actual patterns are in French.

For some relations several difficulties related to
ambiguity appeared. For the relation of location, we can distinguish two kinds of relations depending on the pattern. First the relation r_lieu (hepatocellular carcinoma is at the level of the liver). The second relation is holonomy. It defines the relationship between a term denoting the whole and a term denoting a member of the whole (femur r_holo lower limb). For some connectors (of in caudate lobe of liver) both relations are correct (caudate lobe r_lieu liver and caudate lobe r_holo liver). Note, that we also make use detection of immediate co-occurrences of entities for characteristic relation. For instance multifocal hepatocellular carcinoma (HCC) appears five time together, so we consider multifocal as a probable characteristic of HCC (HCC r_characteristic multifocal).

3.3 Constraint on Patterns

For some linguistic patterns it is very difficult to determine precisely the kind of relation (for example the French connector de (of), sous (under)). So, we have added some semantic constraints on linguistic patterns. A semantic constraint is a condition that should verify the reification of one of variable of the pattern. We may have any number of constraints on $x$ and $y$. We present some examples below:

- $x$ de $y$
  - If $x$ r_isa illness & $y$ r_isa anatomical_location $x$ r_location $y$
  - If $x$ r_isa anatomical_structure and $y$ r_isa anatomical_structure $x$ r_part_of $y$

- $x$ en $y$
  - If $x$ r_isa disease & $y$ r_isa anatomical_location $x$ r_location $y$
  - $x$ avec $y$
  - If $x$ r_isa disease & $y$ r_isa clinical sign $x$ r_symptom $y$
  - $x$ sous $y$
  - If $x$ r_isa disease & $y$ r_isa treatment $x$ r_against-l (is treated by) $y$

- $x$ au niveau du $y$
  - If $x$ r_isa disease & $y$ r_isa anatomical_location $x$ r_location $y$
  - $x$ due à $y$
  - If $x$ r_isa disease & $y$ r_isa microorganism || r_isa environment factor $x$ r_cause $y$

- $x$ porteur de $y$
  - If $x$ r_isa person & $y$ r_isa disease $x$ r_target$y$

- $x$ $y$

If $x$ r_pos Ady & $y$ r_pos Nom $x$ r_carac $y$

- $x$ $y$

If $x$ r_pos Noun & $y$ r_pos Adj $x$ r_carac $y$

In some cases, it is difficult to find some proper rules. Then we decide to incorporate some adjectival or adverbial multiword expressions in the network. For example, to deal with the noun phrase lesion in (en) hyperintensity, the expression in hyperintensity is added to the JDM network.

As mentioned above, we have crafted a small set of semantic patterns for testing purpose (we have, for the moment, identified 12 semantic patterns specific to medicine and 30 more general ones) and especially to detect properly relations for common connectors (de(of), du, en, avec, sous).

3.4 Algorithm

Starting from a given corpus (a radiological in our case), the procedure for extracting relations is informally the following:

Let $S$ the result set, being the empty set at initialization
Finding pattern occurrence in the text by moving a word window of size $n$
(essentially similar to using a Finite State Automata implementing the recognition of the linguistic patterns)
For each pattern occurrence applying constraints to the instantiated variables
if constrains are verified then the associated semantic relation is associated to $x$ and $y$, that is to say added to $S$

Return $S$

The value of $n$ is the length of the longest pattern (including both variables). The result set $S$ is weighted, the weight is the number of time that a given semantic relations between two given terms have been found in the text. 

Let’s take an example with the following semantic patterns:

"$x$ du $y$" | $x$ r_isa disease & $y$ r_isa anatomical_place $x$ r_location $y$

"$x$ du $y$" | $x$ r_isa anatomical_place & $y$ r_isa anatomical_place $x$ r_has_part $y$

From the noun phrase below (in French): « contusion intra-osseuse de l’os sous-chondral en zone portante du condyle latéral » we extract:
contusion intra-osseuse r_location os sous-chondral
os sous-chondral r_has_part condyle latéral

4. Results and Discussion

In order to evaluate the performances of our algorithm, we used classical measures namely precision, recall and F-measure (table 3). From a corpus of more 30 000 medical reports, we extracted a random subset of around 120 000 relation instances for the different relation types. About 800 of these relations were manually checked for evaluating precision. For assessing recall, we manually identified the relations in about 300 medical reports and then we applied our algorithm for comparison.

<table>
<thead>
<tr>
<th>relations</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>cause</td>
<td>74%</td>
<td>60%</td>
<td>66%</td>
</tr>
<tr>
<td>consequence</td>
<td>70%</td>
<td>62%</td>
<td>63.4%</td>
</tr>
<tr>
<td>location</td>
<td>48%</td>
<td>40%</td>
<td>43.6%</td>
</tr>
<tr>
<td>treatment</td>
<td>70%</td>
<td>60%</td>
<td>64.6%</td>
</tr>
<tr>
<td>part-of</td>
<td>32%</td>
<td>30%</td>
<td>31%</td>
</tr>
<tr>
<td>target</td>
<td>45%</td>
<td>40%</td>
<td>42.4%</td>
</tr>
<tr>
<td>characteristic</td>
<td>60%</td>
<td>58%</td>
<td>60%</td>
</tr>
<tr>
<td>lieu</td>
<td>45%</td>
<td>39%</td>
<td>41.7%</td>
</tr>
</tbody>
</table>

Table 3: Results of the extraction of semantic relations without constrains

<table>
<thead>
<tr>
<th>relations</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>cause</td>
<td>90%</td>
<td>60%</td>
<td>72%</td>
</tr>
<tr>
<td>consequence</td>
<td>89%</td>
<td>62%</td>
<td>73%</td>
</tr>
<tr>
<td>location</td>
<td>83%</td>
<td>40%</td>
<td>54%</td>
</tr>
<tr>
<td>treatment</td>
<td>88%</td>
<td>60%</td>
<td>71.3%</td>
</tr>
<tr>
<td>part-of</td>
<td>75%</td>
<td>30%</td>
<td>42.9%</td>
</tr>
<tr>
<td>target</td>
<td>80%</td>
<td>40%</td>
<td>53.3%</td>
</tr>
<tr>
<td>characteristic</td>
<td>88%</td>
<td>58%</td>
<td>70%</td>
</tr>
<tr>
<td>lieu</td>
<td>86%</td>
<td>40%</td>
<td>54.6%</td>
</tr>
</tbody>
</table>

Table 4: Results of the extraction of semantic relations with constrains

Globally the precision measure is quite good when we add constraints on semantic relations. The results show an improvement of the F1-measure. We notice that the method with semantic patterns improves the precision without modifying the recall. We can explain this by the fact that the addition of constraints allows a better characterization of the relation (by consequence an improvement of the precision) while the number of extracted relation does not vary (so the recall is not modified) because we do not add linguistic patterns. A comparison with other works seems a little difficult because we extract relations from specific corpus (radiology reports). Embarek et al., 2008 used linguistic patterns to extract disease-treatment relations with 78% for F1-measure. Abacha et al., 2011, that applied a method similar to our work, obtained 60.46% recall, 75.72% precision and 67.23% F-measure for the extraction of treatment relations. Other works about relation extraction in radiology extraction use machine learning techniques (Rink et al., 2011; Esuli et al., 2013). Some relations cannot be easily extracted because (not so) surprisingly they do not appear in our corpus. For instance, the relation of hypernym or synonymy is rarely present in our corpus which seems normal as radiologist knows already taxonomic information like that a carcinoma is a cancer. Another cause of limitation is the fact that some terms do not belong to the same sentence. Our approach shows much better results with the addition of semantic constraints on relations. In our experiment, these constraints can be assessed thanks to a large knowledge which is the JDM lexical network. We also have applied our proposed method to other corpora. For a cooking corpus of 45 000 recipes, we extracted 245 000 relations with a precision of 95% (manually evaluated on a sample of 755 relations). Furthermore, we extracted 789 000 relations for randomly Wikipedia pages with a precision of 92% (manually evaluated on a sample of 1250 relations). Hypernym extraction on Wikipedia articles has a precision of about 94%.

5. Conclusion

In this article, we have proposed a fast method for extracting semantic relations between entities. The method is based on the use of linguistic patterns with semantic constraints to be checked with the large JDM French lexical network. We showed that adding constraints improve tremendously both recall and precision, without having to rely of a POS tagger or syntactic analyzer.

Future work is to improve the coverage of our system by discovering lexical pattern from texts in an automatic way. A future work is to automate the generation of lexical pattern (Meng et al., 2015). We can take advantage of the knowledge between terms contained in the lexical network to infer the semantic relations hold by some recurrent linguistic patterns identified automatically by high occurrence level.

6. Bibliographical References

for Computational Linguistics.