Taking Advantage of Discursive Properties for Validating Hierarchical Semantic Relations from Parallel Enumerative Structures

Mouna Kamel(✉) and Cassia Trojahn

Institut de Recherche en Informatique de Toulouse, Toulouse, France
{mouna.kamel,cassia.trojahn}@irit.fr

Abstract. This paper presents an approach for automatically validating candidate hierarchical relations extracted from parallel enumerative structures. It relies on the discursive properties of these structures and on the combination of resources of different nature, a semantic network and a distributional resource. The results show an accuracy of between 0.50 and 0.67, with a gain of 0.11 when combining the two resources.

1 Introduction

Relation extraction is a key task in ontology learning from texts. The identification of candidate relations has been the subject of large body of literature and many approaches have been proposed (linguistic, statistical or hybrid approaches, based or not on learning methods). However, this is an error-prone step (imprecise lexico-syntactic patterns, accuracy of learning techniques under 100 %, chaining of NLP tools in pre-processing steps, etc.). Validating candidate relations is a crucial step before integrating them into semantic resources.

This paper concerns the validation of candidate hierarchical relations, the backbone of ontologies. While manual validation is a time-consuming task requiring domain expert judges, automatic ones rely on external semantic resources (such as WordNet, BabelNet), which are usually non domain-specific, or gold standards, which may suffer of imperfections or low domain coverage. The proposal here relies on the extraction of hierarchical semantic relations from parallel enumerative structures (called hereafter PES) [4]. This choice is motivated by the following reasons: (1) PES often carry hierarchical relations; (2) they are frequent in corpora, especially in scientific or encyclopedic texts (rich sources of semantic relations); and (3) they have well-established discursive properties bringing up a semantic unit within the structure. The originality of our approach lies in the discourse properties of PES for disambiguating candidate relations and in the combination of two complementary external resources, a semantic network and a distributional resource. While the semantic network allows for validating the candidate relations with a good level of precision, the distributional resource, which does not specify the nature of the relation but offers a good coverage, allows for emerging new relations, which may enrich the network itself. Although evaluated for the French language, the approach remains reproducible for any other language.

2 Parallel Enumerative Structures

An enumerative structure is a textual structure expressing hierarchical knowledge through different components: a primer, a list of items (at least two) constituting the enumeration, and possibly a conclusion. Different typologies have been proposed [3,5]. Here, we consider enumeratives structures for which the enumeration items are functionally equivalent (from a syntactic and rhetoric point of view) (Fig. 1). From a discursive point of view, the items are independent in a given context: they are in turn connected by a multi-nuclear rhetoric relation (or coordination), the first item being linked to the primer by a nuclear-satellite relation (or subordination) (Fig. 2). According to the RST (Rhetorical Structure Theory) [2], if "DU_j (where DU corresponds to Discourse Unit) is subordinated to DU_i, hence each DU_k coordinated with DU_j is subordinate to DU_i". Thereby, N nuclear-satellite relations between DU_0 and DU_i, for $i=1,...,N$ (if N is the number of items in the ES) can be inferred. These N relations can be specialised in N semantic relations $R(H, h_i)_{i=1,...,N}$ of same nature, where H correspond to a term of DU_0, and h_i to a term of DU_i. From Fig. 2, three relations can be identified: *R(disease, Cholera)*, *R(disease, Colorectal cancer)*, and *R(disease, Diverticulitis)*.

```
There are a number of diseases affecting the gastrointestinal system:
- Cholera
- Colorectal cancer
- Diverticulitis
```

Fig. 1. Example of PES.

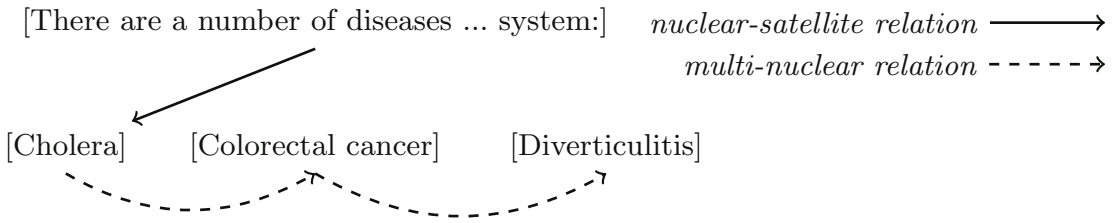

Fig. 2. Discursive representation of the PES of Fig. 1 according to the RST

3 Proposed Approach

The validation principle exploits the discourse properties of PES to jointly validate the relations $R(H, h_i)$ (i = 1, .., N) where R is the hypernym relation:

1. if $R(H, h_i)$ corresponds to an entry in the semantic network SN, $R(H, h_i)$ is validated.
2. if $R(H, h_i)$ has no entry in SN, but an entry corresponding to $R(H, h_j)$ exists in SN and h_i is a neighbour of h_j in the distributional resource DR, then $R(H, h_i)$ is validated.

From SN, we retrieve $Synsets(H)$, the synsets of H, and $SuperHyperyms^k_{SN}(h_i)$, the hypernym synsets of h_i of rank k (k being the maximum length of the path from h_i to one of its hypernym synsets in SN, based on a depth-first search strategy). From DR we retrieve $p(h_i, h_j)$, the semantic proximity between h_i and h_j. This process is described in Algorithm 1.

Algorithm 1. Algorithm for validating a set of relations from a PES

$V \leftarrow \emptyset, \overline{V} \leftarrow \bigcup_{i=1}^{N} R_i$ *// V set of validated relations, $\overline{V}$ set of non validated relations*
for each relation $R_i(H, h_i) \in \overline{V}$ **do**
  if $SuperHyperyms^k_{SN}(h_i) \cap Synsets(H) \neq \emptyset$ **then**
    // H is a hypernym of h_i
    $validate(R_i(H, h_i)) \leftarrow 1$ *// R_i is validated*
    $V = V \bigcup \{R_i\}$
    $\overline{V} = \overline{V} - \{R_i\}$
  end if
end for
if $V \neq \emptyset$ *et* $\overline{V} \neq \emptyset$ **then**
  //at least one relation has been validated and one has not been yet
  for each relation $R_j(H, h_j) \in \overline{V}$ **do**
    $validate(R_j(H, h_j)) = \frac{\sum_{R_i \in V} p(h_i, h_j)}{|V|}$
    //proximity between h_j and the hyponyms h_i from the validated relations
  end for
end if

4 Experimentation

Data set and resources. The evaluation data set[1] is composed of 67 PES involving 262 candidate relations, automatically extracted from Wikipedia pages [4]. These relations have been manually validated by two annotators in a double-blind process. 27 conflicts were identified and resolved. 206 relations were assessed as correct and 56 as incorrect. This set constitutes our *gold standard*. With respect to the resources, we have used the multilingual semantic network *BabelNet* [6] and the distributional resource *Voisins de Wikipédia* [1]. They have been chosen because they support French language and they are built from the same corpus as the one used for constructing the evaluation data set.

Results and discussion. Two sets of candidate relations were considered (Table 1): S, the whole set of true positive relations from the *gold standard* (206 relations) and S_{BN}, the subset of S for which H exists in *BabelNet* (116 relations). For both sets, 76 out of 78 relations were correctly validated by the system. 12 out of 76 have been correctly validated thanks to the distributional resource, what corresponds to an improvement of the performance up to 11 %.

[1] Available at https://www.irit.fr/~Cassia.Trojahn/PES.zip.

Table 1. Overall results of the validation process combining both SN and DR. (+) corresponds to the specific gain of using DR.

	Precision ($+_{DR}$)	Recall ($+_{DR}$)	FMeasure ($+_{DR}$)	Accuracy ($+_{DR}$)
S	.97 (+0.0)	.37 (+.06)	.54 (+.07)	.50 (+.05)
s_{BN}	.97 (+0.0)	.66 (+.11)	.78 (+0.8)	.67 (+.11)

In terms of recall, we have a lower performance (76 relations out of 206 for S but 76 out of 116 for S_{BN}). In terms of accuracy, 130 relations have been validated (out of 262) for the set S and 88 relations (out of 131) for the set S_{BN}.

Although the precision is quite high, we could identify the reasons for the noisy cases. It is due to the fact that we are using BabelSynsets which group terms of similar meaning. For instance, for the candidate relation *R(country,Horn of Africa)*, the BabelSynset *bn:00028934n* = {*land, dry land, earth, ground, terra firma*} belongs to the intersection of the sets $SuperHyperyms^3_{BN}$(*Horn of Africa*) and $Synsets$(*country*). With respect to the low recall, we observed two main phenomena. First, 62 hypernyms (from S) have no entries in *BabelNet*. In this case, no relation within the PES could be validated. Second, considering $k = 3$ (empirically chosen) as maximum length of the path from h_i to one of its hypernyms seems to be insufficient. We could also observe that the distributional resource allows for identifying missing entries in the semantic network. For example, the relation *R(chromosomal abnormality,insertion)* was validated due to the fact that *insertion* and *deletion* are semantically near in the distributional resource. Although the entries in this resource overwhelmingly correspond to single words and 40 % of our hyponyms correspond to compounds, we improved the performance up to 11 % when combining both resources. Distributional resources supporting compounds may further improve our results.

5 Conclusions and Future Work

This paper proposed an approach for automatically validating semantic relations, relying on discursive properties and combining a semantic network and a distributional resource. As future work, we plan to exploit alternative resources (in particular, distributional resources with compounds), analyse the trade-off between depth-first and breath-first search strategies and their computational complexity, exploiting larger semantic networks or combining several resources together. We intent as well to extend our approach to validate other semantic relations like meronymy, synonymy and antonym.

Acknowledgement. Cassia Trojahn is partially supported by the French FUI SparkinData project.

References

1. Adam, C., Fabre, C., Muller, P.: Évaluer et améliorer une ressource distributionnelle: protocole d'annotation de liens sémantiques. TAL **54**(1), 71–97 (2013)
2. Asher, N.: Reference to abstract objects in discourse: a philosophical semantics for natural language metaphysics. In: SLAP, vol. 50. Kluwer (1993)
3. Christophe, L.: Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés. PhD thesis (2000)
4. Fauconnier, J.P., Kamel, M.: Discovering hypernymy relations using text layout. In: Joint Conference on Lexical and Computational Semantics, Denver, pp. 249–258. ACL (2015)
5. Hovy, E., Arens, Y.: Readings in intelligent user interfaces. In: Automatic Generation of Formatted Text, pp. 256–262. Morgan Kaufmann Publishers (1998)
6. Navigli, R., Ponzetto, S.P.: BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif. Intell. **193**, 217–250 (2012)