Discovering Hypernymy Relations using Text Layout

Jean-Philippe Fauconnier, Mouna Kamel

Discovering Hypernymy Relations using Text Layout

Jean-Philippe Fauconnier
Institut de Recherche en
Informatique de Toulouse
118, Route de Narbonne
31062 Toulouse, France
faucon@irit.fr

Mouna Kamel
Institut de Recherche en
Informatique de Toulouse
118, Route de Narbonne
31062 Toulouse, France
kamel@irit.fr

Abstract

Hypernymy relation acquisition has been widely investigated, especially because taxonomies, which often constitute the backbone structure of semantic resources are structured using this type of relations. Although lots of approaches have been dedicated to this task, most of them analyze only the written text. However relations between not necessarily contiguous textual units can be expressed, thanks to typographical or dispositional markers. Such relations, which are out of reach of standard NLP tools, have been investigated in well specified layout contexts. Our aim is to improve the relation extraction task considering both the plain text and the layout. We are proposing here a method which combines layout, discourse and terminological analyses, and performs a structured prediction. We focused on textual structures which correspond to a well defined discourse structure and which often bear hypernymy relations. This type of structure encompasses titles and sub-titles, or enumerative structures. The results achieve a precision of about 60%.

1 Introduction

The hypernymy relation acquisition task is a widely studied problem, especially because taxonomies, which often constitute the backbone structure of semantic resources like ontologies, are structured using this type of relations. Although this task has been addressed in literature, most of the publications report analyses based on the written text only, usually at the phrase or sentence level. However, a written text is not merely a set of words or sentences. When producing a document, a writer may use various layout means, in addition to strictly linguistics devices such as syntactic arrangement or rhetorical forms. Relations between textual units that are not necessarily contiguous can thus be expressed thanks to typographical or dispositional markers. Such relations, which are out of reach of standard NLP tools, have been studied within some specific layout contexts. Our aim is to improve the relation extraction task by considering both the plain text and the layout. This means (1) identifying hierarchical structures within the text using only layout, (2) identifying relations carried by these structures, using both lexico-syntactic and layout features.

Such an approach is deemed novel for at least two reasons. It combines layout, discourse and terminological analyses to bridge the gap between the document layout and lexical resources. Moreover, it makes a structured prediction of the whole hierarchical structure according to the set of visual and discourse properties, rather than making decisions only based on parts of this structure, as usually performed.

The main strength of our approach is its applicability to different document formats as well to several domains. It should be highlighted that encyclopedic, technical or scientific documents, which are often analyzed for building semantic resources, are most of the time strongly structured. Our approach has been implemented for the French language, for which only few resources are currently available. In this paper we focus on specific textual

structures which share the same discourse properties and that are expected to bear hypernymy relations. They encompass for instance titles/sub-titles, or enumerative structures.

The paper is organized as follows. Some related works about hypernymy relation identification are reported in section 2. Section 3 presents the theoretical framework on which the proposed approach is based. Sections 4 and 5 respectively describe transitions from the text layout to its discourse representation and from this discourse structure to the terminological structure. Finally we draw conclusions and propose some perspectives.

2 Related works

The task of extracting hypernymy relations (it may also be denoted as generic/specific, taxonomic, is-a or instance-of relations) is critical for building semantic resources and for semantic content authoring. Several parameters concerning corpora may affect the methods used for this task: the natural language quality (carefully written or informal), the textual genre (scientific, technical documents, newspapers, etc.), technical properties (corpus size, format), the level of precision of the resource (thesaurus, lightweight or full-fledged ontology), the degree of structuring, etc. This task may be carried out by using the proper text and/or external pre-existing resources. Various methods for exploiting plain text exist using techniques such as regular expressions (also known as lexico-syntactic patterns) (Hearst, 1992), classification using supervised or unsupervised learning (Snow et al., 2004; Alfonseca and Manandhar, 2002), distributional analysis (Lenci and Benotto, 2012) or Formal Concepts Analysis (Cimiano et al., 2005). In the Information Retrieval area, the relevant terms are extracted from documents and organized into hierarchies (Sánchez and Moreno, 2005).

Works on the document structure and on the discourse relations that it conveys have been carried out by the NLP community. Among these are the Document Structure Theory (Power et al., 2003), and the DArt_{bio} system (Bateman et al., 2001). These approaches offer strong theoretical frameworks, but they were only implemented from a text generation point of view.

With regard to the relation extraction task using layout, two categories of approaches may be distinguished. The first one encompasses approaches exploiting documents written in a markup language. The semantics of these tags and their nested structure is used to build semantic resources. For instance, collection of XML documents have been analyzed to build ontologies (Kamel and Aussenac-Gilles, 2009), while collection of HTML or MediaWiki documents have been exploited to build taxonomies (Sumida and Torisawa, 2008).

The second category gathers approaches exploiting specific documents or parts of documents, for which the semantics of the layout is strictly defined. Let us mention dictionaries and thesaurus (Jannink and Wiederhold, 1999) or specific and well localized textual structures such as category field (Chernov et al., 2006; Suchanek et al., 2007) or infoboxes (Auer et al., 2007) from Wikipedia pages. In some cases, these specific textual structures are also expressed thanks to a markup language. All these works implement symbolic as well as machine learning techniques.

Our approach is similar to the one followed by Sumida and Torisawa (2008) which analyzes a structured text according to the following steps: (1) they represent the document structure from a limited set of tags (headings, bulleted lists, ordered lists and definition lists), (2) they link two tagged strings when the first one is in the scope of the second one, and (3) they use lexico-syntactic and layout features for selecting hypernymy relations, with the help of a machine learning algorithm. Some attempts have been made for improving these results (Oh et al., 2009; Yamada et al., 2009). However our work differs in two points: we aimed to be more generic by proposing a discourse structure of layout that can be inferred from different document formats, and we propose to find out the relation arguments (hypernym-hyponym term pairs) by analyzing propositional contents. Prior to describing the implemented processes, the underlying principles of our approach will be reported in the next section.

3 Underlying principles of our approach

We rely on principles of discourse theories and on knowledge models for respectively formalizing text layout and identifying hypernymy relations.

3.1 Discourse analysis of the layout

Several discourse theories exist. Their starting point lies in the idea that a text is not just a collection of sentences, but it also includes relations between all these sentences that ensure its coherence (Mann and Thompson, 1988; Asher and Lascarides, 2003). Discourse analysis aims at observing the discourse coherence from a rhetorical point of view (the intention of the author) or from a semantic point of view (the description of the world). A discourse analysis is a three step process: splitting the text into Discourse Units (DU), ensuring the attachment between DUs, and then labeling links between DUs with discourse relations. Discourse relations may be divided into two categories: nucleus-satellite (or subordinate) relations which link an important argument to an argument supporting background information, and multi-nuclear (or coordinate) relations which link arguments of equal importance. Most of discourse theories acknowledge that a discourse is hierarchically structured thanks to discourse relations.

Text layout supports a large part of semantics and participates to the coherence of the text; it thus contributes to the elaboration of the discourse. Therefore, we adapted the discourse analysis to treat the layout, according to the following principles:

- a DU corresponds to a visual unit (a bloc);
- two units sharing the same role (title, paragraph, etc.) and the same typographic and dispositional markers are linked with a multi-nuclear relation; otherwise, they are linked with a nuclear-satellite relation.

An example[1] of document from Wikipedia and the tree which results from the discourse analysis of its layout is given (Figure 1). In the following figures, we represent nucleus-satellite relations with solid lines and multi-nuclear relations with dashed lines.

[1] http://fr.wikipedia.org/wiki/Redécentralisation_d'Internet

We are currently interested in discourse structures displaying the following properties:

- n DUs are linked with multi-nuclear relations;
- one of these coordinated DU is linked to another DU with a nucleus-satellite relation.

Figure 2 gives a representation of such a discourse structure according to the Rhetorical Structure Theory (Mann and Thompson, 1988).

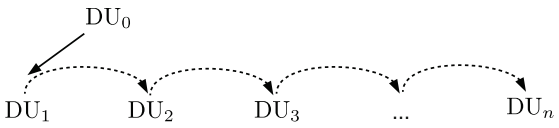

Figure 2: Rhetorical representation of the discourse structure of interest

Although there is only one explicit nucleus-satellite relation, this kind of structure involves n implicit nucleus-satellite relations (between DU_0 and DU_i ($2 \leq i \leq n$)). Indeed, from a discourse point of view, if a DU_j is subordinated to a DU_i, then all DU_k coordinated to DU_j, are subordinated to DU_i. As mentioned above, this kind of discourse structure encompasses textual structures such as titles/sub-titles and enumerative structures which are frequent in structured documents, and which often convey hypernymy relation. In that context, the hypernym is borne by the DU_0 and each DU_i ($1 \leq i \leq n$) bears at least one hyponym.

3.2 Knowledge models for hypernymy relation identification

Hypernymy relation identification is carried out in two stages: specifying if the relation is hypernymic and, if appropriate, identifying its arguments. The first stage relies on linguistic regularities denoting a hypernymy relation, regularities which are expressed thanks to lexical, syntactic, typographical and dispositional clues.

The second stage is based on a graph representation. Rather than independently identifying links between the hypernym and each potential hyponym, we take advantage from the fact that writers use the same syntactic and visual skills (recognized by a textual parallelism) for expressing knowledge units of equal rhetorical importance. Generally, these salient units are semantically linked and belong to a same lexical field.

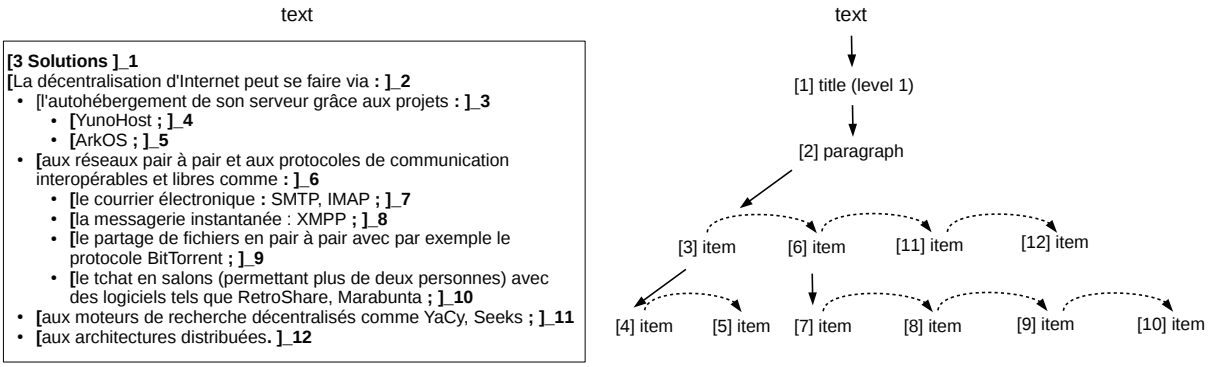

Figure 1: Example of a discourse analysis of text layout

Thus, we represent each discourse structure of interest bearing a hypernymy relation as a directed acyclic graph (DAG), where the nodes are terms and the edges are possible relations between them. This DAG is decomposed into layers, each layer i gathering nodes corresponding to terms of a given DU_i $(0 \leq i \leq n)$. Each node of a layer i $(0 \leq i \leq (n-1))$ is connected by directed edges to all nodes of the layer $i+1$. A root node is added on the top of the DAG. Figure 3 presents an example of this DAG.

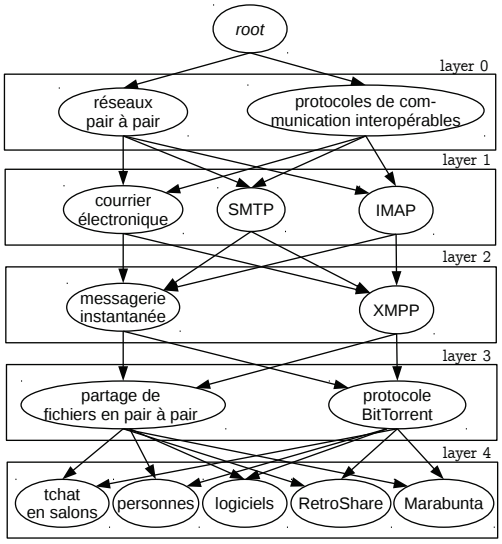

Figure 3: Example of a DAG

We weight the edges according to the inverse similarity of terms they link. Thus, the terms in the lower-cost path starting from the root and ending at the last layer are maximally cohesive. A flatter representation does not allow this structured prediction.

4 From text layout to its discourse representation

To elicit discourse structures from text layout, the system detects visuals units and labels them with their role (paragraph, title, footnote, etc.) in the text. Then, it links the labeled units using discourse relations (nucleus-satellite or multi-nuclear) in order to produce a discourse tree.

We are currently able to process two types of documents: documents written in a markup language and documents in PDF format. It is obvious that tags of markup languages both delimit blocs and give their role. Getting the visual structure is thus straightforward. Conversely, PDF documents do not benefit from such tags. So we used the LAPDF-Text tool (Ramakrishnan et al., 2012) which is based on a geometric analysis for detecting blocs, and we have implemented a machine learning method for labeling these blocs. The features include typographical markers (size of fonts, emphasis markers, etc.) and dispositional one (margins, position in page, etc.).

For labeling relations, we used an adapted version of the shift-reduce algorithm as (Marcu, 1999) did. We thus obtain a dependency tree representing the discourse structure of the text layout. We evaluate this process on a corpus of PDF documents (documents written in a markup language pose no problem). Results are good since we obtain an accuracy of 80.46% for labeling blocs, and an accuracy of 97.23% for labeling discourse relations (Fauconnier et al., 2014). The whole process has been implemented in the LaToe[2] tool.

[2] http://github.com/fauconnier/LaToe

Finally, the extraction of discourse structures of interest may be done easily by means of tree patterns (Levy and Andrew, 2006).

5 From layout discourse structure to terminological structure

We wish to elicit possible hypernymy relations from identified discourse structures of interest. This task involves a two-step process. The first step consists in specifying the nature of the relation borne by these structures. The second step aims at identifying the related terms (the relation arguments). These steps have been independently evaluated on an annotated corpus, while the whole system has been evaluated on another not annotated corpus. Corpora and evaluation protocols are described in the next section.

5.1 Corpora and evaluation protocols

The annotated corpus includes 166 French Wikipedia pages corresponding to urban and environmental planning. 745 discourse structures of interest were annotated by 3 annotators (2 students in Linguistics, and an expert in knowledge engineering) according to a guideline. The annotation task for each discourse structure of interest has consisted in annotating the nucleus-satellite relation as hypernymy or not, and when required, in annotating the terms involved in the relation. For the first stage, we have calculated a degree of inter-annotator agreement (Fleiss et al., 1979) and obtained a kappa of 0.54. The second stage was evaluated as a named entity recognition task (Tateisi et al., 2000) for which we have obtained an F-measure of 79.44. From this dataset, 80% of the discourse structures of interest were randomly chosen to constitute the development set, and the remaining 20% were used for the test set. The tasks described below were tuned on the development set using a k-10 cross-validation. The evaluation is done using the precision, the recall and the F-measure metrics.

A second evaluation for the entire system was led on two corpora respectively made of Wikipedia pages from two domains: *Transport* and *Computer Science*. For each domain, we have randomly selected 400 pages from a French Wikipedia Dump (2014-09-28). Since those copora are not manually annotated, we have only reported the precision.

5.2 Qualifying the nucleus-satellite relation

Hypernymy relations present lexical, syntactic, typographical and dispositional regularities in the text. The recognition of these relations is thus based on the analysis of these regularities within the two DUs explicitly linked by the nucleus-satellite relation. We consider this problem as a binary classification one: each discourse structure is assigned to either the *Hypernymy-Structure* class or the *nonHypernymy-Structure* class. The *Hypernymy-Structure* class encompasses discourse structures with a nucleus-satellite relation bearing a hypernymy, whereas the *nonHypernymy-Structure* one gathers all others discourse structures. In the example given in figure 1, the discourse structures constituted of DUs {3,4,5} and {6,7,8,9,10} would be classified as *Hypernymy-Structure*, while this constituted of DUs {2,3,6,11,12} would be assigned to the *nonHypernymy-Structure* class.

For this purpose, we applied feature functions (summarized in table 1) in order to map the two DUs linked by the explicit nucleus-satellite relation into a numerical vector which is submitted to a classifier. The feature functions were defined according to background knowledge and were selected on the basis of a Pearson's correlation.

Features	Description
POS	Unigrams of parts of speech
Position	Position of a token in a DU
Markers	Boolean indicating whether a token belongs to a predefined lexicon
Gram	Boolean indicating whether the last sentence of a DU shows a syntactic hole
Punc	Returns the last punctuation of a DU
NbToken	Number of tokens in a DU
NbSent	Number of sentences in a DU

Table 1: Main features for qualifying the relation

We have compared two types of classifiers: a linear one which generalizes well, but may produce more misclassifications when data distribution presents a large spread, and a non-linear one which may lead to a model separating well the training set but with an overfitting risk. We respectively used a Maximum Entropy classifier (MaxEnt) (Berger et al., 1996) and a Support Vector Machine (SVM) with a Gaussian kernel (Cortes and Vapnik, 1995).

The morphological and lexical information used were obtained from the French dependency parser Talismane (Urieli, 2013). For the classifiers, we have used the OpenNLP[3] library for the MaxEnt and the LIBSVM implementation of the SVM[4]. This task has been evaluated against a majority baseline which better reflects the reality because of the asymmetry of the relation distribution. Table 2 presents the results. The two supervised strategies outperform significantly the baseline (p-values$<$0.01)[5].

Strategies	Prec.	Rec.	F1
MaxEnt	**78.01**	84.78	81.25
SVM	74.77	**90.22**	81.77
Baseline	63.01	100.0	77.31

Table 2: Results for qualifying the relation

Regarding the F-measure metric, the difference between the MaxEnt and the SVM is not significant. We observe that the MaxEnt achieves the best precision, while the SVM reaches the best recall. These results are not surprising since the SVM decision boundary seems to be biased by outliers, thus increasing the false positive rate on unseen data.

5.3 Identifying the terms linked by the hypernymy relation

We have now to identify terms linked by the hypernymy relation. As previously mentioned we build a DAG reflecting all possible relations between terms of the DUs, to find the lower-cost path which represents the most cohesive sequence of terms.

If we consider the discourse structure constituted of DUs {6,7,8,9,10} in figure 1, the retrieved path from the corresponding DAG (figure 3) would be ["protocoles de communication interopérables" (*interoperable communication protocols*), "courrier électronique" (*email*), "messagerie instantanée" (*instant messaging*), "partage de fichiers en pair à pair" (*peer-to-peer file sharing*), "tchat en salons" (*chat room*)]. Then, an example of hypernymy relation would be "courrier électronique" (*email*) is a kind of "protocoles de communication interopérables" (*interoperable communication protocols*).

[3] http://opennlp.apache.org/

[4] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[5] The p-values are calculated using a paired t-test.

The cost of an edge is defined using the following function:

$$\text{cost}(< T_i^j, T_{i+1}^k >) = 1 - p(y|T_i^j, T_{i+1}^k)$$

where T_i^j is the j-th term of DU_i. The probability assigned to the outcome y measures the likeliness that both terms are linked. This probability is conditioned by lexical and dispositional clues. Since it is expected that terms involved in the relation share the same lexical field, we also consider the cosine similarity between the term vectors. All those clues are mapped into a numerical vector using feature functions summarized in table 3.

Features	Description
POS_c	Context of a term (bigrams and unigrams of parts of speech)
POS_t	Parts of speech of a term
Role	Role of a DU
Visual	Boolean indicating whether a pair of terms share the same visual properties
Position_t	Value indicating a term position
Position_d	Position of a DU in the whole document
Coord	For a DU, presence of coordinated DUs
Sub	For a DU, presence of subordinated DUs
Level	Value indicating the level of a DU in the structure of document
Punc	Returns the last punctuation of a DU
NbToken	Number of tokens in a DU
NbSent	Number of sentences in a DU
COS	Cosine similarity for a pair of terms

Table 3: Main features for the terms recognition

We built two models based on supervised probabilistic classifiers since characteristics of links between a hypernym and a hyponym are different from those between two hyponyms. The first model considers only the edges between layer 0 and layer 1 (hypernym-hyponym link), whereas the second one is dedicated to the edges of remaining layers (hyponym-hyponym link).

For this step, we used ACABIT (Daille, 1996) and YaTeA (Aubin and Hamon, 2006) for extracting terms. The cosine similarity is based on a distributional model constructed with the word2vec tool (Mikolov et al., 2013) and the French corpus FrWac (Baroni et al., 2009). We have learned the models using a Maximum Entropy classifier.

For computing the lower-cost path, we use an A* search algorithm because it can handle large search space with an admissible heuristic. The estimated cost of a path P, a sequence of edges from the root to a given term, is defined by:

$$f(P) = g(P) + h(P)$$

The function $g(P)$ calculates the real cost along the path P and it is defined by:

$$g(P) = \sum_{<T_i^j, T_{i+1}^k> \in P} \text{cost}(< T_i^j, T_{i+1}^k >)$$

The heuristic $h(P)$ is a greedy function which picks a new path with the minimal cost over d layers and returns its cost:

$$h(P) = g(l_d(P))$$

The function $l_d(P)$ is defined recursively: $l_0(P)$ is the empty path. Assume $l_d(P)$ is defined and $T_{i_d}^{j_d}$ is the last node reached on the path formed by the concatenation of P and $l_d(P)$, then we define:

$$l_{d+1}(P) = l_d(P).< T_{i_d}^{j_d}, T_{i_d+1}^m >$$

where m is the index of the term with the lower cost edge and belonging to the layer $i_d + 1$:

$$m = \underset{k<|\text{layer } i_d+1|}{\text{argmin}} \text{cost}(< T_{i_d}^{j_d}, T_{i_d+1}^k >)$$

This heuristic is admissible by definition. We set d=3 because it is a good tradeoff between the number of operations and the number of iterations during the A* search.

In order to evaluate this task, we compare it to a baseline and two vector-based approaches. The baseline works on the assumption that two related terms belong to a same window of words; then it takes the last term of the layer 0 as hypernym, and the first term of each layer i ($1 \leq i \leq n$) as hyponym. The two other strategies use a cosine similarity (calculated with respectively 200- and 500-dimensional vectors) for the costs estimation. Table 4 presents the results.

The MaxEnt achieves the best F-measure and outperforms the others proposed strategies. The vector-based strategies present interesting precisions, which seems to confirm a correlation between the lexical cohesion of terms and their likelihood of being involved in a relation.

Strategies	Prec.	Rec.	F1
MaxEnt	78.98	69.09	**73.71**
w2v-200	66.52	30.10	41.45
w2v-500	83.71	30.10	44.28
Baseline	48.37	69.09	56.91

Table 4: Results for terms recognition

To lead additional evaluations we define the score of a path as the mean of its costs, and we select results using a list of threshold values: only the paths with a score lower than a given threshold are returned. Figure 4 shows the Precision-Recall curves using the whole list of threshold values.

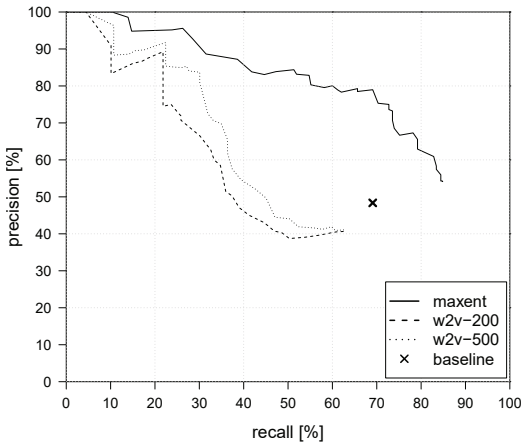

Figure 4: Comparison between the baseline, the vector-based strategies and the MaxEnt

5.4 Evaluation of the whole system

In this section, we report the results for the whole process applied on two corpora made of Wikipedia pages from two domains: *Transport* and *Computer Science*. For each of them, we applied a discourse analysis of the layout, and we extracted the hypernym-hyponym pairs. This extraction was done with a Maximum Entropy classifier which has shown a good precision for the two tasks described before. The retrieved pairs were ranked according to the score of the path they belong to. Finally, we

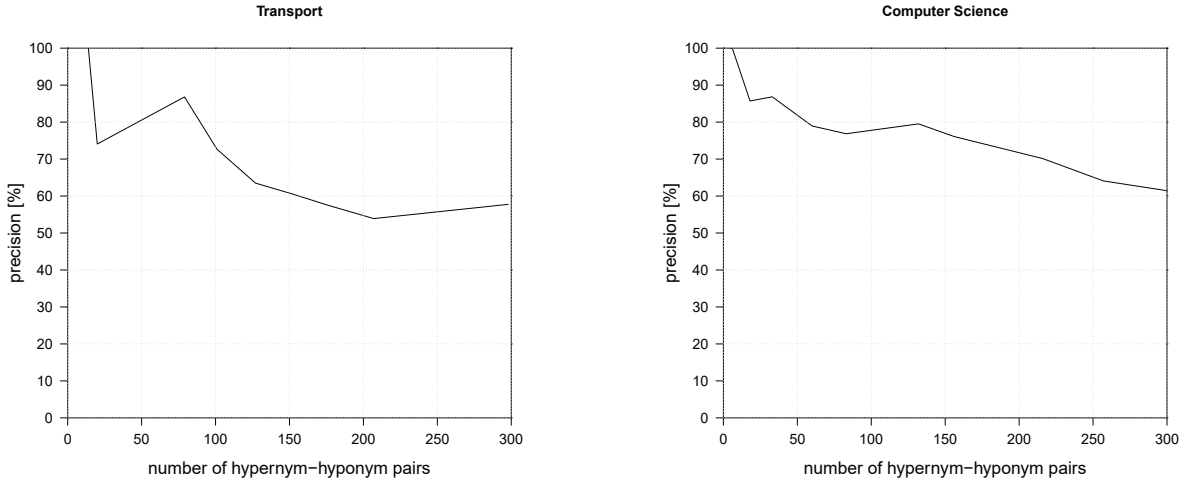

Figure 5: Precision curves for two domains of Wikipedia

manually checked the first 500 pairs. The curves in figure 5 indicate the precision. For the two domains, around 300 pairs were retrieved with a precision of about 60% for the highest threshold. Table 5 presents examples of extracted relations. The terms noted with a symbol '*' are considered as errors.

hypernyms	hyponyms
transporteurs frigorifiques (refrigerated transporters)	STEF, transporteur*, Groupe Delanchy, Norbert Dentressangle, Groupe Malherbe, Madrias
pôles d'échanges (interchange stations)	Gare de la Part Dieu, Centre intermodal d'échanges de Limoges, Union Station à Toronto
transmission (transmission)	Courte distance*, Moyenne distance*, Longue distance*

Table 5: Examples of extracted relations

We have identified the main sources of error. The most common arises from nested discourse structures. In this case, intermediate DUs often specify contexts, and therefore do not contain the searched hyponyms. This is the case in the last example of table 5 where the retrieved hyponyms for "transmission" (transmission) are "Courte distance" (*Short distance*), "Moyenne distance" (*Medium distance*) and "Longue distance" (*Long distance*).

Another error comes from a confusion between hypernymy and meronymy relations, which are both hierarchical. The fact that these two relations share the same linguistic properties may explain this confusion (Ittoo and Bouma, 2009). Furthermore we are still faced with classical linguistic problems which are out of the scope of this paper: anaphora, ellipse, coreference, etc.

Finally, we ignore cases where the hypernymy relation is reversed, i.e. when the hyponym is localized into the nucleus DU and its hypernym into a satellite DU. Clues that we use are not enough discriminating at this level.

6 Conclusion

In this paper we investigate a new way for extracting hypernymy relations, exploiting the text layout which expresses hierarchical relations and for which standard NLP tools are not suitable.

The system implements a two steps process: (1) a discourse analysis of the text layout, and (2) a hypernymy relation identification within specific discourse structures. We first evaluate each module independently (discourse analysis of the layout, identification of the nature of the relation, and identification of arguments of the relation), and we obtain accuracies of about 80% and 97% for the discourse analysis, and F-measures of about 81% and 73% for the relation extraction. We then evaluate the whole process and we obtain a precision of about 60%.

One way to improve this work is to extend this analysis to other hierarchical relations. We plan to investigate more advanced techniques offered by distributional semantic models in order to discriminate hypernymy relation from meronymy ones.

Another way is to extend the scope of investigation of the layout to take into account new discursive structures. Moreover, a subsequent step to this work is its large scale application on collections of structured web documents (such as Wikipedia pages) in order to build semantic resources and to share them with the community.

References

Enrique Alfonseca and Suresh Manandhar. 2002. Improving an ontology refinement method with hyponymy patterns. *cell*, 4081:0–0087.

N. Asher and A. Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

John Bateman, Thomas Kamps, Jörg Kleinz, and Klaus Reichenberger. 2001. Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, 27(3):409–449.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Sergey Chernov, Tereza Iofciu, Wolfgang Nejdl, and Xuan Zhou. 2006. Extracting semantics relationships between wikipedia categories. *SemWiki*, 206.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, 24:305–339.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Béatrice Daille. 1996. Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act: Combining symbolic and statistical approaches to language*, 1:49–66.

Jean-Philippe Fauconnier, Laurent Sorin, Mouna Kamel, Mustapha Mojahid, and Nathalie Aussenac-Gilles. 2014. Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux. In *Actes de la 21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, pages 340–351.

Joseph L Fleiss, John C Nee, and J Richard Landis. 1979. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86(5):974–977.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545. Association for Computational Linguistics.

Ashwin Ittoo and Gosse Bouma. 2009. Semantic selectional restrictions for disambiguating meronymy relations. In *proceedings of CLIN09: The 19th Computational Linguistics in the Netherlands meeting, to appear*.

Jan Jannink and Gio Wiederhold. 1999. Thesaurus entry extraction from an on-line dictionary. In *Proceedings of Fusion*, volume 99. Citeseer.

Mouna Kamel and Nathalie Aussenac-Gilles. 2009. How can document structure improve ontology learning? In *Workshop on Semantic Annotation and Knowledge Markup collocated with K-CAP*.

Alessandro Lenci and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234. Citeseer.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1999. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 365–372. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa. 2009. Bilingual co-training for monolingual hyponymy-relation acquisition. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNL*, pages 432–440. Association for Computational Linguistics.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Cartic Ramakrishnan, Abhishek Patnia, Eduard H Hovy, Gully APC Burns, et al. 2012. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1):7.

David Sánchez and Antonio Moreno. 2005. Web-scale taxonomy learning. In *Proceedings of Workshop on Extending and Learning Lexical Ontologies using Machine Learning (ICML 2005)*, pages 53–60, Bonn, Germany.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Asuka Sumida and Kentaro Torisawa. 2008. Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, volume 8, pages 883–888. Citeseer.

Yuka Tateisi, Tomoko Ohta, Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. 2000. Building an annotated corpus in the molecular-biology domain. In *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content*, pages 28–36. Association for Computational Linguistics.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse.

Ichiro Yamada, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, Masaki Murata, Stijn De Saeger, Francis Bond, and Asuka Sumida. 2009. Hypernym discovery based on distributional similarity and hierarchical structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 929–937. Association for Computational Linguistics.