A Supervised Machine Learning Approach for Taxonomic Relation Recognition through Non-linear Enumerative Structures

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

Bernard Rothenburger
Institut de Recherche en
Informatique de Toulouse
118, Route de Narbonne
31062 Toulouse, France
rothenburger@irit.fr

ABSTRACT

Improving relation extraction process requires to have a better insight of the proper text or to use external resources. Our work lies in the first term of this alternative, and aim at extending works about semantic relation identification in texts for building taxonomies which constitute the backbone of ontologies on which Semantic Web applications are built. We consider a specific discursive structure, the enumerative structure, as it bears explicit hierarchical knowledge. This structure is expressed with the help of lexical or typo-dispositional markers whose role is to introduce hierarchical levels between its components. Typo-dispositional markers are unfortunately not integrated into most parsing systems used for information extraction tasks. In order to extend the taxonomic relation identification process, we thus propose a method for recognizing this relation through enumerative structures which benefit from typo-dispositional markers (we called them non-linear enumerative structures). Our method is based on supervised machine learning. Two strategies have been applied: a linear classification with a MaxEnt and a non-linear one with a SVM. The results obtained in each of these approaches are close, with respectively an F1 of 81.25% and of 81.77%.

Categories and Subject Descriptors

I.2.7 [**Artificial intelligence**]: Natural language processing—*Information extraction*; I.2.6 [**Artificial intelligence**]: Learning—*Knowledge acquisition*

General Terms

Algorithms, Experimentation

SAC'15 April 13-17, 2015, Salamanca, Spain.
Copyright 2015 ACM 978-1-4503-3196-8/15/04…$15.00.
http://dx.doi.org/10.1145/2695664.2695988

Keywords

relation extraction, taxonomic relation, enumerative structure, text layout, supervised machine learning

1. INTRODUCTION

Improving relation extraction process requires to have a better insight of the proper text or to use external resources. Our work lies in the first term of this alternative, and aims at extending works about semantic relation identification in texts for building taxonomies which constitute the backbone of semantic resources on which Semantic Web applications are built. Indeed, the task of extracting taxonomic relation (also denoted as generic/specific, isA or instanceOf relations) is for example critical for ontology construction, enrichment or population. A lot of parameters may affect the type of methods used for this task, resulting in several proposed approaches and methods. Current implementations like lexico-syntactic patterns [4], clustering or machine learning algorithms (mostly unsupervised ones [3]), assume that related concepts are expressed in plain noun phrases and they only work when syntactic parsers produce relevant analysis. More recently, some works exploit the text hierarchical layout, without analyzing the plain text [8, 7, 6].

The contribution of this work is a fine-grained method to identify taxonomic relation from a structured discursive structure, the enumerative structure (ES), as it bears explicit hierarchical knowledge: a primer introduces and confers unity to a list of items. In many cases, the primer expresses an entity which is connected with a generic/specific relation to entities expressed in items. These discursive structures may be expressed: they may be expressed with the help of lexical or typo-dispositional markers whose role is to introduce hierarchical levels between its components, and they may also be expressed in the linearity of the text (a) or in a bi-dimensional space (b).

Non-linear ESs, i.e. those which benefit from typo-dispositional markers as in (b), are common in encyclopedic, technical or scientific corpora because they clarify the presentation of domain entities, alleviating the cognitive effort of the reader. However, analyzing these non-linear structures is not straightforward as the semantic part usually carried out by lexical markers is then supported by extralinguistic markers. So far these extralinguistic markers have not been integrated into most parsing systems used for information extraction tasks. We thus propose a method for recognizing the taxonomic relation borne by non-linear ESs. Our previous work relies on formal regularities [5], but it has been shown that some complex configurations and ambiguities limit the quality of such approaches. Thus, we propound hereafter a supervised machine learning approach for inferring taxonomic relations from non-linear ESs.

(a) Shoes are mainly composed of a sole, a vamp and a heel.

(b) Non-spoken forms of communication are:
- Written language
- Sign language
- Whistled language
- Non-verbal language

2. LEARNING TASKS

Identifying the taxonomic relation in ESs is a complex problem. Applying a non-linear method may conduct to an overfitted model, which does not generalize well on unseen data. In contrast, a linear model may have more misclassifications, but often generalizes better. We decided to apply these two strategies in order to compare them: a linear classification with a maximum entropy classifier (MaxEnt) [1] and a non-linear one with a Support Vector Machine (SVM) [2]. Both algorithms are supervised and make the assumption that the ESs present properties, modeled by a feature vector $\mathbf{x}$ with d dimensions, which can be learned from a training set. In this case, each feature often supplies an additional information for identifying the relation.

In our problem, the class of ESs which bore a taxonomic relation is represented as a discrete value y and the MaxEnt provides the conditional probability that an ES falls into this class given a parameter vector θ. On a log scale, the MaxEnt is linear in the feature space: the first term is the dot product for a given x and the second one is independent of the hold y (1). This equivalent to a logistic regression.

$$\log p(y = u|\mathbf{x}) = \theta_u^T\mathbf{x} - \log\sum_{v=1}^{c}\exp(\theta_v^T\mathbf{x}) \qquad (1)$$

The SVM provides a decision function which separates the dataset into two classes (2). For modeling non-linear dependencies between the features of ESs, we applied a radial basis function kernel, such as $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma||\mathbf{x} - \mathbf{x}'||^2)$. This provides the similarity between the $\mathbf{x}$ being classified and the support vector $\mathbf{x}'$. In the dual space, the α are the Lagrange multipliers and the primal-dual relationship allows a link with θ. The γ defines the sensibility of the kernel whereas the b reflects the bias term.

$$f(x) = sign(\sum_{i=1}^{n} y_i\alpha_i K(\mathbf{x}, \mathbf{x}^{(i)}) + b) \qquad (2)$$

Our experiments were run using the OpenNLP[1] library for the MaxEnt and the LIBSVM[2] implementation for the SVM. A *cut-off* of 100 was used for the MaxEnt and we applied a $\gamma = 1/d$ for the SVM.

3. EVALUATION

We manually built a corpus of 745 non-linear ESs extracted from 168 French Wikipedia pages. Indeed, Wikipedia is a goldmine of information since each page describes properties of a single entity. Furthermore, those properties are often expressed through non-linear enumerative structures.

In a guideline, we specified an annotation task in two steps: (i) for a given ES, the annotator had to characterize the relation as ontological or not, and (ii) if the ES bore an ontological relation, the annotator had to specify whether this relation was taxonomic or not. The distribution of this annotated corpus is reported in table 1.

Then, this corpus has been pre-processed for the learning tasks. Morphological and syntactic information have been added using the dependency parser Talismane [9]. 599 of the ESs were randomly chosen to constitute the development set (dev set), and the remaining 146 were used for the test set (test set). The whole corpus is available and can be used under the terms of the Creative Commons license[3].

Table 1: Distribution of the annotated corpus

Type		Obs.	Cover.
Ontological	Taxonomic	464	62.3%
	nonTaxonomic	81	10.9%
nonOntological		200	26.8%
Corpus		745	100%

We propose two binary classification tasks T_Taxo1 and T_Taxo2, for which we first conducted a feature selection through a Pearson's correlation. The two tasks classify an ES as taxonomic or not. The difference is that T_Taxo2 first implements an auxiliary binary classification task which classifies an ES as ontological or not, this first classification then being considered as an additional feature.

Except for this auxiliary classification, the set of features is the same for the two tasks. We distinguish two families of features: the first one is applied on the primer and the items of an ES, the second one concerns only the primer. The main features are summarized in the table 2.

Table 2: Feature set description

General Features	Description
POS	The presence of a part of speech in the primer or in the items
Start/End	The first or last part of speech in the primer or in the items
Plural	Boolean indicating the presence of a plural noun
Form	The number of tokens and the number of sentences
Primer's features	
Marker	Boolean indicating the presence of a relational marker
Syntactic	Boolean indicating if the last sentence is not syntactically complete, i.e. it ends with a subordinating conjunction, a preposition, a verb, etc.
Punctuation	Returns the last punctuation

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

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3] http://www.github.com/jfaucon/LARAt

Two evaluations have been carried out: an internal one on the dev set, and an external one on the test set. The internal evaluation was performed through a 10-fold cross-validation, and the external one was done with a holdout method. In the latter case, the model was trained on the entire dev set and evaluated on the test set. We propose a majority baseline, which better reflects the reality than a random baseline does, since the two classes are here unequally distributed.

The tables 3 and 4 present the results in terms of precision, recall, F1 and accuracy for the two tasks T_Taxo1 and T_Taxo2 on, respectively, the dev set and the test set. In both cases, the baseline shows a good precision thanks to the distribution of the corpus, making it difficult to beat.

Table 3: Internal evaluation on the dev set

Tasks	Strategies	Prec.	Rec.	F1	Acc.
T_Taxo1	MaxEnt	70.89	81.18	75.69	67.61
	SVM	72.53	88.71	79.81	72.12
T_Taxo2	MaxEnt	73.03	82.26	77.37	70.12
	SVM	73.57	89.78	**80.87**	**73.62**
Baseline	Majority	62.10	100.0	76.62	62.10

Table 4: External evaluation on the test set

Tasks	Strategies	Prec.	Rec.	F1	Acc.
T_Taxo1	MaxEnt	70.59	78.26	74.23	65.75
	SVM	71.05	88.04	78.64	69.86
T_Taxo2	MaxEnt	78.01	84.78	81.25	**75.34**
	SVM	74.77	90.22	**81.77**	74.66
Baseline	Majority	63.01	100.0	77.31	63.01

On the test set, the SVM from the T_Taxo2 task achieves an F1 of 81.77. This score decreases to a value of 78.64 for T_Taxo1, when the additional feature relative to the prior ontological classification is not considered. As shown in table 5, this feature led to significant improvements[4] in the prediction, though with a lesser gain for the SVM, which seems to learn easily without any additional information. This is expected since the SVM learns a non-linear hypothesis and has a greater flexibility to fit its training set. In contrast, the MaxEnt in T_Taxo2 reaches a better precision, since the feature space is well separated with the pre-classification. The gain of accuracy for this model reflects this property.

Table 5: Comparisons of the tasks on the test set

Comparisons			p-values
T_Taxo2 MaxEnt	vs.	T_Taxo1 MaxEnt	< 0.01
T_Taxo2 MaxEnt	vs.	Baseline	< 0.01
T_Taxo2 SVM	vs.	T_Taxo1 SVM	< 0.02
T_Taxo2 SVM	vs.	Baseline	< 0.01

As a conclusion, the obtained results on the test set resulting from our linear and non-linear classifiers show a good generalization. That suggests that the feature selection conducted on the dev set achieves a good bias–variance tradeoff, whatever the algorithm used.

As shown, the MaxEnt and the SVM led to close results but with some variations in terms of precision and recall. An investigation revealed that our classifiers seem to learn different representations of the problem. One way to improve our system would be to combine both models using ensemble approaches such as bagging or boosting.

4. PERSPECTIVES

First of all, in order to extract full-fledged taxonomic relations, we have to learn the concepts or the instances linked by these relations. Once this has been achieved, we have to confirm the results we have got by experimenting our system on other corpora. We may then plan to tackle larger scale applications of our work as Sumida *et al.* [8] do when they extract hyponymy relations from Japanese Wikipedia pages. We also plan to extend our machine learning approach in order to differentiate, within non-linear ES, the other ontological relations (including the meronymic relations and the non-hierarchical ones) from the lexical ones (including synonymy, antonymy, etc.). These more precise relations could then be involved both in terminology or ontology building processes.

5. REFERENCES

[1] A. Berger, V. Pietra, and S. Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[3] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, 24:305–339, 2005.

[4] M. A. Hearst. 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, 1992.

[5] M. Kamel, B. Rothenburger, and J.-P. Fauconnier. Identification de relations sémantiques portées par les structures énumératives paradigmatiques. *Revue d'Intelligence Artificielle*, Ingénierie des Connaissances, 2014.

[6] J.-H. Oh, K. Uchimoto, and K. Torisawa. 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, 2009.

[7] S. Ravi and M. Paşca. Using structured text for large-scale attribute extraction. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1183–1192. ACM, 2008.

[8] A. Sumida and K. Torisawa. Hacking wikipedia for hyponymy relation acquisition. In *IJCNLP*, volume 8, pages 883–888. Citeseer, 2008.

[9] A. Urieli. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* PhD thesis, Université de Toulouse, 2013.

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