A Supervised Learning for the Identification of Semantic Relations in Parallel Enumerative Structures

Jean-Philippe Fauconnier — **IRIT**
Co-Advisors : Nathalie Aussenac-Gilles and Mouna Kamel — **{firstname}.{lastname}@irit.fr**

1. Overview

Fields

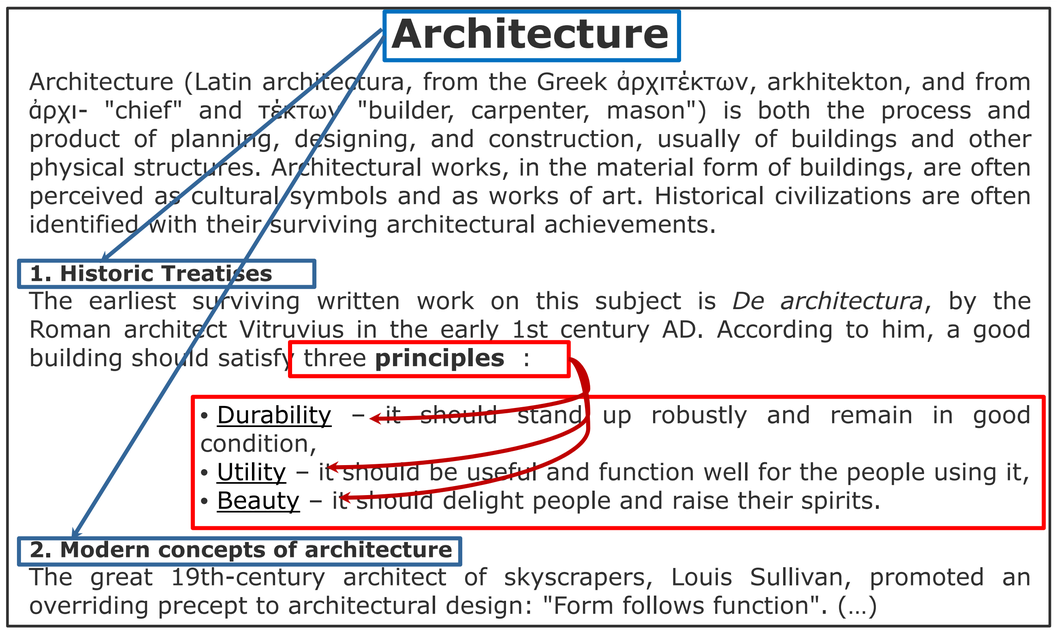
- Relation extraction
- Ontology learning

Purpose

Take into account text layout and structure to extract :

- binary relations
- n-ary relations

Method : use machine learning (classification) on syntactically tagged texts.



Example

(1) Hierarchical relation between the title and subtitles

(2) Hyperonymy relation between a concept and its 3 hyponyms

First experiment

We aim top extract relationships from enumerative structures,

textual objects that usually carry hierarchical relations

2. Parallel Enumerative Structures

Composition of a PES

Under IAU definitions, in the Solar System and in order of [increasing distance from the sun], there are *eight* **planets** :

- **Mercury** → **Item**
- **Venus**
- **Earth**
- **Mars**
- …

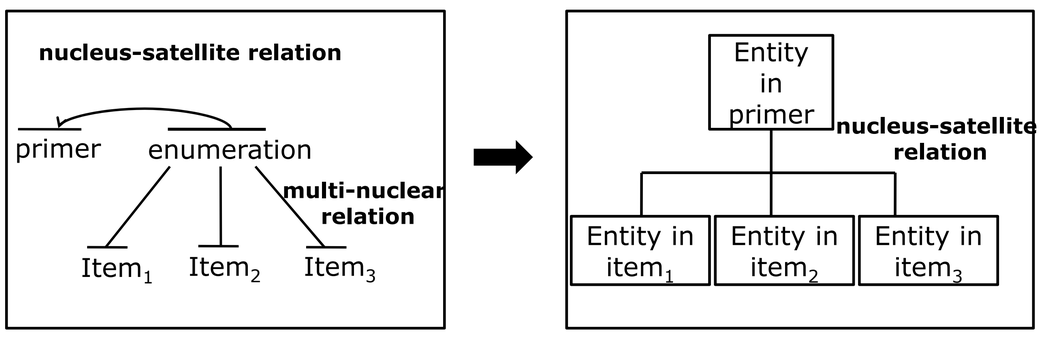
primer | **enumeration**

A parallel enumerative structure is :

- Homogeneous (items have the same layout),
- Paradigmatic (items are independent),
- Isolated (that do not require outside elements).

Semantic representation of a PES

The rhetorical structure can be translated into a hierarchical structure as follows :



3. Corpus and Annotation

1. Projection of *OntoTopo* concepts on Wikipedia
2. Extraction of parallel enumerative structures
3. Distribution of each primer on each of its items

Two sub-corpora : 1000 PES & 4317 pairs primer-item

Classes	Description	Kappa
IsA	Hyperonymy relation	0.509
PartOf	Meronymy relation	0.493
InstanceOf	Rel. between a concept and its instance	0.652
OtherOntological	Non-taxonomic relation (isCauseOf, …)	**0.299**
Lexical	Relation between terms (synonymy, …)	0.636
Other	Ambiguous cases, headings, …	0.641

Average : 0.56

4. Process

Pre-Processing

Analysis with Talismane, a statistical dependency parser

Two approaches

1. Use of linguistic and paralinguistic features
   e.g : presence of a verb, numeral, punctuation, infinitive in beginning of item, …
2. Trigrams of Lemma-POSTag
   e.g : The_DT dog_NN runs_VBZ

For classification, using of maximum entropy algorithm :

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} w_i f_i(x, y)\right)$$

5. Evaluation

Task	Approaches	Accuracy	σ	CI 95 %
Task 1	Ling. & paraling.	**61.10**	0.0154	[58.08 ;64.11]
	Trigrams	59.80	0.0155	[56.76 ;62.83]
	Baseline *Other*	38.00	0.0153	[35.00 ;40.99]
Task 2	Ling. & paraling.	58.70	0.0074	[57.25 ;60.15]
	Trigrams	**59.50**	0.0074	[58.04 ;60.95]
	Baseline *Other*	29.30	0.0069	[27.94 ;30.65]
Task 3	Ling. & paraling.	58.50	0.0155	[58.08 ;64.11]
	Trigrams	**59.00**	0.0155	[56.76 ;62.83]
	Baseline *Other*	38.00	0.0153	[35.00 ;40.99]

Three tasks of evaluation

1. Classifying each PES with its primer and its first item
2. Classifying each pair primer-item
3. Classifying each PES with weighted average of its pairs

Conclusion and future work

- Task 1 with linguistic and paralinguistic approach gives the best results for PES classification
- A hybrid approach combining machine learning and a rule-based system ?
- More paralinguistic features?

Université Paul Sabatier TOULOUSE III — IRIT CNRS - INPT - UPS - UT1 - UTM — MELODI