

Annotation of Semantic Relations in Patent Documents

Valentina Bartalesi Lenzi, Rachele Sprugnoli, and Emanuele Pianta



Outline

- Introduction
 - Annotation process
 - Annotation of Part-Whole Relations
 - Annotation of Motion Relations
 - An experiment on the Inter-Coder Agreement
 - Conclusions and future work
-

Introduction

- Manual annotation **Part–Whole** and **Motion** relations in patent documents
 - **PATExpert** was an European project whose goal was to change the traditional approach to patent processing from textual to semantic
 - The aim of our activity in the PATExpert project was to create a gold standard for the evaluation of an automatic relation extraction tool developed by FBK–irst.
 - Part–Whole and Motion relations are numerically relevant in patent documents
 - Possible queries to the relation extraction tool:
 - “Retrieve all patent sentences mentioning an optical head that includes a Wollaston prism”*
 - “Retrieve all patent sentences mentioning the movement of a cam member from the first to the second position”*
-

Annotation process (1)

- Two corpora made up of 16 patents each, 8 about optical devices and 8 regarding machine tools:
 - i. Part-Whole relations in a corpus of about 50,000 words;
 - ii. Motion relations in a corpus of about 100,000 words.

- Development of specific annotation guidelines

- Two annotators, 4.5 person/months.

Annotators, after a training phase, interacted and negotiated common solutions to controversial annotations

- Callisto was used as annotation tool:
 - input files were raw texts in UTF-8 encoding
 - output data were XML files in the ACE Program Format (APF)
-

Annotation process (2)

- We adopted the annotation scheme created for the Relation Detection and Recognition task of the Automatic Content Extraction (ACE) Program, adapting attribute values to our specific needs.

Attributes:

- *Extent*, the smallest portion of text in which the relation is expressed
 - *Arguments*, the two entities involved in the relation
 - *Semantic Type*, used to distinguish between Part-Whole and Motion relations
 - *Semantic Subtype*, used to characterize the semantic roles of motion relations
 - *LexicalCondition*, the syntactic class of the lexical item expressing the relation
 - *Modality*, indicates if the relation is asserted or not
-

Annotation of Part–Whole Relations (1)

- A **Part–Whole** relation is the relation between a segment or a portion in which an entity is divided and the entity itself.
- Winston et al. (1987) taxonomy was our starting point.
- Given the characteristics of the patent documents in our corpus, we choose to annotate only three of the six classes* defined by Winston:
 - **Component–Object**, the relation between a component and the integral object to which it belongs
e.g. “the rotary holder has a shaft”;
 - **Portion–Mass**, the whole is considered as a homogeneous mass and its portions are similar to each other and to the whole
e.g. “excessive portion of the adhesive”;
 - **Place–Area**, spatial relation among regions in a geometrical sense
e.g. “a surface of an optical disc”.

* The other three classes are *Member–Collection*, *Stuff–Object*, *Feature–Activity*

Annotation of Part–Whole Relations (2)

- Special effort to make a distinction between Part–Whole relations and other relations which can be expressed in a similar way.

Relations NOT to be annotated:

- *Topological Inclusion*: relation between a content and its container
e.g. “there is a foam in the adhesive layer”;
 - *Attribution*: relation between an object and its attribute
e.g. “the thickness of the Wollaston prism”;
 - *Attachment*: two components are attached to each other
e.g. “the front panel attached to the casing”;
 - *Ownership*: relation between a person or an institution and something that they own
e.g. “a woodworker have several boards”.
-

The *LexicalCondition* attribute

- Indicates the syntactic class of the element that provides justification for the tagging of each relation.
 - Five values:
 - 1) *Verbal*, e.g. “tool clamping device comprising a draw bar”
 - 2) *Preposition*, e.g. “optical recording layers of an optical disk”
 - 3) *Possessive*, e.g. “optical disks bonded at their faces”
 - 4) *Adverbial*, e.g. “the optical disc is read from one side thereof”
 - 5) *PreModifier*, e.g. “the substrate side” → noun modifiers that precede the head noun and can be the part or the whole of that head
-

Part-Whole Relations: Data (1)

- Annotated relations: 1,015
- Occurrences of the *Modality* attribute

Modality	Count	Percentage
Asserted	963	94.88%
Possibility	46	4.53%
Negated	6	0.59%

→ “a recording film **having** grooves”

→ “the DVD-RAM **may have** a rom area”

→ “the rom area **has no** reflective layer”

- Occurrences of the *LexicalCondition* attribute

LexicalCondition	Count	Percentage
Verbal	499	49.16%
Preposition	440	43.35%
PreModifier	37	3.65%
Adverbial	20	1.97%
Possessive	19	1.87%

Part-Whole Relations: Data (2)

- Verbs determining Part-Whole relations

Verbs	Count
have	116
provide	94
include	85
comprise	74
form	72
locate	13
coexist	11
characterize	10
make, compose, incorporate, place	< 10

- Prepositions in Part-Whole relations

Preposition	Percentage	Examples
of	80.34%	a cam face <i>of</i> a spring disk
in	9.57%	a recording pit <i>in</i> the data area
on	5.42%	a tab <i>on</i> the slide block
with	3.03%	a bore <i>with</i> a stop shoulder
within, at, into	< 1.00%	the region <i>within</i> the data area the pin is <i>at</i> the lower end the pipes <i>into</i> one unit

Comparison with Girju classification

▪ Girju et al. (2006) groups Part-Whole lexico-syntactic patterns in four clusters on the basis of their semantic similarity:

- 1) Genitives and the verb “to have”
- 2) Noun compounds
- 3) Prepositions
- 4) Other (verbs different from “to have” and expressions like “X is member of Y” and “X is a branch of Y”)

Girju patterns	Percentage	LexicalCondition	Percentage
Noun Compounds	16.07%	PreModifier	3.65%
Other	6.36%	Verbal	49.16%

ignoring “to have” the occurrences are more than 38%

Annotation of Motion Relations

- Motion relations are not binary:

- more than two arguments

- the optical head is moved in a radial direction by a moving means*

- just one argument

- the optical head is raised*

- We used “Callisto relations” to annotate the link between a lexical item expressing a motion and one of its arguments.

- Talmy’s theory (1985, 1991) on the concept of motion and the analysis of motion relations developed within the FrameNet project were taken as a reference point for our work.

- We used the *Semantic Subtype* attribute to indicate the semantic role of each argument involved in the relation.

- The semantic roles adopted in our annotation belong to the *Motion* and *Cause_motion* frames in FrameNet.
-

Motion Relations: Semantic Roles

List of semantic roles (the semantic role filler is underlined):

- Theme: e.g. “the optical pickup *moves* in the radial direction”;
 - Source: e.g. “an optical head *moves* from the current position”;
 - Goal: e.g. “an optical head *moves* to a target track”;
 - Path: e.g. “the collar *moved* along the drill”;
 - Direction: e.g. “the pick-up *moves* in that direction”;
 - Distance: e.g. “the optical pickup *moves* to a certain extent”;
 - Manner: e.g. “the optical head *moves* at a low speed”;
 - Cause: e.g. “the object *moved* by these motors”.
-

How Motion is Expressed

The *LexicalCondition* attribute indicates how the motion is syntactically expressed:

- 1) **Verbal-Direct**: verbs accompanied by a direct argument
e.g. “the puller *moves* to a target position”;
 - 2) **Verbal-Prep**: verbs with prepositional phrases
e.g. “the operative gear *moves* along the guides”;
 - 3) **Nominal-Prep**: nominal heads followed by prepositional phrase
e.g. “the *movement* of the guide”;
 - 4) **Nominal-Poss**: nominal heads modified by possessives
e.g. “its *movement*”;
 - 5) **Nominal-PreMod**: pre-modifiers followed by nominal heads
e.g. “the guide *movement*”;
 - 6) **Nominal-Head**: the head of a nominal phrases expresses an attribute of motion e.g. “*the movement* distance”;
 - 7) **Adverbial**: adverbs encoding manner of motion
e.g. “the guide quickly *moves*”.
-

Motion Verbs (1)

- To identify motion verbs we took into consideration the lexical units reported for the following frames of FrameNet:
 - Motion, “the puller is *moved* in the cross direction”
 - Cause_motion, “the lens actuator is *driven* by the control unit”
 - Arriving, “the cam member *reaches* the second position”
 - Departing, “the reading head *departing* from the first position”
 - Traversing, “the guide rod *passes through* a guide hole”
 - Motion_directional, “the front guide pins *rise up* along the grooves”
-

Motion Verbs (2)

- A motion interpretation can be assigned also to some verbs not included in the mentioned frames.

E.g. “the drill guide assembly is *raised* to a higher position”

“the cam member is *displaced* from the first position to the second position”

- Although in English verbs can conflate a manner component, we annotated only linguistic elements that explicitly express manner of motion.

E.g. “The cam member slides slowly along the restricting groove”.

Semantic Roles	Linguistic elements
Theme	The cam member
Manner	slowly
Path	the restricting groove

The Role of Prepositions

- Spatial prepositions turned out to play a crucial role in identifying semantic roles fillers.
- The literature distinguishes between **directional** prepositions, like *into* and *through*, and **locative** prepositions, like *in* and *under*.
- Annotation problem: correctly interpreting directional prepositions which can encode both goal and path roles.

→ Zwarts (2008) classification of directional prepositions

Prepositions	Examples
Source: e.g. <i>from</i> , <i>out of</i>	– the optical pick-up is moved <i>from</i> the inner side
Goal: e.g. <i>into</i> , <i>to</i> , <i>onto</i>	– the drill bit slides <i>into</i> the drilling guide bushings
Route: e.g. <i>via</i> , <i>through</i> , <i>across</i>	– guide pins which slide <i>across</i> slide bushings
Comparative: e.g. <i>towards</i>	– the slider moves <i>toward</i> the turn table
Constant: e.g. <i>along</i>	– the cam follower slides <i>along</i> the support shaft
Periodic: e.g. <i>up and down</i>	– the drill guide assembly is moved <i>up and down</i>

Motion Relation: Data (1)

- Annotated relations: 624
- Annotated semantic role fillers: 1,324

Semantic Roles	Count	Percentage
Theme	590	44.56%
Goal	166	12.54%
Direction	159	12.01%
Cause	134	10.12%
Path	97	7.32%
Manner	81	6.12%
Source	63	4.76%
Distance	34	2.57%

No Theme → “the movement caused by the driving mechanism”

- Occurrences of the *Modality* attribute

Modality	Count	Percentage
Asserted	1213	91.61%
Possibility	99	7.48%
Negated	12	0.91%

Motion Relation: Data (2)

- Occurrences of the *LexicalCondition* attribute

LexicalCondition	Count
Verbal-Direct	631 (47.66%)
Verbal-Prep	452 (34.14%)
Nominal-Prep	118 (8.91%)
Adverbial	69 (5.21%)
Nominal-Head	36 (2.72%)
Nominal-Poss	13 (0.98%)
Nominal-PreMod	5 (0.38%)

- Occurrences of motion verbs

Verbs	Count
move	388
drive	59
position	37
reach	35
push	25
guide	22
displace	14
slide	11
pass, lift, carry, ride, depart, go, overrun, place, pull, raise, reposition, run, transfer, rise, transport	< 10

Motion: comparison with other corpora

- To our knowledge, there is not another corpus in which all motion relations have been systematically annotated with semantic roles
 - Potential comparison with SemCor
 - Pre-condition: mapping WordNet synset annotation into FrameNet annotation
(see ongoing work at FBK on WordNet to FrameNet mapping)
-

Inter-Coder Agreement

- Work in progress:
 - one patent about machine tools
 - about 1,400 words
 - two expert annotators
 - Part-Whole Relation annotation
 - Relation Detection: 0.993 (Dice Coefficient)
 - *LexicalCondition* assignment: 0.983 (k statistic)
 - Motion Relation annotation
 - Relation Detection: 1 (Dice Coefficient)
 - *LexicalCondition* assignment: 0.939 (k statistic)
 - Semantic Role assignment: 0.940 (k statistic)
-

Conclusions and Future Work

- **Difficulties:**
 - text understanding;
 - identification of verbs encoding Part–Whole relations
 - choice of the correct Semantic Role for each argument of Motion relations
 - **Future work:**
 - to complete the experiment on the inter–coder agreement
 - to extent the linguistic analysis of motion verbs
 - to exploit results for lexical typological analysis?
(e.g. Verb–Framed versus Satellite–Framed languages)
-

A decorative vertical bar on the left side of the slide, consisting of a light gray top section and a dark gray bottom section. Two horizontal red lines extend from the right edge of the bar across the slide.

Thanks for your attention