Towards deep content extraction from specialized discourse: The case of verbal relations in patent claims

Gabriela Ferraro

Department of Communication and Information Technologies
Universitat Pompeu Fabra

November 2012

Outline

Outline

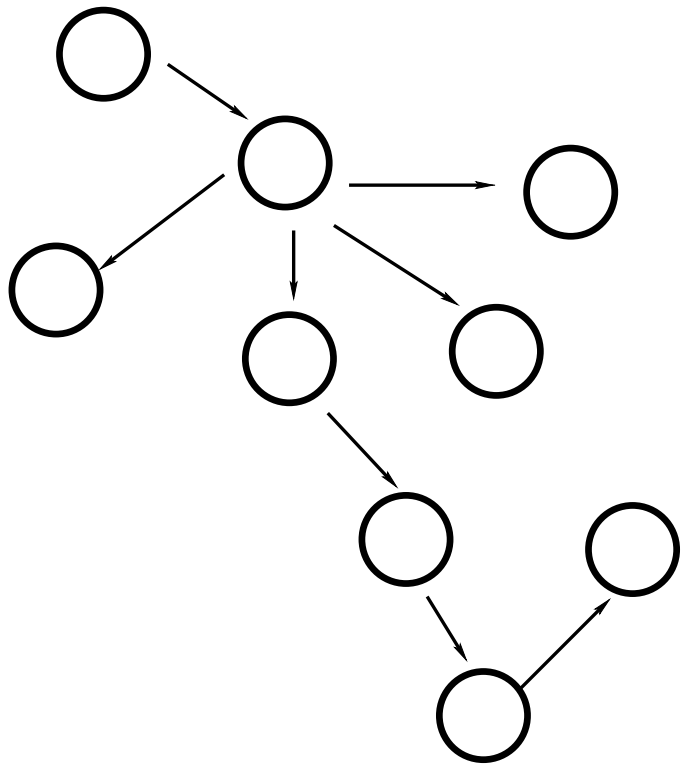
Problem definition

An automatic focusing device comprising: an objective lens for focusing a light beam emitted by a light source on a track of an information recording medium; a beam splitter for separating a reflected light beam reflected by the information recording medium at a focal spot thereon and through the objective lens from the light beam emitted by the light source; an astigmatic optical system including an optical element capable of causing the astigmatic aberration of the separated reflected light beam; a light detector having a light receiving surface divided, except the central portion thereof, into a plurality of light receiving sections which are arranged symmetrically with respect to a first axis extending in parallel to the axial direction of the optical element and to a second axis extending perpendicularly to the first axis and adapted to receive the reflected beam transmitted through the optical element and to give a light reception output signal corresponding to the shape of the spot of the reflected light beam formed on the light receiving surface; a focal position detecting circuit capable of giving an output signal corresponding to the displacement of the objective lens from the focused position, on the basis of the output signal given by the light detector; and a lens driving circuit which drives the objective lens along the optical axis on the basis of the output signal given by the focal position [...]

Problem definition

An automatic focusing device **comprising**: an objective lens for focusing a light beam emitted by a light source on a track of an information recording medium; a beam splitter for separating a reflected light beam reflected by the information recording medium at a focal spot thereon and through the objective lens from the light beam **emitted** by the light source; an astigmatic optical system including an optical element capable of causing the astigmatic aberration of the separated reflected light beam; a light detector having a light receiving surface divided, except the central portion thereof, into a plurality of light receiving sections which are arranged symmetrically with respect to a first axis extending in parallel to the axial direction of the optical element and to a second axis extending perpendicularly to the first axis and adapted to receive the reflected beam transmitted through the optical element and to give a light reception output signal corresponding to the shape of the spot of the reflected light beam formed on the light receiving surface; a focal position detecting circuit capable of giving an output signal corresponding to the displacement of the objective lens from the focused position, on the basis of the output signal given by the light detector; and a lens driving circuit which drives the objective lens along the optical axis on the basis of the output signal given by the focal position [...]

Proposal



Why patent claims?

- Patents contain up-to-date scientific and technical information that is usually not published elsewhere

- Patent claims form the central section of a patent document as they define the boundaries of legal protection

- Patent claim genre is a challenge to NLP

Why patent claims?

- Patents contain up-to-date scientific and technical information that is usually not published elsewhere

- Patent claims form the central section of a patent document as they define the boundaries of legal protection

- Patent claim genre is a challenge to NLP

Why patent claims?

- Patents contain up-to-date scientific and technical information that is usually not published elsewhere
- Patent claims form the central section of a patent document as they define the boundaries of legal protection
- Patent claim genre is a challenge to NLP

Why patent claims?

- Patents contain up-to-date scientific and technical information that is usually not published elsewhere
- Patent claims form the central section of a patent document as they define the boundaries of legal protection
- Patent claim genre is a challenge to NLP

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style

- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing

- Provide a framework for the generalization of verbal relations

  via clustering and cluster labeling techniques

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style

- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing

- Provide a framework for the generalization of verbal relations

  via clustering and cluster labeling techniques

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style

- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing

- Provide a framework for the generalization of verbal relations

  via clustering and cluster labeling techniques

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style

- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing

- Provide a framework for the generalization of verbal relations

  via clustering and cluster labeling techniques

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style

- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing

- Provide a framework for the generalization of verbal relations

  via clustering and cluster labeling techniques

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style
- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing
- Provide a framework for the generalization of verbal relations

  via clustering and cluster labeling techniques

Objectives

- Provide a detailed description of the patent claim genre

  via analysis of the linguistic idiosyncrasies of the claim style
- Provide a framework for flexible, unlexicalized and n-ary verbal relation extraction

  via deep dependency parsing
- Provide a framework for the generalization of verbal relations

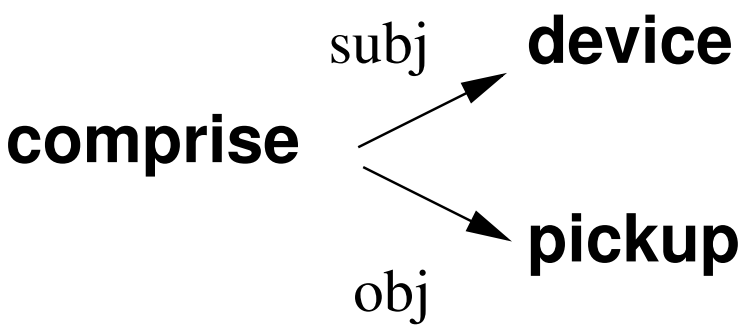
  via clustering and cluster labeling techniques

Outline

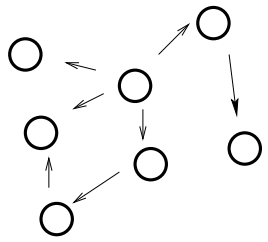
Linguistic dependency

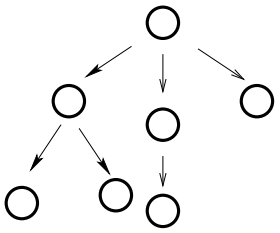
Dependency relations are:

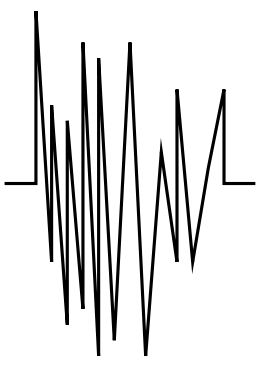
- binary directed asymmetrical relations that take two arguments
- one argument is the *head*
- the other argument is the *dependent* of the *head*



The Meaning-Text Theory







Semantic Representation

Deep Syntactic Representation

Surface Syntactic Representation

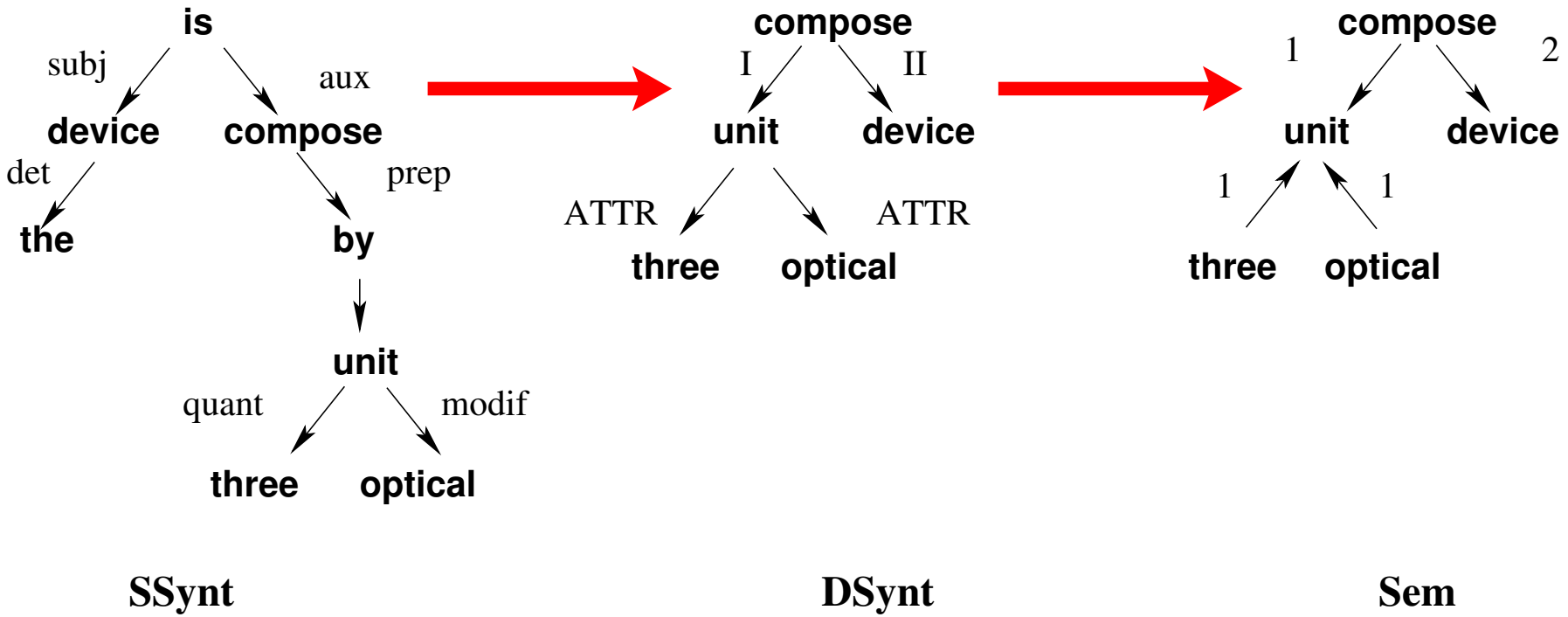
Deep Morphological Representation

Surface Morphological Representation

Deep Phonetic Representation

Surface Phonetic Representation

MTT: from syntax to semantics



The device is composed by three optical units.

Surface syntax vs. Deep syntax

Surface syntax:

- dependencies capture the grammatical relations of lexical units
- comprises around 50 language-dependent relations
- all the lexemes of the sentence are nodes of the tree

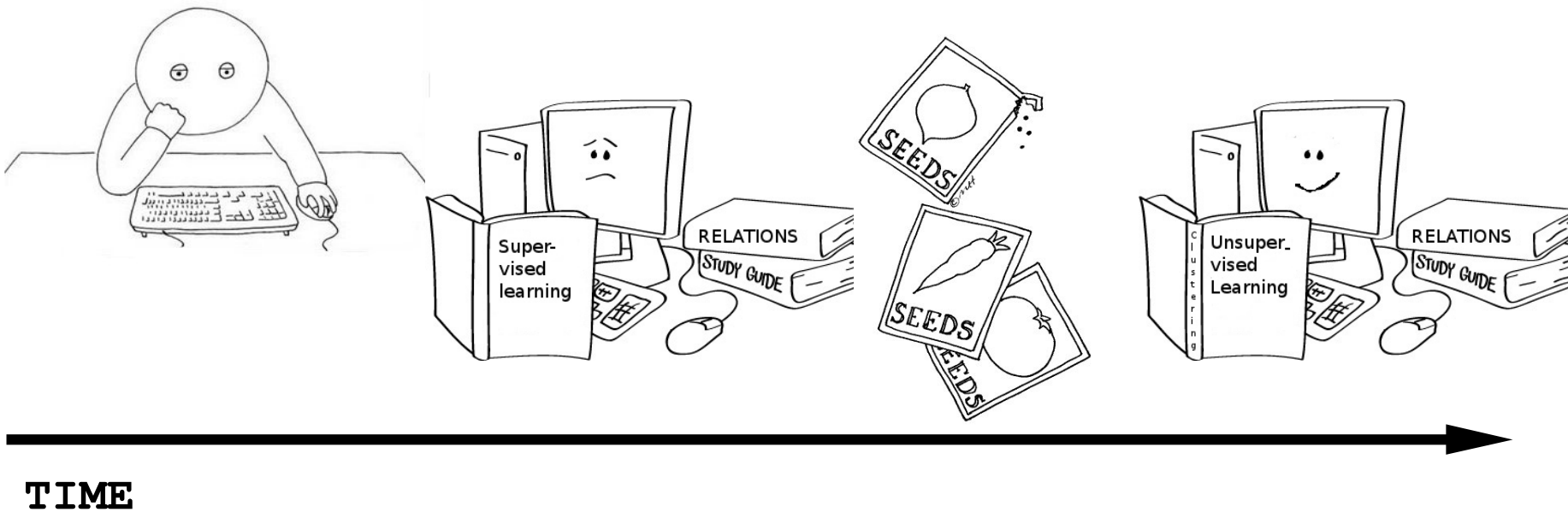
Deep syntax:

- dependencies capture the predicate argument structure
- comprises only 9 language-independent relations
- only semantically full lexemes appear in the tree

Outline

Relation extraction

- lexico-syntactic patterns
- purely syntactic patterns
- word co-occurrences



Relation clustering

Relation clustering deals with grouping relations according to some similarity criteria

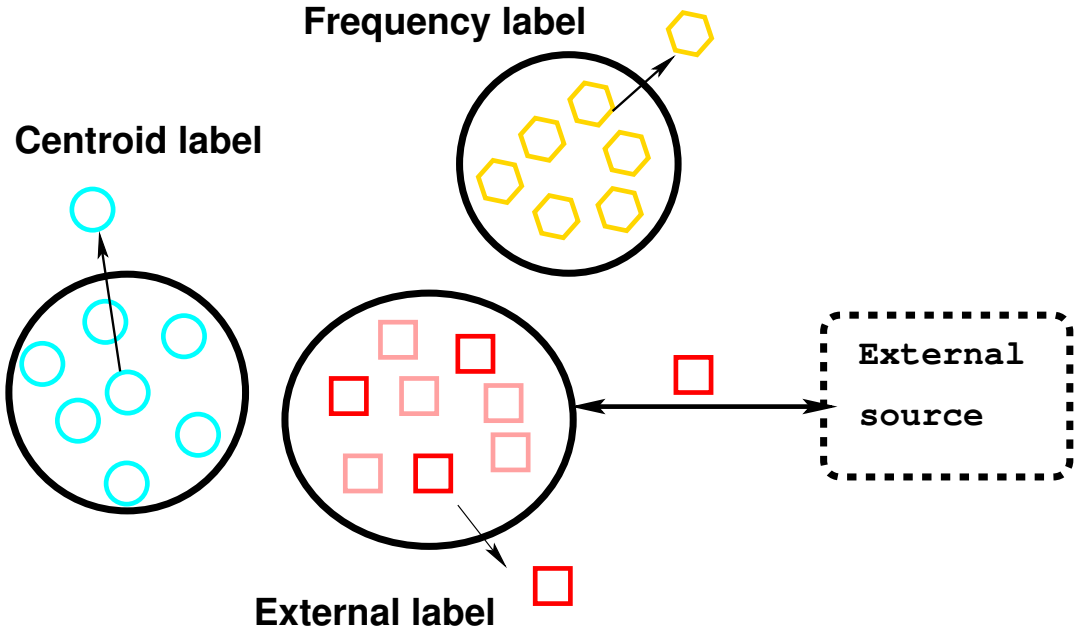
- Over a taxonomy or set of predefined relations (*verbs of putting*, *verbs of removing*, *verbs of motion*, etc.)

- Using similarity measures (MI, LRA, etc.)

- Using the WordNet hierarchy

Cluster labeling

- Internal cluster labeling
- Differential cluster labeling

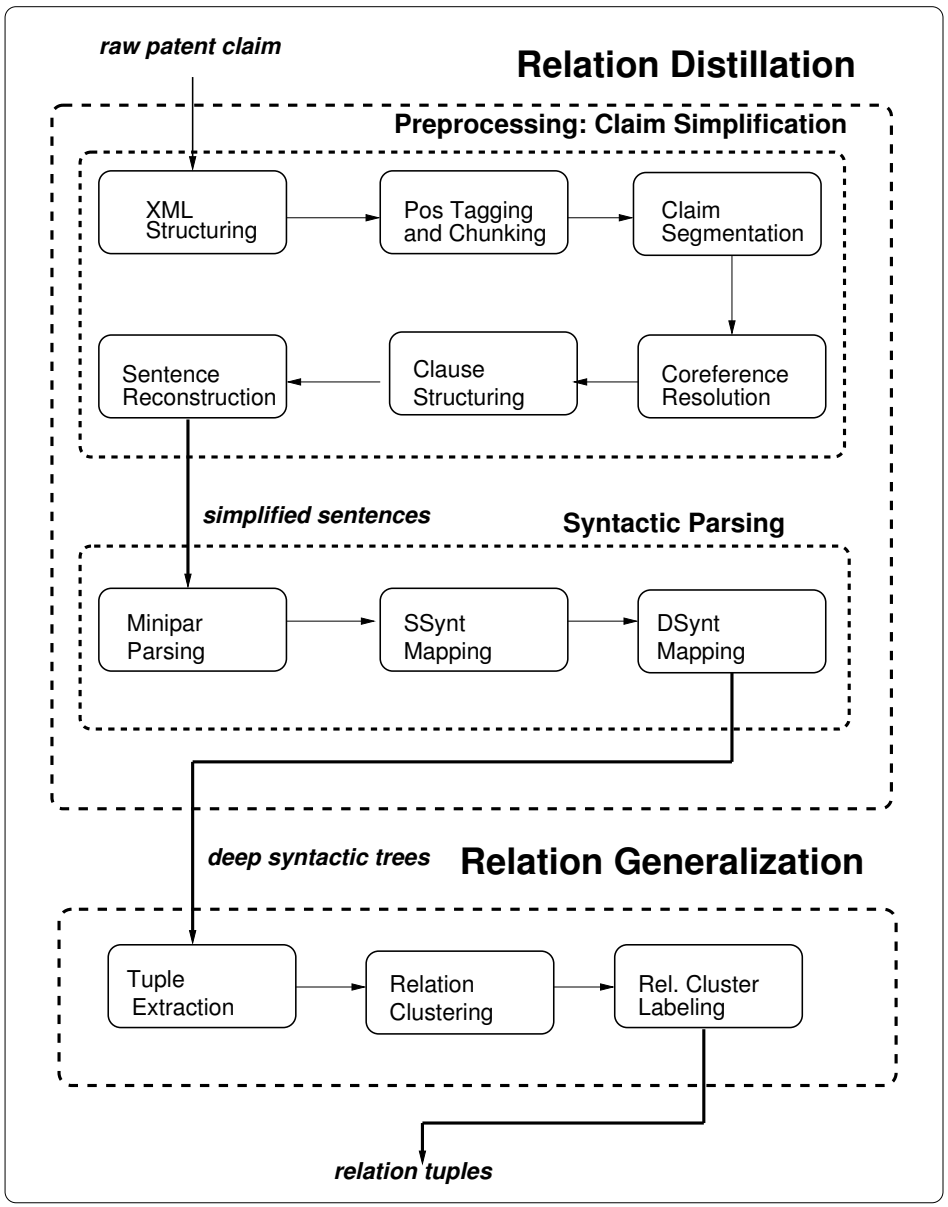


A step further in RE

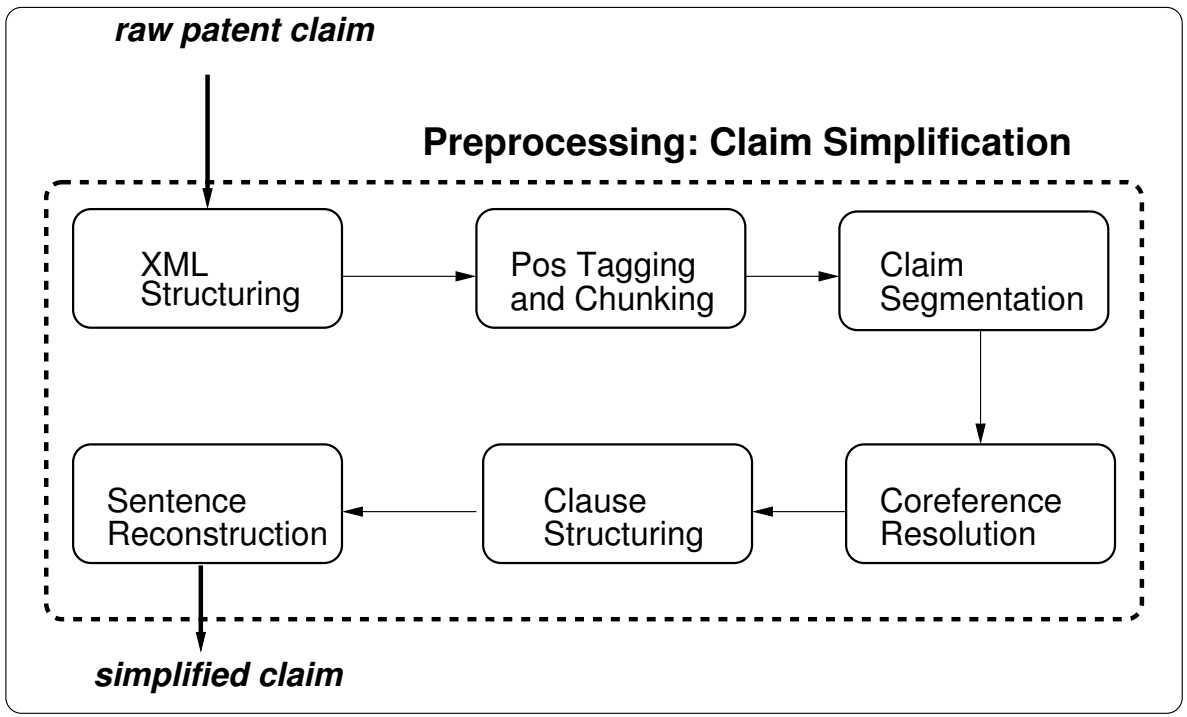
- **Relation Extraction**
  - open & n-ary
  - use deep-syntactic trees to identify the arguments of relations
- **Relation Classification**
  - unsupervised methods with simple features

Outline

Architecture



Preprocessing: claim simplification



- Linguistic preprocessing: TreeTagger
- Clause segmentation: Machine learning (ML) approach
- Coreference resolution between NPs
- Clause tree building: ML approach
- Clause reconstruction: rule-based approach

Simplification input example

An automatic focusing device comprising: an objective lens for focusing a light beam emitted by a light source on a track of an information recording medium; a beam splitter for separating a reflected light beam reflected by the information recording medium at a focal spot thereon and through the objective lens from the light beam emitted by the light source; an astigmatic optical system including an optical element capable of causing the astigmatic aberration of the separated reflected light beam; a light detector having a light receiving surface divided, except the central portion thereof, into a plurality of light receiving sections which are arranged symmetrically with respect to a first axis extending in parallel to the axial direction of the optical element and to a second axis extending perpendicularly to the first axis and adapted to receive the reflected beam transmitted through the optical element and to give a light reception output signal corresponding to the shape of the spot of the reflected light beam formed on the light receiving surface; a focal position detecting circuit capable of giving an output signal corresponding to the displacement of the objective lens from the focused position, on the basis of the output signal given by the light detector; and a lens driving circuit which drives the objective lens along the optical axis on the basis of the output signal given by the focal position [...]

Simplification output example

An automatic focusing device comprises: an objective lens; a beam splitter; an astigmatic optical system; a light detector; a focal position detecting circuit capable of giving an output signal and a lens driving circuit.

The objective lens focusses light beam.

The light source emits a light beam on a track of an information recording medium.

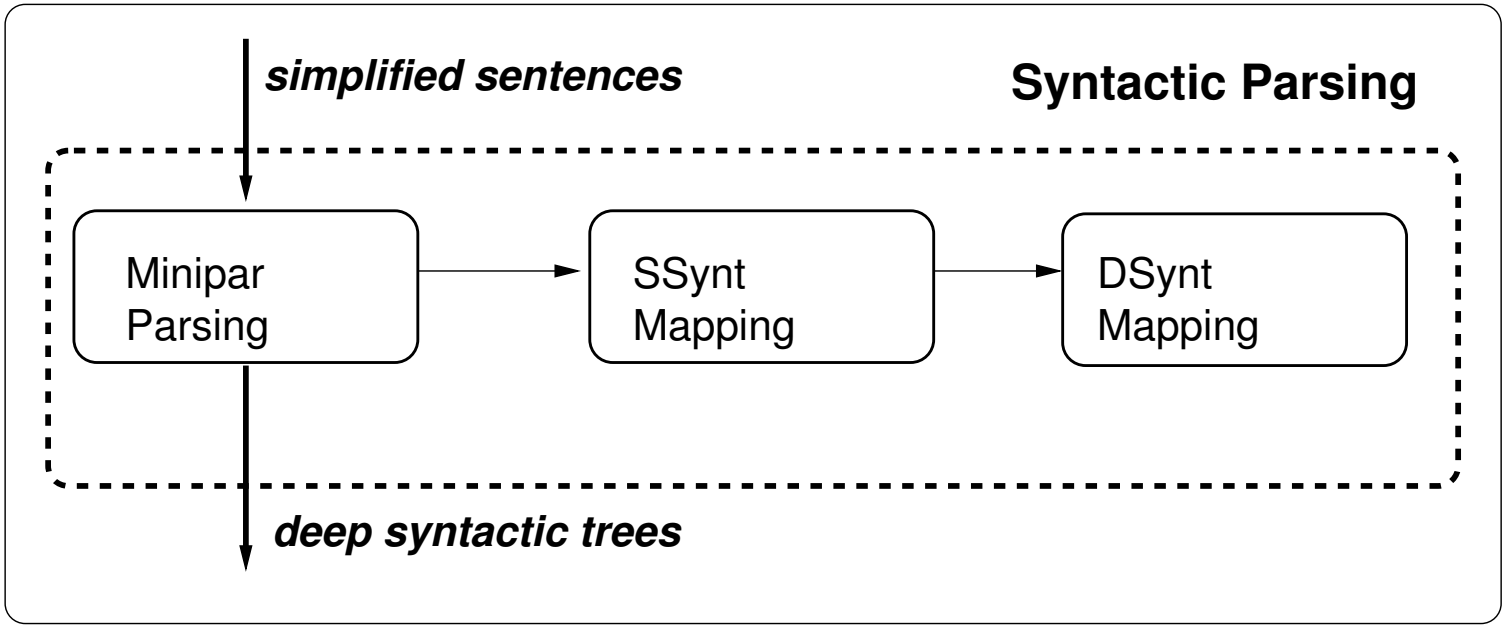
The beam splitter separates the reflected light beam.

The information recording medium reflects the reflected light beam at a focal spot thereon and through the objective lens from the light beam.

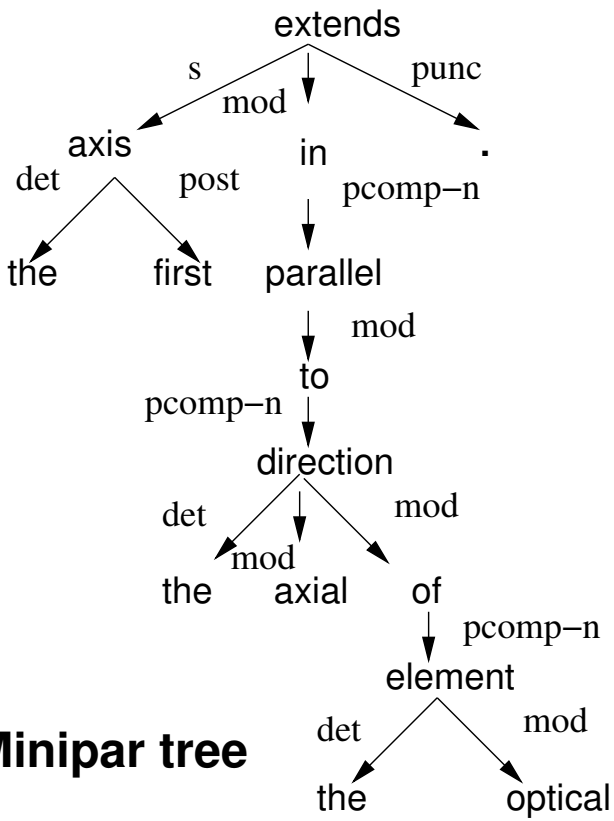
The light source emits the light beam.

The astigmatic optical system includes an optical element.

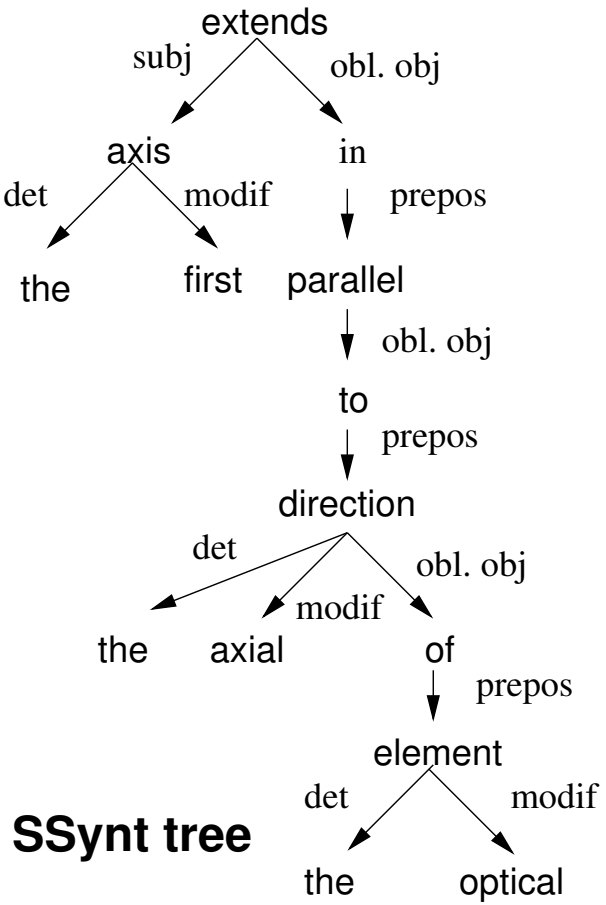
Syntactic parsing



Syntactic parsing: Minipar to MTT Surface Syntax

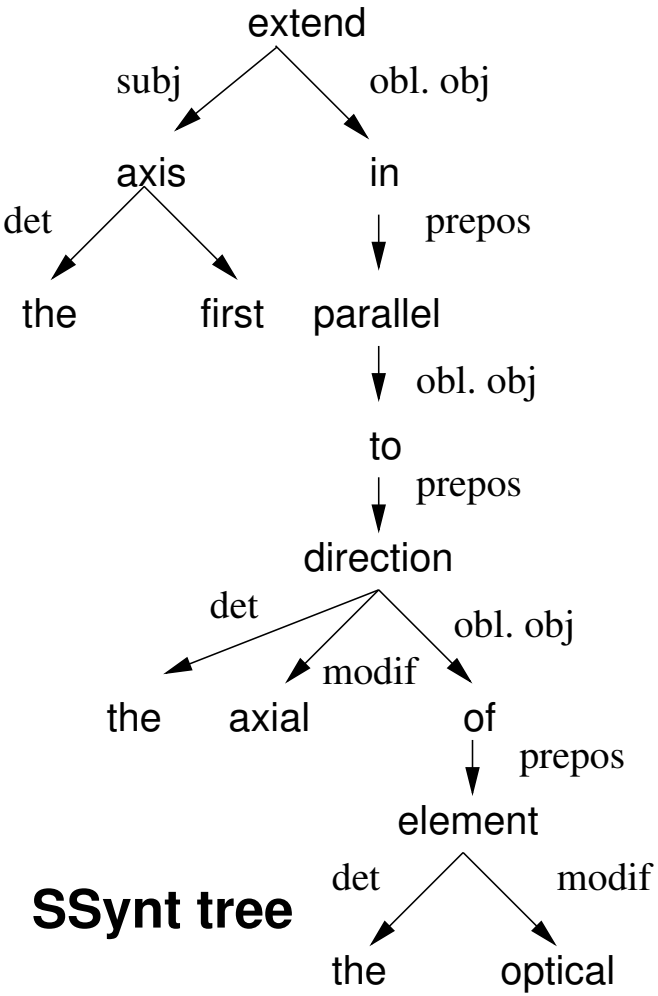


Minipar tree

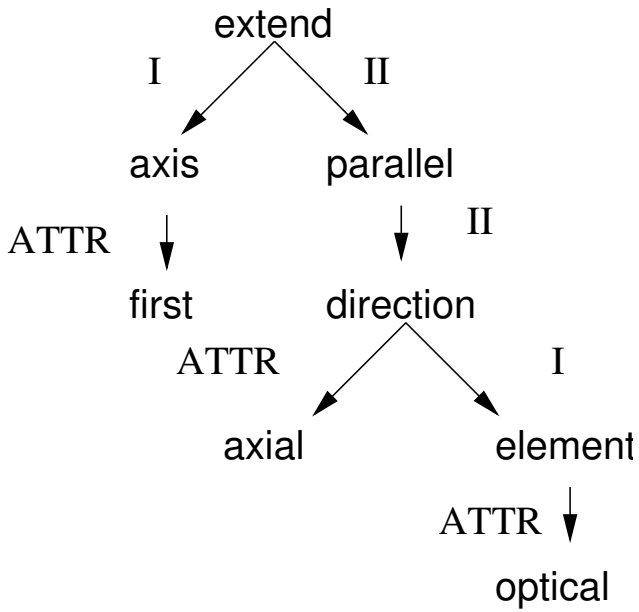


SSynt tree

Syntactic parsing: Surface Syntax to Deep Syntax

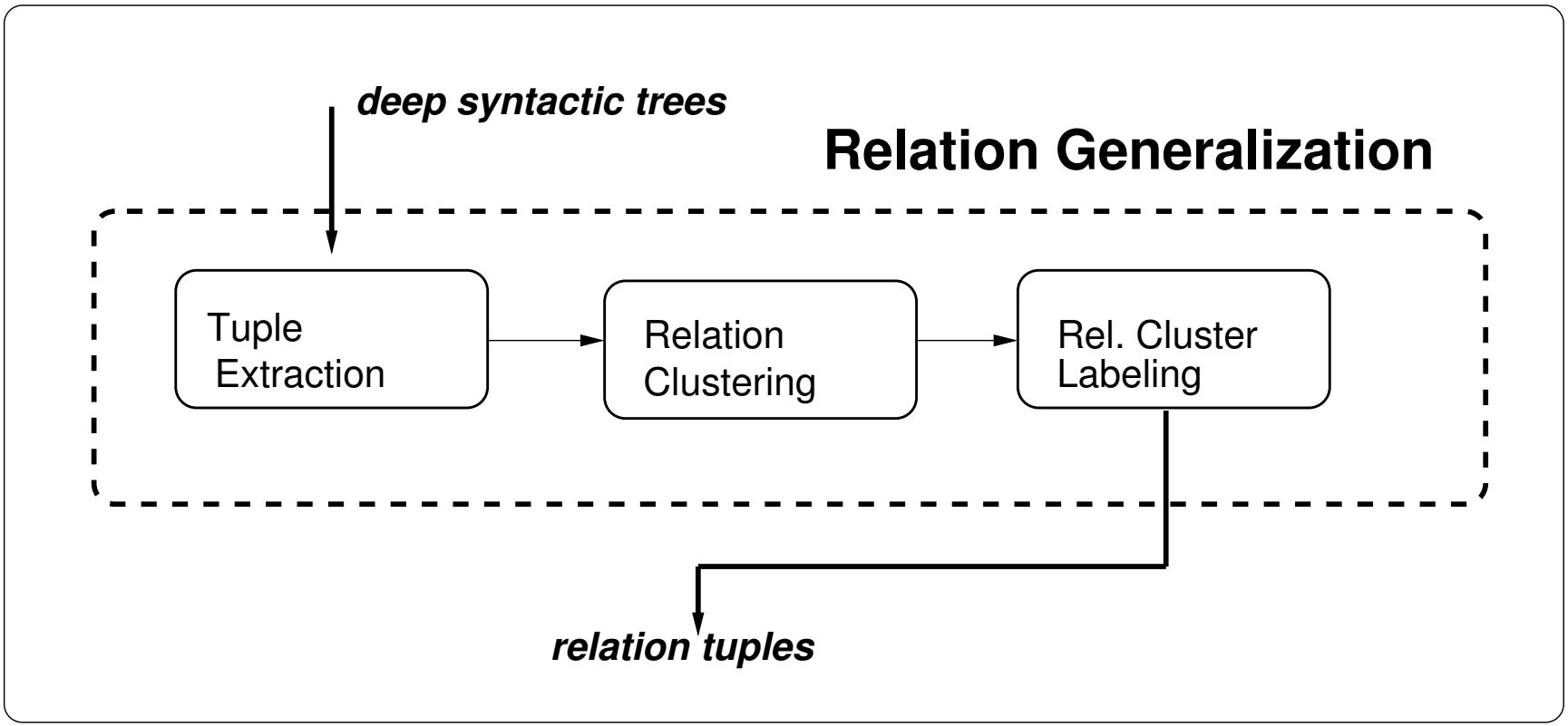


SSynt tree



DSynt tree

Relation generalization

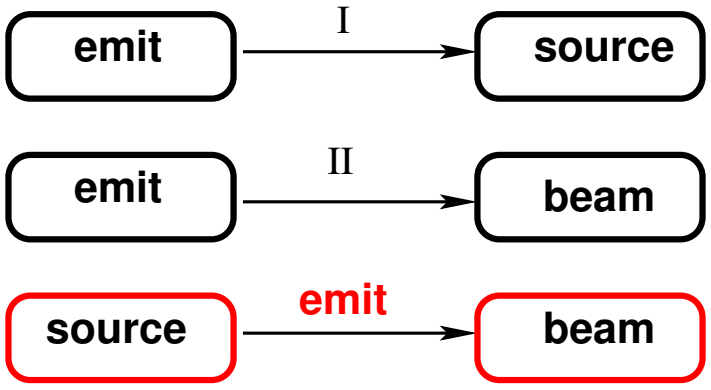


Obtaining relation tuples

- argument tags are eliminated
- actantial relations are merged
- attributive relations are merged

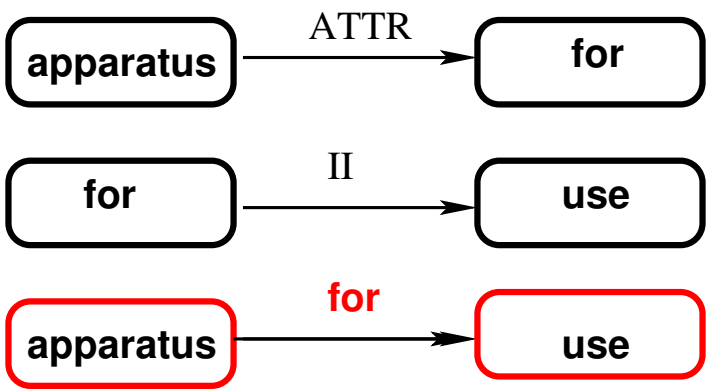
Actantial tuples merging rule:

IF (I(V_i, A_1) & II(V_i, A_2))
SET (V_i(A_1, A_2))
FI



Attributive tuples merging rule:

IF (ATTR(A_1, A_2) & II(A_2, A_3))
SET (A_2(A_1, A_3))
FI



Relation clustering

- Each verbal relation is represented as a vector of its WordNet-synonyms, e.g.:

construct1(construct1, build1, make17)
make24(make24, build1, construct1)

Step 1: calculate the similarity between synonym vectors

$$\cos(\vec{v_i}, \vec{w_j}) = \frac{\vec{v_i} \bullet \vec{w_j}}{|\vec{v_i}| \bullet |\vec{w_j}|} = \frac{\sum_{s=1}^{N} v_{i,s} w_{j,s}}{\sqrt{\sum_{s=1}^{N} v_{i,s}^2} \sqrt{\sum_{s=1}^{N} w_{j,s}^2}}$$

Step 2: cluster the vectors according to the obtained similarities

- Best clustering algo.: *optimized repeated bisections*; $K = 60$

Relation clustering

- Each verbal relation is represented as a vector of its WordNet-synonyms, e.g.:

construct1(construct1, build1, make17)
make24(make24, build1, construct1)

Step 1: calculate the similarity between synonym vectors

$$\cos(\vec{v_i}, \vec{w_j}) = \frac{\vec{v_i} \bullet \vec{w_j}}{|\vec{v_i}| \bullet |\vec{w_j}|} = \frac{\sum_{s=1}^{N} v_{i,s} w_{j,s}}{\sqrt{\sum_{s=1}^{N} v_{i,s}^2} \sqrt{\sum_{s=1}^{N} w_{j,s}^2}}$$

Step 2: cluster the vectors according to the obtained similarities

- Best clustering algo.: *optimized repeated bisections*; $K = 60$

Relation clustering

- Each verbal relation is represented as a vector of its WordNet-synonyms, e.g.:

construct1(construct1, build1, make17)
make24(make24, build1, construct1)

Step 1: calculate the similarity between synonym vectors

$$\cos(\vec{v_i}, \vec{w_j}) = \frac{\vec{v_i} \bullet \vec{w_j}}{|\vec{v_i}| \bullet |\vec{w_j}|} = \frac{\sum_{s=1}^{N} v_{i,s} w_{j,s}}{\sqrt{\sum_{s=1}^{N} v_{i,s}^2} \sqrt{\sum_{s=1}^{N} w_{j,s}^2}}$$

Step 2: cluster the vectors according to the obtained similarities

- Best clustering algo.: *optimized repeated bisections*; $K = 60$

Relation clustering

- Each verbal relation is represented as a vector of its WordNet-synonyms, e.g.:

construct1(construct1, build1, make17)
make24(make24, build1, construct1)

Step 1: calculate the similarity between synonym vectors

$$\cos(\vec{v_i}, \vec{w_j}) = \frac{\vec{v_i} \bullet \vec{w_j}}{|\vec{v_i}| \bullet |\vec{w_j}|} = \frac{\sum_{s=1}^{N} v_{i,s} w_{j,s}}{\sqrt{\sum_{s=1}^{N} v_{i,s}^2} \sqrt{\sum_{s=1}^{N} w_{j,s}^2}}$$

Step 2: cluster the vectors according to the obtained similarities

- Best clustering algo.: *optimized repeated bisections*; $K = 60$

Relation clustering

- Each verbal relation is represented as a vector of its WordNet-synonyms, e.g.:

construct1(construct1, build1, make17)
make24(make24, build1, construct1)

Step 1: calculate the similarity between synonym vectors

$$\cos(\vec{v_i}, \vec{w_j}) = \frac{\vec{v_i} \bullet \vec{w_j}}{|\vec{v_i}| \bullet |\vec{w_j}|} = \frac{\sum_{s=1}^{N} v_{i,s} w_{j,s}}{\sqrt{\sum_{s=1}^{N} v_{i,s}^2} \sqrt{\sum_{s=1}^{N} w_{j,s}^2}}$$

Step 2: cluster the vectors according to the obtained similarities

- Best clustering algo.: *optimized repeated bisections*; $K = 60$

Relation clustering

- Each verbal relation is represented as a vector of its WordNet-synonyms, e.g.:

construct1(construct1, build1, make17)
make24(make24, build1, construct1)

Step 1: calculate the similarity between synonym vectors

$$\cos(\vec{v_i}, \vec{w_j}) = \frac{\vec{v_i} \bullet \vec{w_j}}{|\vec{v_i}| \bullet |\vec{w_j}|} = \frac{\sum_{s=1}^{N} v_{i,s} w_{j,s}}{\sqrt{\sum_{s=1}^{N} v_{i,s}^2} \sqrt{\sum_{s=1}^{N} w_{j,s}^2}}$$

Step 2: cluster the vectors according to the obtained similarities

- Best clustering algo.: *optimized repeated bisections*; $K = 60$

Examples of relational clusters

- comprise, contain, stop, block
- bound, limit, restrain, inhibit
- curve, cut, reduce, trim
- differentiate, distinguish, separate
- extract, pull-out
- delete, erase
- enter, insert, introduce
- tighten, fasten, secure
- associate, connect, join, link, relate
- become, come, release, turn

Cluster labeling

Internal cluster labeling:

- Frequency oriented labeling
- Verb hyperonym-oriented labeling
- Thesaurus frequency-oriented labeling

Differential cluster labeling:

- VHyper MI-oriented labeling
- VHyper χ^2-oriented labeling
- Thesaurus MI-oriented labeling
- Thesaurus χ^2-oriented labeling

Cluster labeling

Internal cluster labeling:

- Frequency oriented labeling
- Verb hyperonym-oriented labeling
- Thesaurus frequency-oriented labeling

Differential cluster labeling:

- VHyper MI-oriented labeling
- VHyper χ^2-oriented labeling
- Thesaurus MI-oriented labeling
- Thesaurus χ^2-oriented labeling

Cluster labeling

Internal cluster labeling:

- Frequency oriented labeling
- Verb hyperonym-oriented labeling
- Thesaurus frequency-oriented labeling

Differential cluster labeling:

- VHyper MI-oriented labeling
- VHyper χ^2-oriented labeling
- Thesaurus MI-oriented labeling
- Thesaurus χ^2-oriented labeling

Cluster labeling

Internal cluster labeling:

- Frequency oriented labeling
- Verb hyperonym-oriented labeling
- Thesaurus frequency-oriented labeling

Differential cluster labeling:

- VHyper MI-oriented labeling
- VHyper χ^2-oriented labeling
- Thesaurus MI-oriented labeling
- Thesaurus χ^2-oriented labeling

Cluster labeling

Internal cluster labeling:

- Frequency oriented labeling
- Verb hyperonym-oriented labeling
- Thesaurus frequency-oriented labeling

Differential cluster labeling:

- VHyper MI-oriented labeling
- VHyper χ^2-oriented labeling
- Thesaurus MI-oriented labeling
- Thesaurus χ^2-oriented labeling

Internal cluster labeling

C_i = [bound, limit, restrain, inhibit, fasten, fix, secure, lock]

Frequency oriented labeling

C_i = [bound:63, limit:74, restrain:21, inhibit:101, fasten:49, fix:53, secure:13, lock:28]

Verb hyperonym-oriented labeling

fasten = [fasten1, fix2, fix firmly1, fasten2, link1, put together1, tie3, [...]]

Thesaurus frequency-oriented labeling

fasten = [fix, secure, attach, tighten, change, alter, modify, [...]]

Internal cluster labeling

C_i = [bound, limit, restrain, inhibit, fasten, fix, secure, lock]

Frequency oriented labeling

C_i = [bound:63, limit:74, restrain:21, inhibit:101, fasten:49, fix:53, secure:13, lock:28]

Verb hyperonym-oriented labeling

fasten = [fasten1, fix2, fix firmly1, fasten2, link1, put together1, tie3, [...]]

Thesaurus frequency-oriented labeling

fasten = [fix, secure, attach, tighten, change, alter, modify, [...]]

Internal cluster labeling

C_i = [bound, limit, restrain, inhibit, fasten, fix, secure, lock]

Frequency oriented labeling

C_i = [bound:63, limit:74, restrain:21, inhibit:101, fasten:49, fix:53, secure:13, lock:28]

Verb hyperonym-oriented labeling

fasten = [fasten1, fix2, fix firmly1, fasten2, link1, put together1, tie3, [...]]

Thesaurus frequency-oriented labeling

fasten = [fix, secure, attach, tighten, change, alter, modify, [...]]

Internal cluster labeling

C_i = [bound, limit, restrain, inhibit, fasten, fix, secure, lock]

Frequency oriented labeling

C_i = [bound:63, limit:74, restrain:21, inhibit:101, fasten:49, fix:53, secure:13, lock:28]

Verb hyperonym-oriented labeling

fasten = [fasten1, fix2, fix firmly1, fasten2, link1, put together1, tie3, [...]]

Thesaurus frequency-oriented labeling

fasten = [fix, secure, attach, tighten, change, alter, modify, [...]]

Differential cluster labeling

- **Based on Mutual Information**:
  - VHyper MI-oriented labeling
  - Thesaurus MI-oriented labeling

$$I(X,Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y)\ log2\ \frac{p(x,y)}{p_1(x)p_2(y)}$$

- **Based on χ^2 test**:
  - VHyper χ^2-oriented labeling
  - Thesaurus χ^2-oriented labeling

$$\chi^2 = \sum_{a \epsilon A} \sum_{b \epsilon B} \frac{(O_{a,b} - E_{a,b})^2}{E_{a,b}}$$

Differential cluster labeling

- **Based on Mutual Information**:
  - VHyper MI-oriented labeling
  - Thesaurus MI-oriented labeling

$$I(X,Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y)\ log2\ \frac{p(x,y)}{p_1(x)p_2(y)}$$

- **Based on χ^2 test**:
  - VHyper χ^2-oriented labeling
  - Thesaurus χ^2-oriented labeling

$$\chi^2 = \sum_{a \epsilon A} \sum_{b \epsilon B} \frac{(O_{a,b} - E_{a,b})^2}{E_{a,b}}$$

Examples of the internal cluster labeling strategies

Gold Standard Clusters	GS	Freq	VHyper	ThesFreq
{comprise, contain, have, include}	contain	comprise	comprise	get
{bound, limit, restrain, inhibit}	limit	inhibit	determine	bind
{tighten, fasten, fix, secure, deposit}	fix	fix	fix	attach
{compress, trim, reduce, minimize}	reduce	reduce	cut	lessen
{extract, pull-out}	extract	extract	remove	take-out
{ colorredremove, cut, delete, erase, exclude}	remove	remove	remove	take-out
{enter, insert, interpose, introduce, enclose}	insert	insert	connect	introduce
{apply, feed, provide, give, use, supply, render}	produce	provide	provide	give
{hold , maintain, retain, support, prevent}	keep	support	maintain	hold
{accord, allow, let, permit}	let	accord	have	permit

Gold Standard Clusters	GS	VHyper-MI	VHyperχ^2	ThesMI	Thesχ^2
{comprise, contain, have, include}	contain	comprise	incorporate	incorporate	incorporate
{bound, limit, restrain, inhibit}	limit	restrict	restrict	throttle	restrict
{tighten, fasten, fix, secure, deposit}	fix	put	lay	find out	localise
{compress, trim, reduce, minimize}	reduce	trim down	thin-out	find-out	minify
{extract, pull-out}	extract	move forcibly	pull-up	pull-up	press-out
{remove, cut, delete, erase, exclude}	remove	erase	kill	cancel	take-out
{enter, insert, interpose, introduce, enclose}	insert	shut-it	enclose	pull-in	pull-in
{apply, feed, provide, give, use, supply, render}	produce	administer	furnish	furnish	furnish
{hold , maintain, retain, support, prevent}	keep	hold on	hold on to	defend	defend
{accord, allow, let, permit}	let	grant	grant	consent	consent

Cluster labeling fallback strategies

- Frequency labeling

Label candidate	MI	Corpus frequency
come near	1390.48	2
fill	1390.48	**87**
get together	1390.48	0

- Random fallback

Cluster	Random label
converge, meet, satisfy	get together
record, file	impeach
accord, allow, let, permit	grant

Cluster labeling fallback strategies

- Frequency labeling

Label candidate	MI	Corpus frequency
come near	1390.48	2
fill	1390.48	**87**
get together	1390.48	0

- Random fallback

Cluster	Random label
converge, meet, satisfy	get together
record, file	impeach
accord, allow, let, permit	grant

Cluster labeling fallback strategies

- Frequency labeling

Label candidate	MI	Corpus frequency
come near	1390.48	2
fill	1390.48	**87**
get together	1390.48	0

- Random fallback

Cluster	Random label
converge, meet, satisfy	get together
record, file	impeach
accord, allow, let, permit	grant

Cluster labeling fallback strategies

- Frequency labeling

Label candidate	MI	Corpus frequency
come near	1390.48	2
fill	1390.48	**87**
get together	1390.48	0

- Random fallback

Cluster	Random label
converge, meet, satisfy	get together
record, file	impeach
accord, allow, let, permit	grant

Cluster labeling fallback strategies

- Frequency labeling

Label candidate	MI	Corpus frequency
come near	1390.48	2
fill	1390.48	**87**
get together	1390.48	0

- Random fallback

Cluster	Random label
converge, meet, satisfy	get together
record, file	impeach
accord, allow, let, permit	grant

Illustration of the approach

comprise (automatic focusing device, lens driving circuit)

focus(objective lens, light beam)

emit (light source, light beam, track of an information recording medium)

reflect(information recording medium, reflected light beam, focal spot, objective lens)

emit(light source, light beam)

be-capable(optical element, **cause**(optical element, astigmatic **aberration**(separated reflected light beam)))

include(astigmatic optical system, optical element)

extend(second axis, **perpendicularly**(first axis))

Illustration of the approach

---

comprise (automatic focusing device, lens driving circuit)

focus(objective lens, light beam)

point (light source, light beam, track of an information recording medium)

reflect(information recording medium, reflected light beam, focal spot, objective lens)

point(light source, light beam)

be-capable(optical element, **cause**(optical element, astigmatic **aberration**(separated reflected light beam)))

comprise(astigmatic optical system, optical element)

extend(second axis, **perpendicularly**(first axis))

Outline

Evaluation: simplification

- **Evaluation of the claim segmentation:**

	# automatic segments	# 1:1 alignments	F-score
Baseline	4078	3282	0.52
Our system	5342	4327	**0.73**

- **Evaluation of coreference resolution:**

# manual coref.	# automatic coref.	# automatic correct	F-score
199	190	159	**0.81**

Evaluation: simplification

- **Evaluation of claim structuring:**

	# automatic spans	# correct spans	F-score
Perfect input	201	114	**0.56**
Raw	143	75	0.42
Baseline	227	79	0.35

Evaluation: relation extraction

- Comparative evaluation
- Syntactic parsing vs. chunking
- Blohm & Cimiano (2007)
- Small evaluation set (5 patents)

Manual extraction	**Blohm & Cimiano**	**Our approach**
# 94	# 51 (54%)	# **67** (71%)

Evaluation: relational clustering

- **Gold standard**:
  - Corpus of 2076 English patent claims
  - 193 most frequent verbs
  - 54 classes
- **Clustering baselines**
  - (a) Verbs are randomly assigned to a cluster (50 repetitions)
  - (b) The total of manual clusters and their category sizes are preserved and verbs are randomly assigned to them (50 repetitions)
- **Clustering evaluation measures**:
  - Purity
  - Adjusted Rand Index
  - Pair-wise evaluation

Evaluation: relation clustering results

Eval. measures	**Baseline (a)**	**Baseline (b)**	**Our clustering**
Purity	0.33	0.34	0.73
ARI	0.02	0.01	0.34
Pair-wise Recall	0.01	0.03	0.32
Pair-wise Precision	0.02	0.03	0.33
Pair-wise F-score	0.02	0.03	0.32

Evaluation: WSD-based similarity vs. automatic similarity

Eval. measures	**WSD Verbs**	**WSD Our Approach**
Purity	0.6	0.7
ARI	0.2	0.3
Pair wise Recall	0.4	0.4
Pair wise Precision	0.06	0.1
Pair wise F-score	0.1	0.1

Evaluation: relational cluster labeling

Results of internal labeling:

	% Correct	% Partially correct	% Incorrect
Gold standard	77%	17%	7%
Freq	**78%**	20%	2%
VHyper	43%	25%	32%
ThesFreq	58%	26%	16%

Results of differential labeling:

	% Correct	% Partially correct	% Incorrect
Gold standard	77%	17%	7%
VHyper-MI	50%	45%	5%
VHyper χ^2	60%	27%	13%
ThesMI	**70%**	25%	5%
Thes χ^2	67%	22%	11%

Evaluation: relational cluster labeling

Results of internal labeling:

	% Correct	% Partially correct	% Incorrect
Gold standard	77%	17%	7%
Freq	**78%**	20%	2%
VHyper	43%	25%	32%
ThesFreq	58%	26%	16%

Results of differential labeling:

	% Correct	% Partially correct	% Incorrect
Gold standard	77%	17%	7%
VHyper-MI	50%	45%	5%
VHyper χ^2	60%	27%	13%
ThesMI	**70%**	25%	5%
Thes χ^2	67%	22%	11%

Outline

Contributions

The first linguistic study about patent claim style

- facilitates the development of NLP technologies
- supports patent users

A framework for flexible, *n*-ary and unlexicalized RE

- distilled from deep syntactic structures
- not limited in type, number or arity

A framework for unsupervised relation generalization

- the approach is simple and independent of the corpus
- relation classes are provided by cluster labeling

A claim paraphrasing application

- relies on surface linguistic analysis
- claims are more readable

Contributions

The first linguistic study about patent claim style

- facilitates the development of NLP technologies
- supports patent users

A framework for flexible, *n*-ary and unlexicalized RE

- distilled from deep syntactic structures
- not limited in type, number or arity

A framework for unsupervised relation generalization

- the approach is simple and independent of the corpus
- relation classes are provided by cluster labeling

A claim paraphrasing application

- relies on surface linguistic analysis
- claims are more readable

Contributions

The first linguistic study about patent claim style

- facilitates the development of NLP technologies
- supports patent users

A framework for flexible, *n*-ary and unlexicalized RE

- distilled from deep syntactic structures
- not limited in type, number or arity

A framework for unsupervised relation generalization

- the approach is simple and independent of the corpus
- relation classes are provided by cluster labeling

A claim paraphrasing application

- relies on surface linguistic analysis
- claims are more readable

Contributions

The first linguistic study about patent claim style

- facilitates the development of NLP technologies
- supports patent users

A framework for flexible, *n*-ary and unlexicalized RE

- distilled from deep syntactic structures
- not limited in type, number or arity

A framework for unsupervised relation generalization

- the approach is simple and independent of the corpus
- relation classes are provided by cluster labeling

A claim paraphrasing application

- relies on surface linguistic analysis
- claims are more readable

Contributions

The first linguistic study about patent claim style

- facilitates the development of NLP technologies
- supports patent users

A framework for flexible, *n*-ary and unlexicalized RE

- distilled from deep syntactic structures
- not limited in type, number or arity

A framework for unsupervised relation generalization

- the approach is simple and independent of the corpus
- relation classes are provided by cluster labeling

A claim paraphrasing application

- relies on surface linguistic analysis
- claims are more readable

Limitations

- claims vs. other patent sections
- some claim types have not been covered
- our approach is English-specific

Limitations

- claims vs. other patent sections
- some claim types have not been covered
- our approach is English-specific

Limitations

- claims vs. other patent sections
- some claim types have not been covered
- our approach is English-specific

Applications

- Content visualization
- Ontology learning
- Search of similar documents
- Question answering

* Support patent users

* Support patent technologies development

Future work

- **Improvement of the current approach**
  - Dispense with the claim simplification module
  - Use a more powerful and patent-adapted syntactic dependency parser

- **Extension of the current approach**
  - Apply the RE approach to distil relations expressed by predicative nouns and adjectives
  - Propose an approach to group similar nouns and similar adjectives

Future work

- **Improvement of the current approach**
  - Dispense with the claim simplification module
  - Use a more powerful and patent-adapted syntactic dependency parser

- **Extension of the current approach**
  - Apply the RE approach to distil relations expressed by predicative nouns and adjectives
  - Propose an approach to group similar nouns and similar adjectives

Future work

- **Improvement of the current approach**
  - Dispense with the claim simplification module
  - Use a more powerful and patent-adapted syntactic dependency parser

- **Extension of the current approach**
  - Apply the RE approach to distil relations expressed by predicative nouns and adjectives
  - Propose an approach to group similar nouns and similar adjectives

Thank you!

Questions?