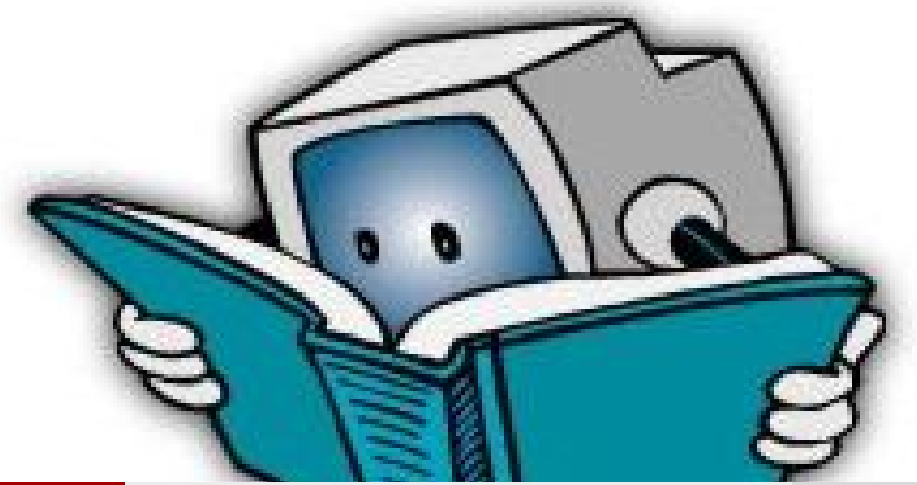
ClausIE: Clause-Based Open Information Extraction

Luciano Del Corro Rainer Gemulla

Max-Planck-Institut für Informatik

May 2013



Open Information Extraction: From sentences to propositions

GOAL: Extract information from natural text

Open Information Extraction: From sentences to propositions

GOAL: Extract information from natural text

Sentence

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Open Information Extraction: From sentences to propositions

GOAL: Extract information from natural text

Sentence

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Extractions/Propositions

(Bell,	'is',	a telecommunication company)
(Bell,	is based in,	Los Angeles)
(Bell,	makes,	electronic products)
(Bell,	distributes,	electronic products)
	...	

Open Information Extraction: From sentences to propositions

GOAL: Extract information from natural text

Sentence

Bell, a telecommunication company, which is based in Los Angeles, makes and distributes electronic, computer and building products.

Extractions/Propositions

(Bell,	'is',	a telecommunication company)
(Bell,	is based in,	Los Angeles)
(Bell,	makes,	electronic products)
(Bell,	distributes,	electronic products)
	...	

Most OIE extractors

- Propositions expressed as triples $(arg_1, relation, arg_2)$
- Verb based relation
- Arguments restricted to noun phrases

Open Information Extraction: challenges and applications

Challenges/Requirements

- Domain independent
- Unbounded set of relations
- No filtering of information
- Structured output
- Scalable

Open Information Extraction: challenges and applications

Challenges/Requirements

- Domain independent
- Unbounded set of relations
- No filtering of information
- Structured output
- Scalable

Applications

- Structured search
- Automatic ontology construction
- Question answering
- Semantic role labeling, discourse parsing, ... ?

Outline

Outline

Information and Representation: a two-step approach

Information

- What information is expressed?
- How much to retain?
- How to identify it? (e.g. non-verb mediated propositions‘)
  - ⋆ Messi, a golden ball winner, plays in Barcelona

Information and Representation: a two-step approach

Information

- What information is expressed?
- How much to retain?
- How to identify it? (e.g. non-verb mediated propositions‘)
  - ⋆ Messi, a golden ball winner, plays in Barcelona

Representation

- What is the form of the relation?
  - ⋆ Messi plays in Barcelona $\rightarrow$ *plays* or *plays in*
- Triples or n-ary propositions?
  - ⋆ *(Messi, plays football in, Barcelona)* or *(Messi, plays, football, in Barcelona)*
- What should be the scope of the arguments?
  - ⋆ Gandhi was vegetarian

Information and Representation: a two-step approach

Information

- What information is expressed?
- How much to retain?
- How to identify it? (e.g. non-verb mediated propositions‘)
  - ⋆ Messi, a golden ball winner, plays in Barcelona

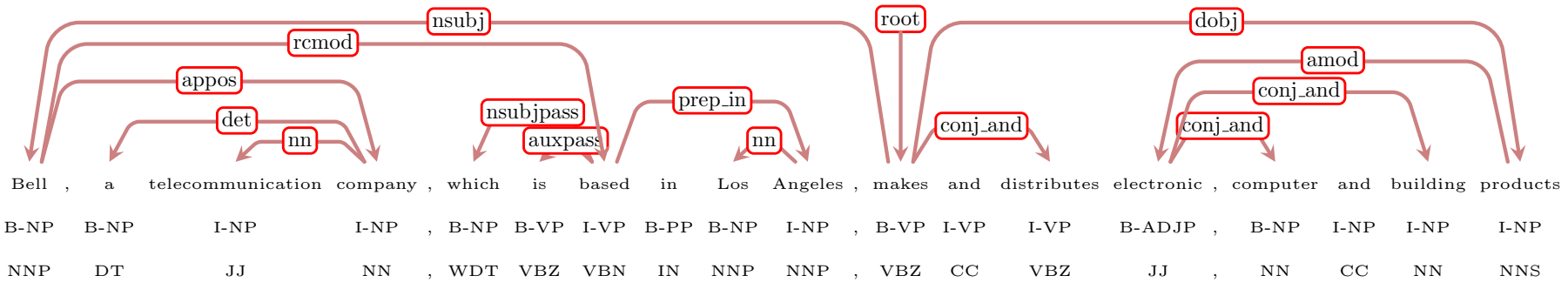
Representation

- What is the form of the relation?
  - ⋆ Messi plays in Barcelona → *plays* or *plays in*
- Triples or n-ary propositions?
  - ⋆ *(Messi, plays football in, Barcelona)* or *(Messi, plays, football, in Barcelona)*
- What should be the scope of the arguments?
  - ⋆ Gandhi was vegetarian

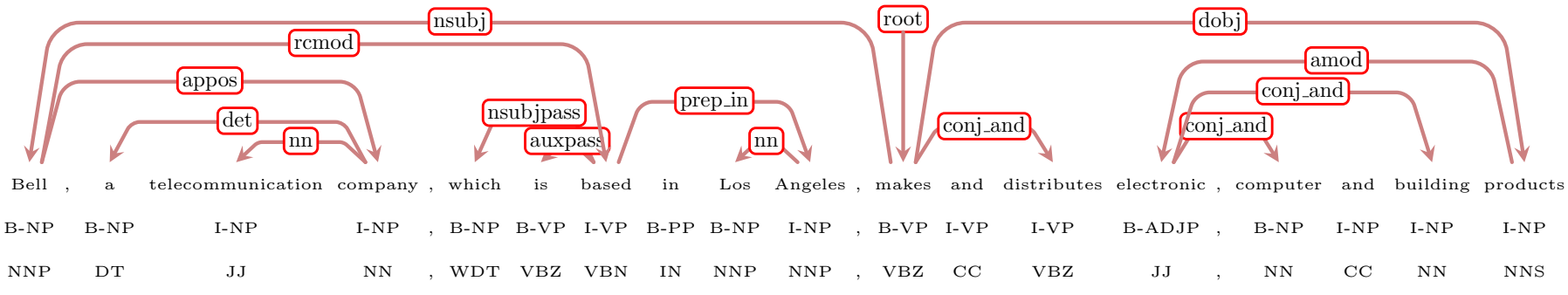
We aim to separate these two phases

Outline

Open Information Extractors and Language Technology



Open Information Extractors and Language Technology



Chunks/POS

- TextRunner
- WOE^{pos}
- Reverb

Dependency Parser

- Wanderlust
- WOE^{parse}
- KrakeN
- OLLIE

Outline

Clause Essentials

- A clause is like a simple sentence
  - ⋆ Paul eats a chocolate bar

Clause Essentials

- A clause is like a simple sentence
  - ⋆ Paul eats a chocolate bar
- A sentence can be composed by more than one clause
  - ⋆ Anna drinks coffee and Bob plays football

Clause Essentials

- A clause is like a simple sentence
  - ⋆ Paul eats a chocolate bar
- A sentence can be composed by more than one clause
  - ⋆ Anna drinks coffee and Bob plays football
- Each clause encodes one or more propositions

Clause Essentials

- A clause is like a simple sentence
  - ⋆ Paul eats a chocolate bar
- A sentence can be composed by more than one clause
  - ⋆ Anna drinks coffee and Bob plays football
- Each clause encodes one or more propositions
- Clauses can have optional adverbials
  - ⋆ He will take the exam in May

Clause Essentials

- A clause is like a simple sentence
  - ⋆ Paul eats a chocolate bar
- A sentence can be composed by more than one clause
  - ⋆ Anna drinks coffee and Bob plays football
- Each clause encodes one or more propositions
- Clauses can have optional adverbials
  - ⋆ He will take the exam in May
- A minimal clause is a clause without its optional adverbials
  - ⋆ He will take the exam

The seven clauses

1. SV_i → Albert Einstein died.

S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object

The seven clauses

1. SV_i → Albert Einstein died.
2. SV_e A → Albert Einstein remained in Princeton.

S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object

The seven clauses

1. SV_i → Albert Einstein died.
2. $SV_e\ A$ → Albert Einstein remained in Princeton.
3. $SV_c\ C$ → Albert Einstein is smart.

S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, $\mathbf{O_i}$**:** Indirect Object, **O:** Direct Object

The seven clauses

1. SV_i → Albert Einstein died.
2. $SV_e\ A$ → Albert Einstein remained in Princeton.
3. $SV_c\ C$ → Albert Einstein is smart.
4. $SV_{mt}\ O$ → Albert Einstein has won the Nobel Prize.
5. $SV_{dt}\ O_i\ O_d$ → RSAS gave Albert Einstein the Nobel Prize.
6. $SV_{ct}\ O\ A$ → The doorman showed Albert Einstein to his office.
7. $SV_{ct}\ O\ C$ → Albert Einstein declared the meeting open.

S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object

The seven clauses

1. SV_i → Albert Einstein died.
2. SV_e A → Albert Einstein remained in Princeton.
3. SV_c C → Albert Einstein is smart.
4. SV_{mt} O → Albert Einstein has won the Nobel Prize.
5. SV_{dt} O_i O_d → RSAS gave Albert Einstein the Nobel Prize.
6. SV_{ct} O A → The doorman showed Albert Einstein to his office.
7. SV_{ct} O C → Albert Einstein declared the meeting open.

S: Subject, **V**

By identifying each minimal clause in a sentence we can identify the essential information

ect

The seven clauses: optional adverbials

Pattern	Clause Type	Example	Derived clauses
		Some extended patterns	
$SV_i AA$	SV	AE died in Princeton in 1955.	(AE, died) (AE, died, in Princeton) (AE, died, in 1955) (AE, died, in Princeton, in 1955)

S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object

The seven clauses: optional adverbials

Pattern	Clause Type	Example	Derived clauses
		Some extended patterns	
SV_iAA	SV	AE died in Princeton in 1955.	(AE, died) (AE, died, in Princeton) (AE, died, in 1955) (AE, died, in Princeton, in 1955)
SV_eAA	SVA	AE remained in Princeton until his death.	(AE, remained, in Princeton) (AE, remained, in Princeton, until his death)

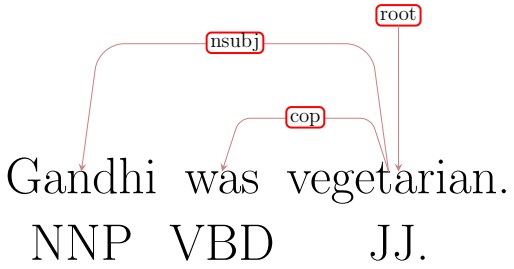
S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object

The seven clauses: optional adverbials

Pattern	Clause Type	Example	Derived clauses
		Some extended patterns	
SV_iAA	SV	AE died in Princeton in 1955.	(AE, died) (AE, died, in Princeton) (AE, died, in 1955) (AE, died, in Princeton, in 1955)
SV_eAA	SVA	AE remained in Princeton until his death.	(AE, remained, in Princeton) (AE, remained, in Princeton, until his death)
SV_cCA	SVC	AE is a scientist of the 20th century.	(AE, is, a scientist) (AE, is, a scientist, of the 20th century)
$SV_{mt}OA$	SVO	AE has won the Nobel Prize in 1921.	(AE, has won, the Nobel Prize) (AE, has won, the Nobel Prize, in 1921)
$ASV_{mt}O$	SVO	In 1921, AE has won the Nobel Prize.	(AE, has won, the Nobel Prize) (AE, has won, the Nobel Prize, in 1921)

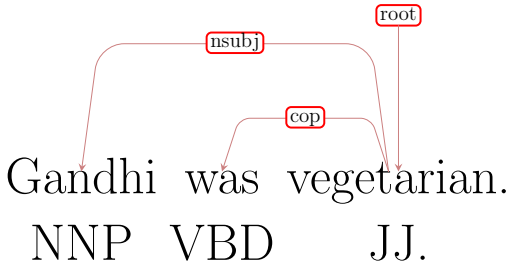
S: Subject, **V:** Verb, **A:** Adverbial, **C:** Complement, **O_i:** Indirect Object, **O:** Direct Object

From clauses to clause types (I)



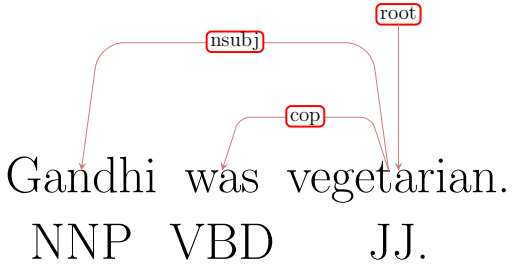
DP

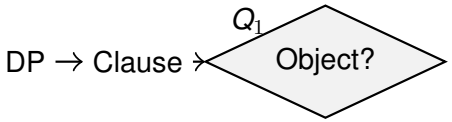
From clauses to clause types (I)



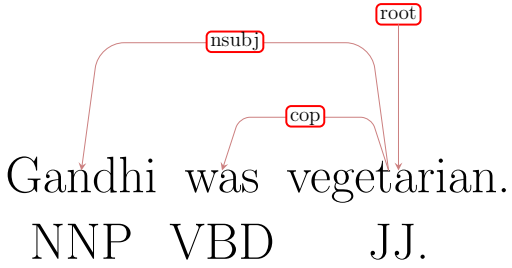
DP → Clause

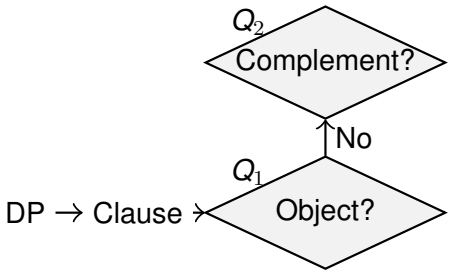
From clauses to clause types (I)



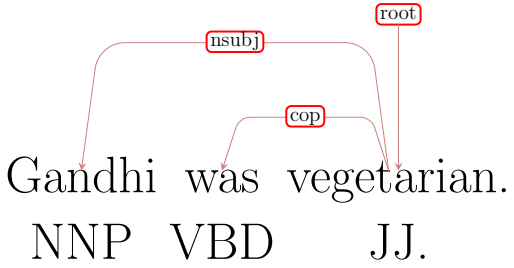


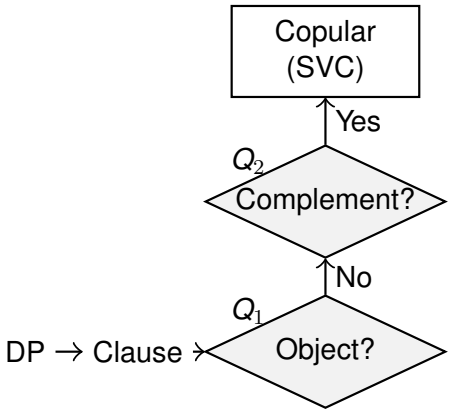
From clauses to clause types (I)



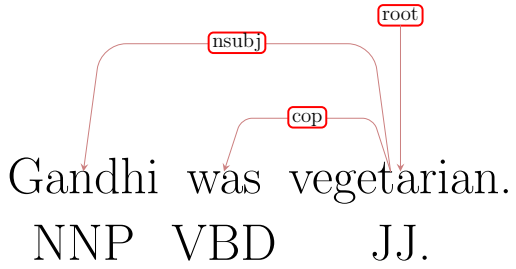


From clauses to clause types (I)

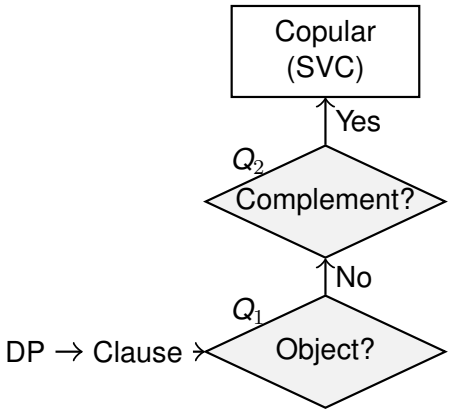




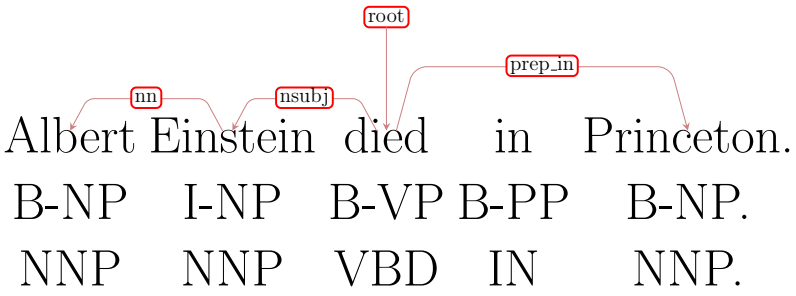
From clauses to clause types (I)



→ (**S:** *Gandhi*, **V:** *was*, **C:** *vegetarian*)

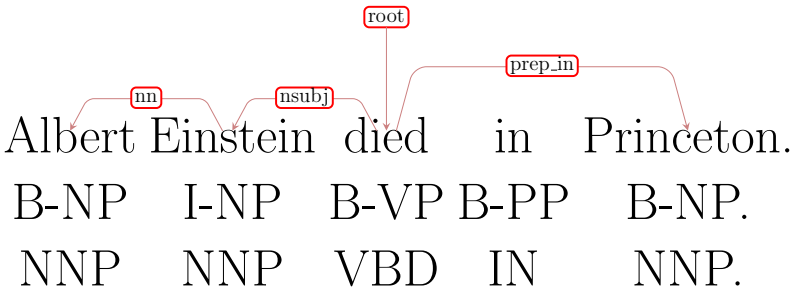


From clauses to clause types (II)



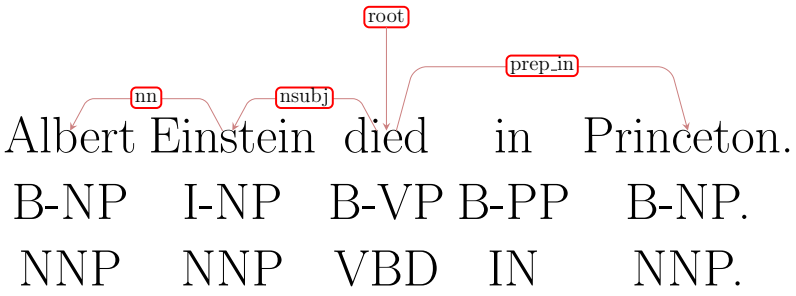
DP

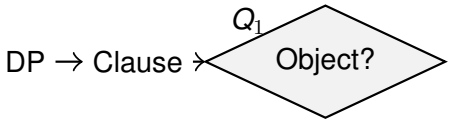
From clauses to clause types (II)



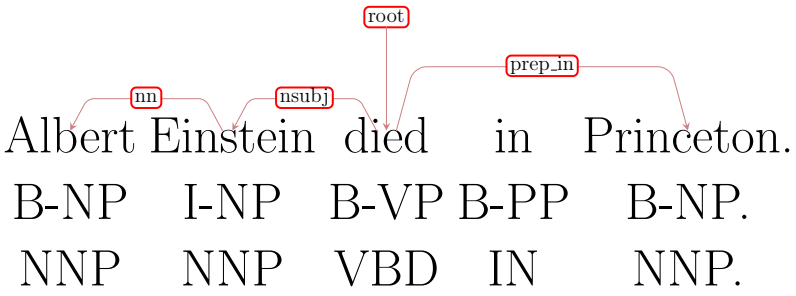
DP → Clause

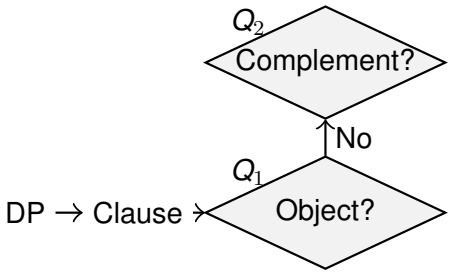
From clauses to clause types (II)



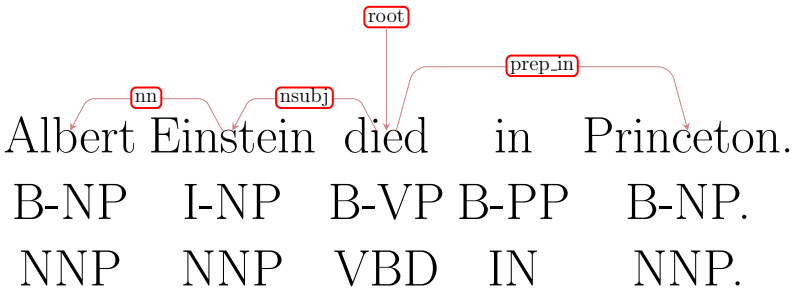


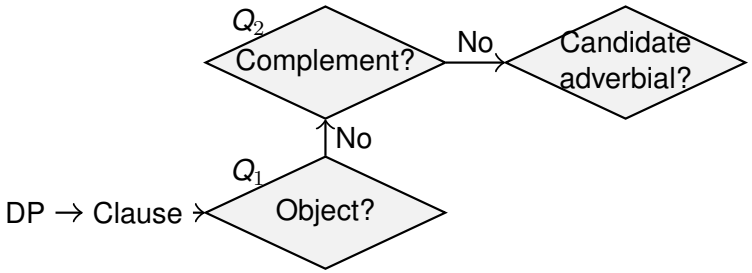
From clauses to clause types (II)



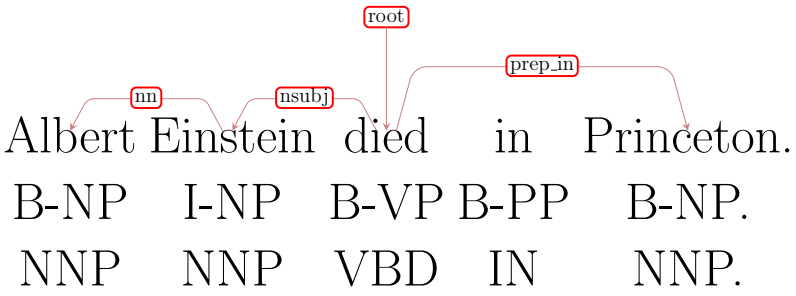


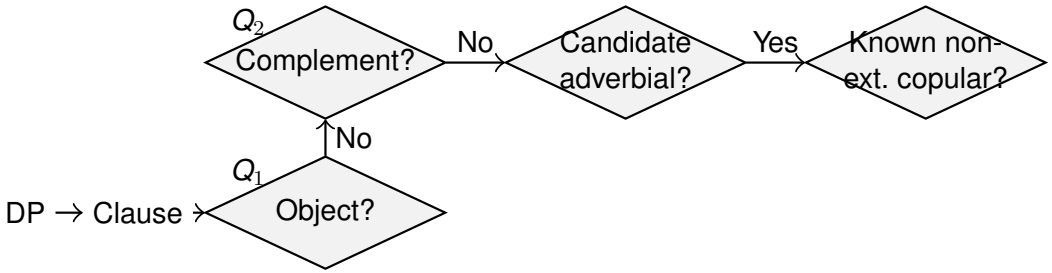
From clauses to clause types (II)



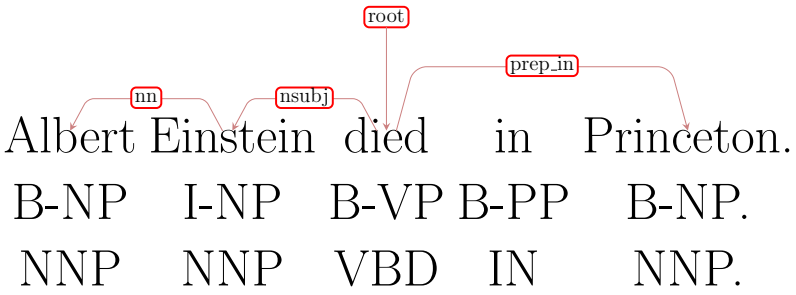


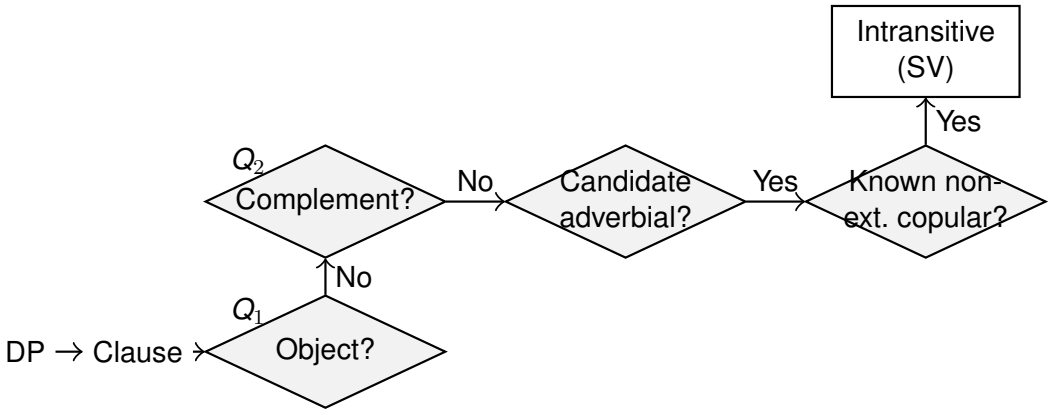
From clauses to clause types (II)



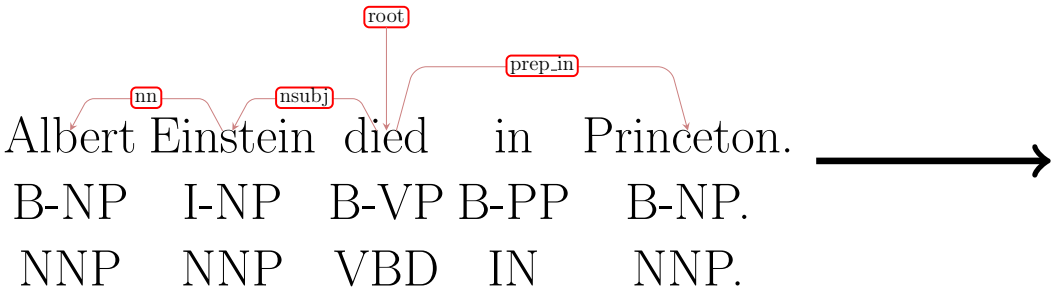


From clauses to clause types (II)

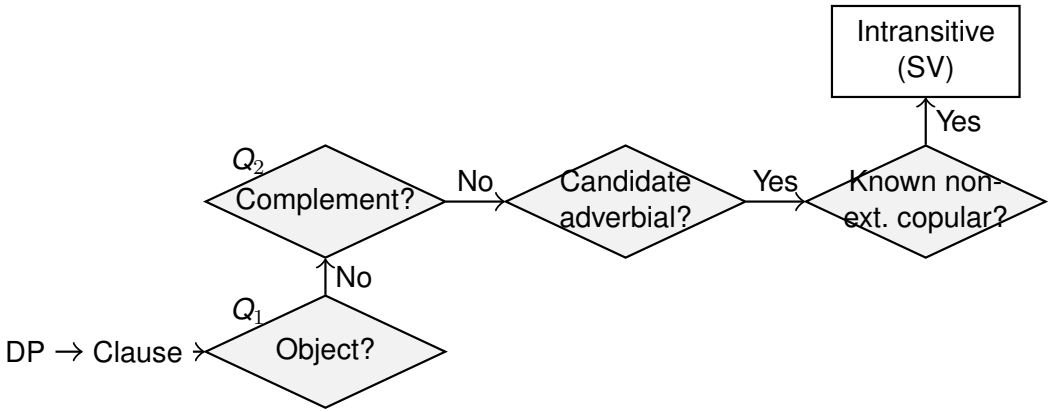




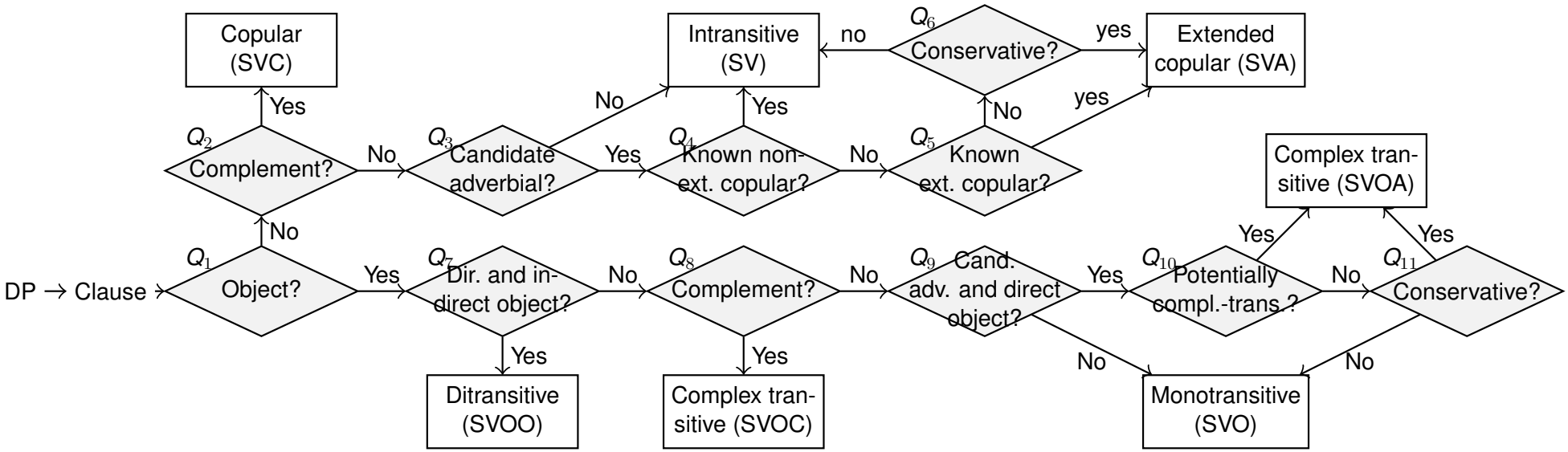
From clauses to clause types (II)



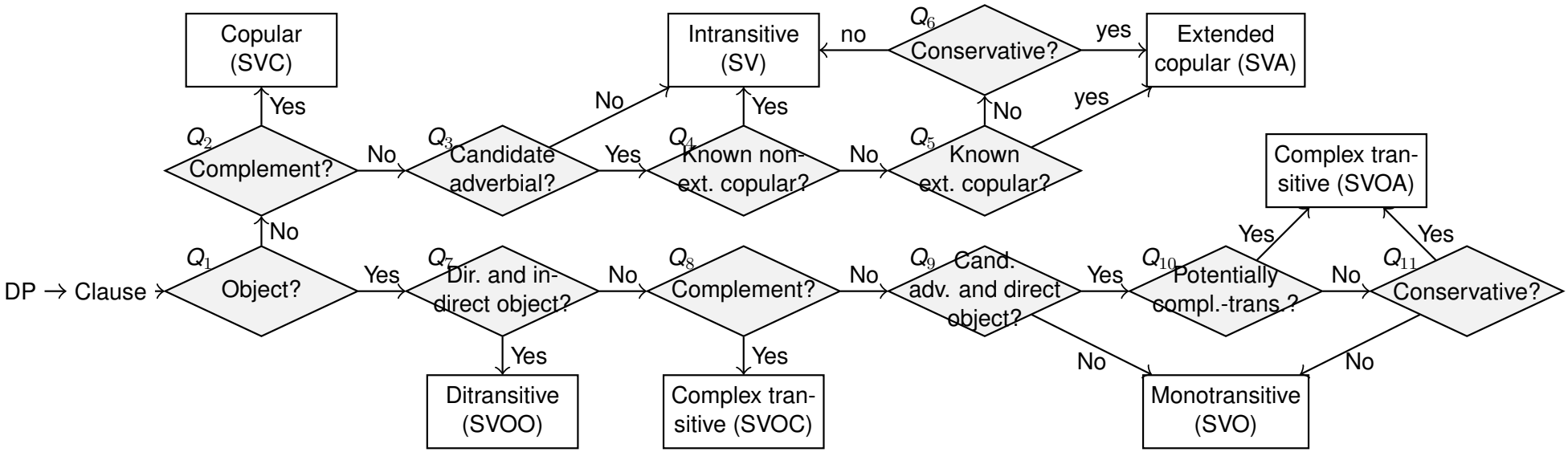
*(**S:** AE, **V:** died,)*
*(**S:** AE, **V:** died, **A:** in Princeton)*



From clauses to clause types (II)

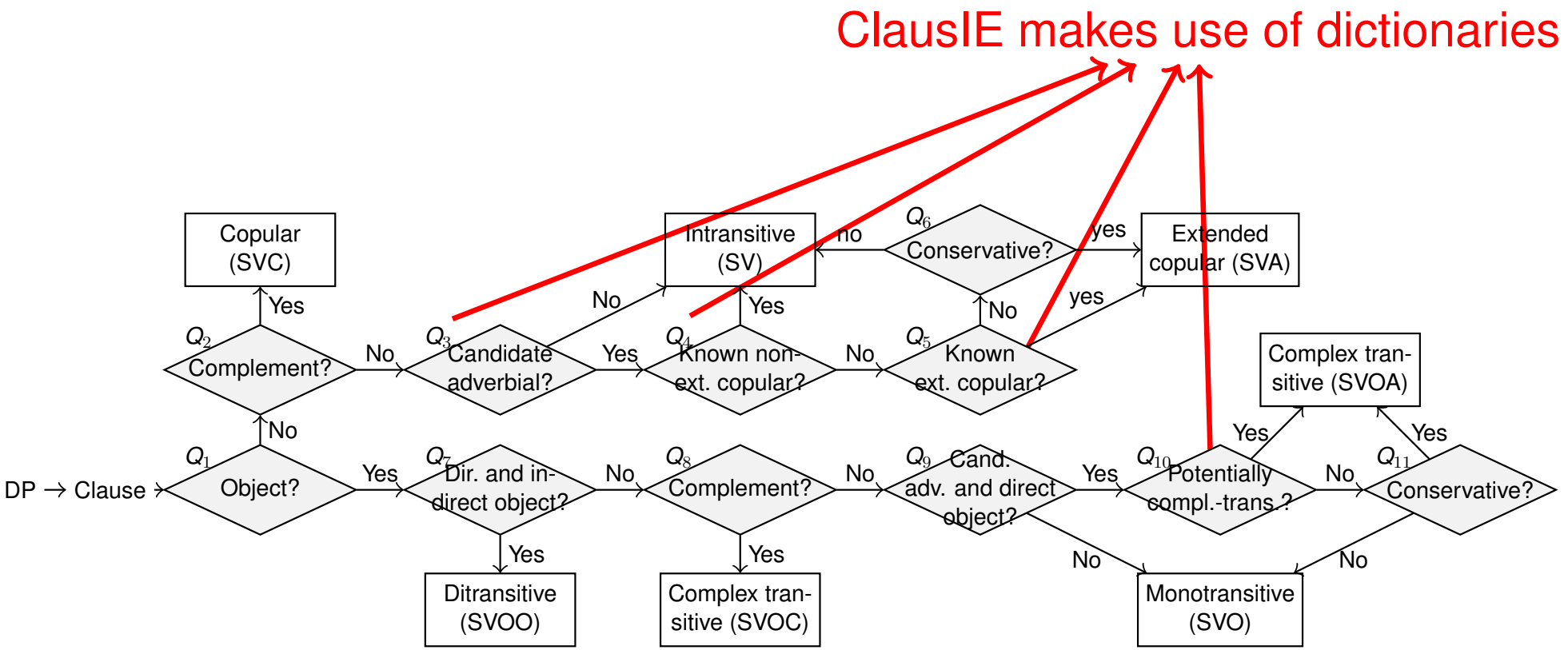


From clauses to clause types (II)



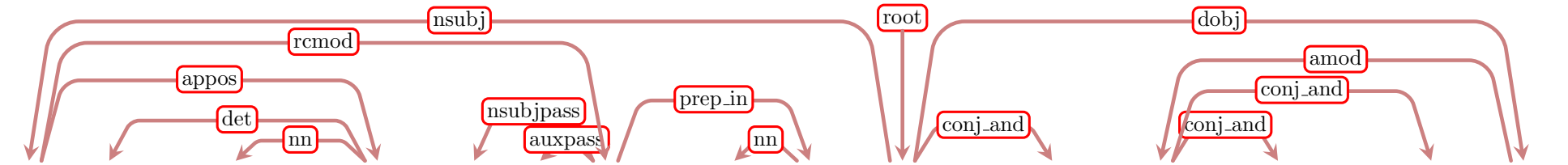
We first identify the information and then generate the proposition.

From clauses to clause types (II)



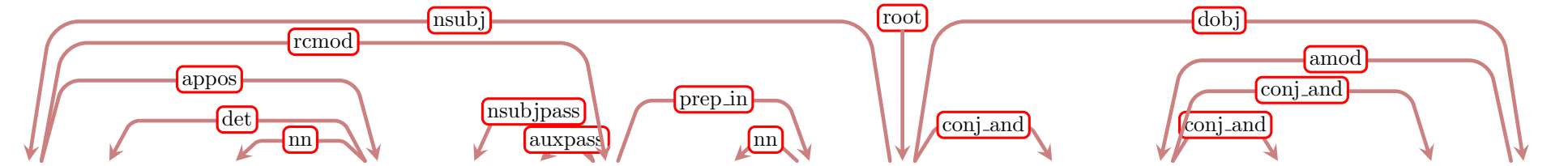
We first identify the information and then generate the proposition.

Example



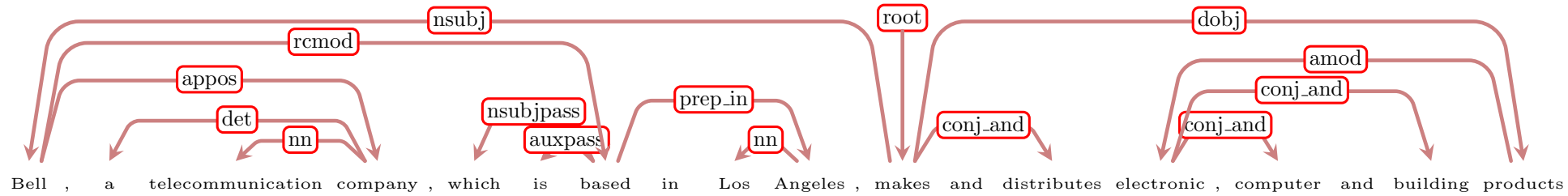
Bell	,	a	telecommunication	company	,	which	is	based	in	Los	Angeles	,	makes	and	distributes	electronic	,	computer	and	building	products
B-NP		B-NP	I-NP	I-NP	,	B-NP	B-VP	I-VP	B-PP	B-NP	I-NP	,	B-VP	I-VP	I-VP	B-ADJP	,	B-NP	I-NP	I-NP	I-NP
NNP		DT	JJ	NN	,	WDT	VBZ	VBN	IN	NNP	NNP	,	VBZ	CC	VBZ	JJ	,	NN	CC	NN	NNS

Example



Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

Example

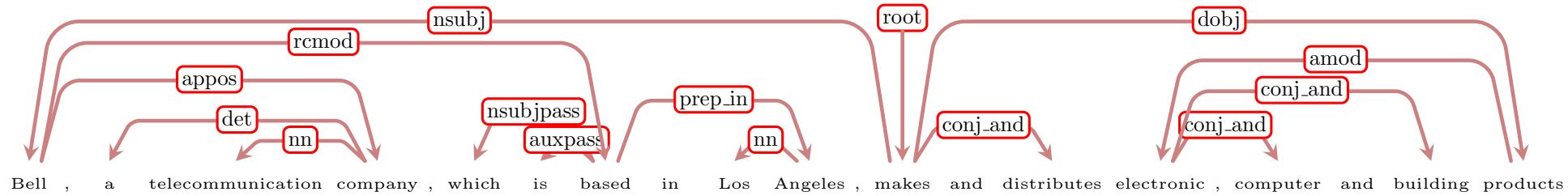


Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

Reverb $\rightarrow$ (*a telecommunication company, is based in, Los Angeles*)

Ollie $\rightarrow$ (*Bell, distributes, electronic , computer and building products*)

Example



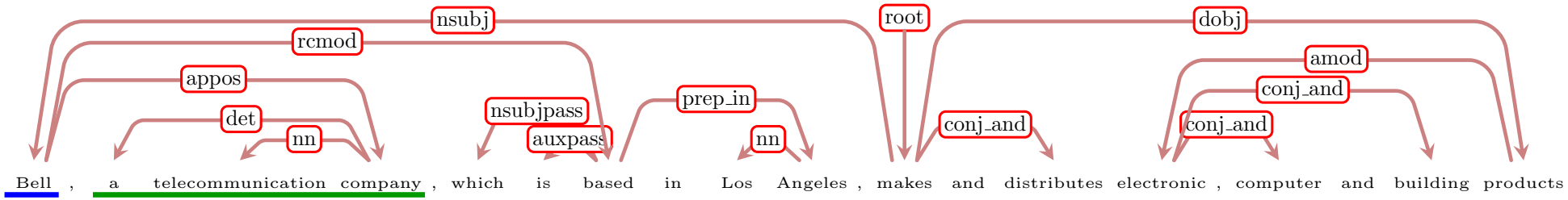
Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

Reverb → (*a telecommunication company,* *is based in,* *Los Angeles*)

Ollie → (*Bell,* *distributes,* *electronic , computer and building products*)

ClausIE →
(**S:** *Bell,* **V:** *'is',* **C:** *a telecommunication company*)
(**S:** *Bell,* **V:** *is based,* **A:** *in Los Angeles*)
(**S:** *Bell,* **V:** *makes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *makes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *makes,* **O:** *building products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *building products*)

Example



Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

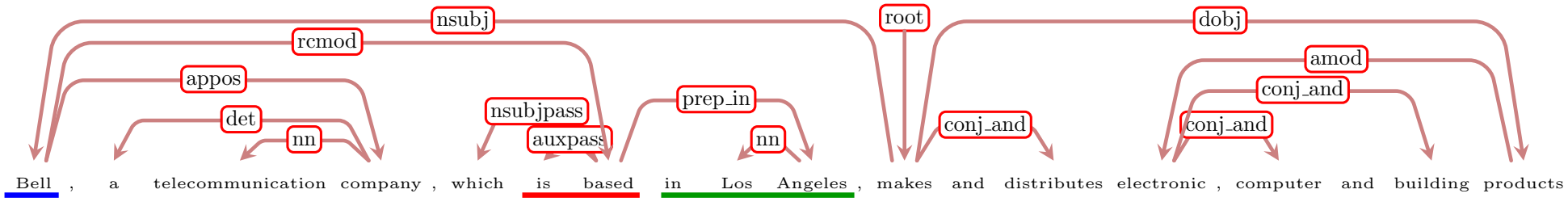
Reverb → (*a telecommunication company,* *is based in,* *Los Angeles*)

Ollie → (*Bell,* *distributes,* *electronic , computer and building products*)

ClausIE →

(**S:** *Bell,*	***V*:** *'is',*	**C:** *a telecommunication company*)
(**S:** *Bell,*	**V:** *is based,*	**A:** *in Los Angeles*)
(**S:** *Bell,*	**V:** *makes,*	**O:** *electronic products*)
(**S:** *Bell,*	**V:** *makes,*	**O:** *computer products*)
(**S:** *Bell,*	**V:** *makes,*	**O:** *building products*)
(**S:** *Bell,*	**V:** *distributes,*	**O:** *electronic products*)
(**S:** *Bell,*	**V:** *distributes,*	**O:** *computer products*)
(**S:** *Bell,*	**V:** *distributes,*	**O:** *building products*)

Example



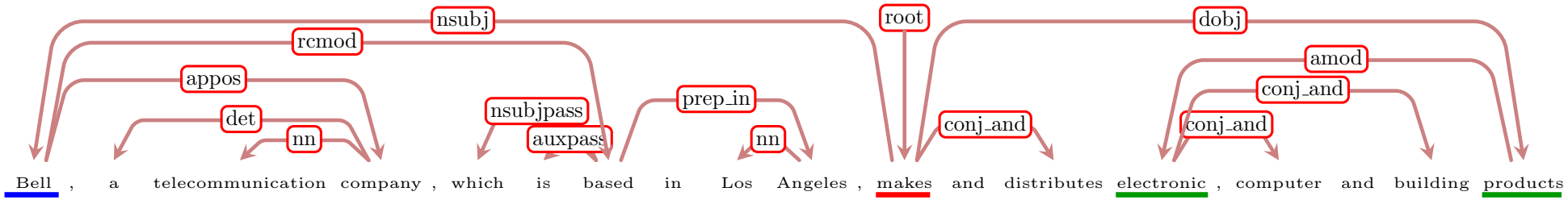
Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

Reverb → (*a telecommunication company,* *is based in,* *Los Angeles*)

Ollie → (*Bell,* *distributes,* *electronic , computer and building products*)

ClausIE →
(**S:** *Bell,* ***V*:** *'is',* **C:** *a telecommunication company*)
(**S:** *Bell,* **V:** *is based,* **A:** *in Los Angeles*)
(**S:** *Bell,* **V:** *makes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *makes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *makes,* **O:** *building products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *building products*)

Example



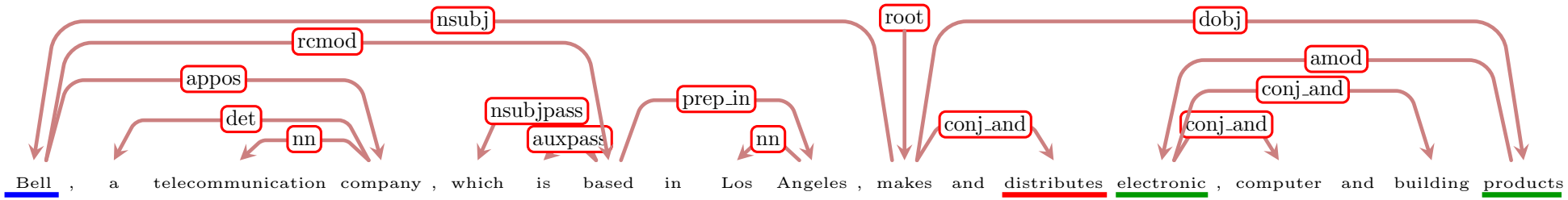
Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

Reverb → (*a telecommunication company,* *is based in,* *Los Angeles*)

Ollie → (*Bell,* *distributes,* *electronic , computer and building products*)

ClausIE →
(**S:** *Bell,* **V:** *'is',* **C:** *a telecommunication company*)
(**S:** *Bell,* **V:** *is based,* **A:** *in Los Angeles*)
(**S:** *Bell,* **V:** *makes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *makes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *makes,* **O:** *building products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *building products*)

Example



Bell, a telecommunication company, which is based in Los Angeles , makes and distributes electronic, computer and building products.

Reverb → (*a telecommunication company,* *is based in,* *Los Angeles*)

Ollie → (*Bell,* *distributes,* *electronic , computer and building products*)

ClausIE →
(**S:** *Bell,* **V:** *'is',* **C:** *a telecommunication company*)
(**S:** *Bell,* **V:** *is based,* **A:** *in Los Angeles*)
(**S:** *Bell,* **V:** *makes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *makes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *makes,* **O:** *building products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *electronic products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *computer products*)
(**S:** *Bell,* **V:** *distributes,* **O:** *building products*)

Identifying information

- ClausIE separates the identification of the information from its representation

Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause

Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data

Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
- Initial support non-verb mediated relations

Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
- Initial support non-verb mediated relations
- Processing of conjunctions (in verbs and subject/arguments)
  - ⋆ Messi and Iniesta play in Barcelona $\rightarrow$ *(Messi, plays, in Barcelona)*, *(Iniesta, plays, in Barcelona)*

Identifying information

- ClausIE separates the identification of the information from its representation
- Identifies essential and optional arguments in a clause
- No training data
- Initial support non-verb mediated relations
- Processing of conjunctions (in verbs and subject/arguments)
  - ⋆ Messi and Iniesta play in Barcelona → *(Messi, plays, in Barcelona)*, *(Iniesta, plays, in Barcelona)*
- Resolution of relative clauses
  - ⋆ I saw the man whose house you like → *(I, saw, the man)*, *(You, like, the man's house)* ...

Proposition Generation: a flexible process

- Arbitrary form of relations
  - ⋆ *(Messi, plays football in, Barcelona)* or *(Messi, plays, football in Barcelona)*

Proposition Generation: a flexible process

- Arbitrary form of relations
  - ⋆ *(Messi, plays football in, Barcelona)* or *(Messi, plays, football in Barcelona)*
- Propositions can be customized (e.g. triple, n-ary, etc)
  - ⋆ *(Messi, plays, football in Barcelona)* or *(Messi, plays, football, in Barcelona)*

Proposition Generation: a flexible process

- Arbitrary form of relations
  - ⋆ *(Messi, plays football in, Barcelona)* or *(Messi, plays, football in Barcelona)*
- Propositions can be customized (e.g. triple, n-ary, etc)
  - ⋆ *(Messi, plays, football in Barcelona)* or *(Messi, plays, football, in Barcelona)*
- Arbitrary argument types (e.g. noun phrases, adjectives, etc)
  - ⋆ *(Gandhi, was, vegetarian)* or *(Gandhi, was, a vegetarian)* or *(Gandhi from Porbandar, was, a vegetarian)*

Proposition Generation: a flexible process

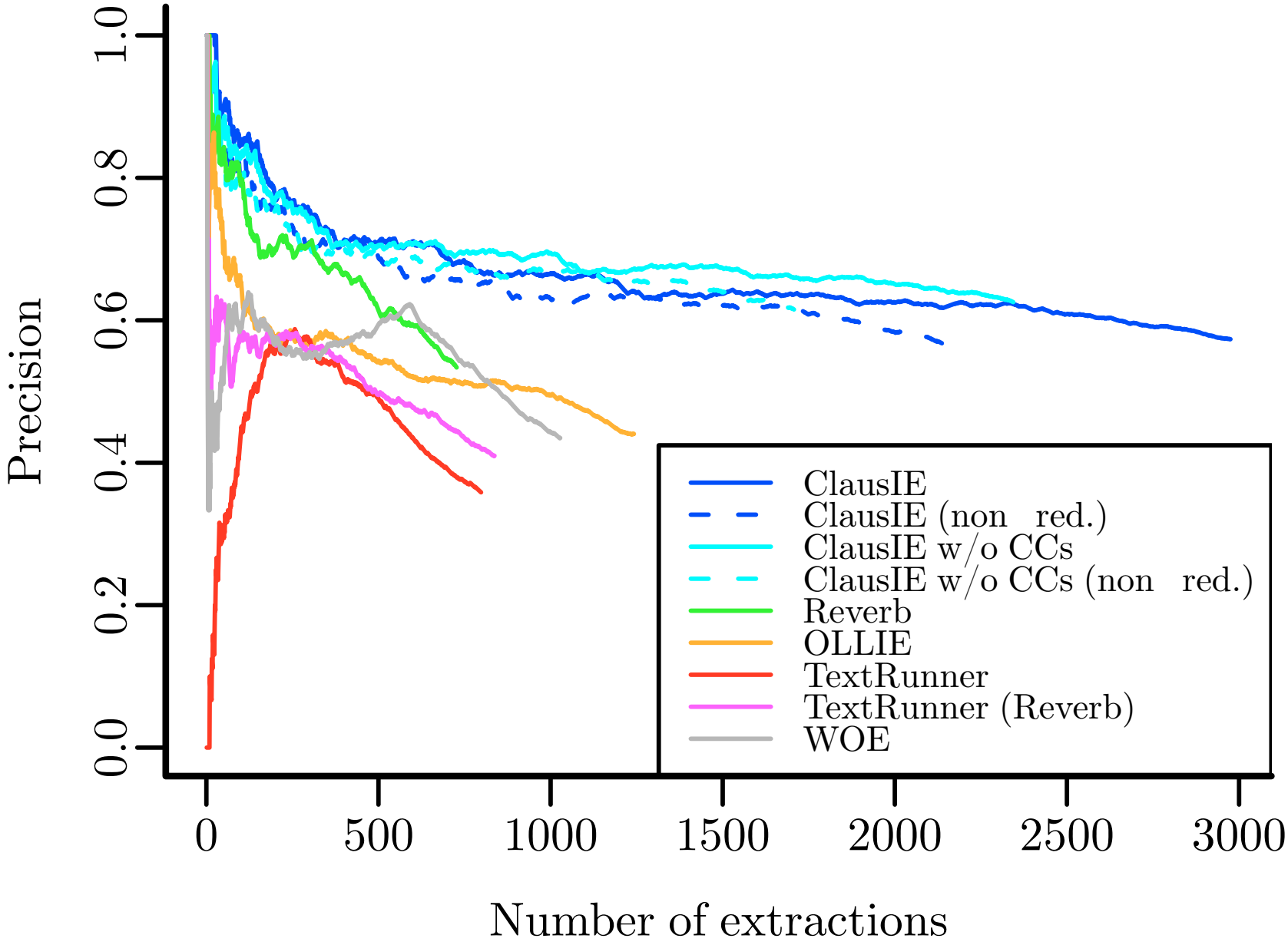
- Arbitrary form of relations
  - ⋆ *(Messi, plays football in, Barcelona)* or *(Messi, plays, football in Barcelona)*
- Propositions can be customized (e.g. triple, n-ary, etc)
  - ⋆ *(Messi, plays, football in Barcelona)* or *(Messi, plays, football, in Barcelona)*
- Arbitrary argument types (e.g. noun phrases, adjectives, etc)
  - ⋆ *(Gandhi, was, vegetarian)* or *(Gandhi, was, a vegetarian)* or *(Gandhi from Porbandar, was, a vegetarian)*
- Optional arguments can be used to generate new propositions
  - ⋆ *(Paul, takes, a shower, in the morning)* or *(Paul, takes, a shower)*

Outline

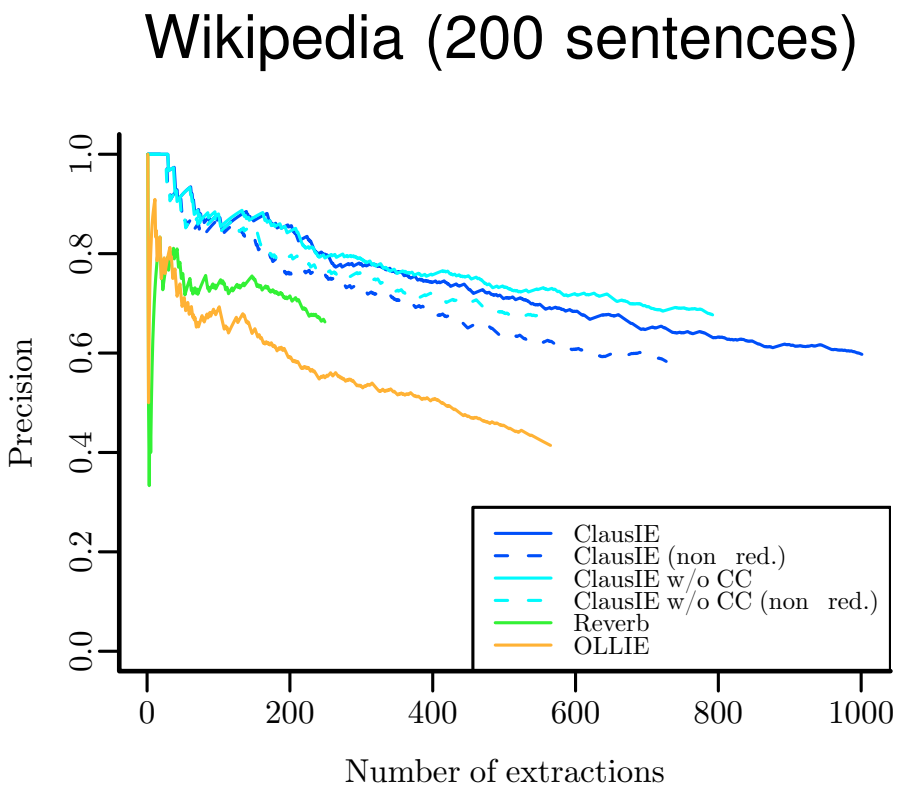
Evaluation

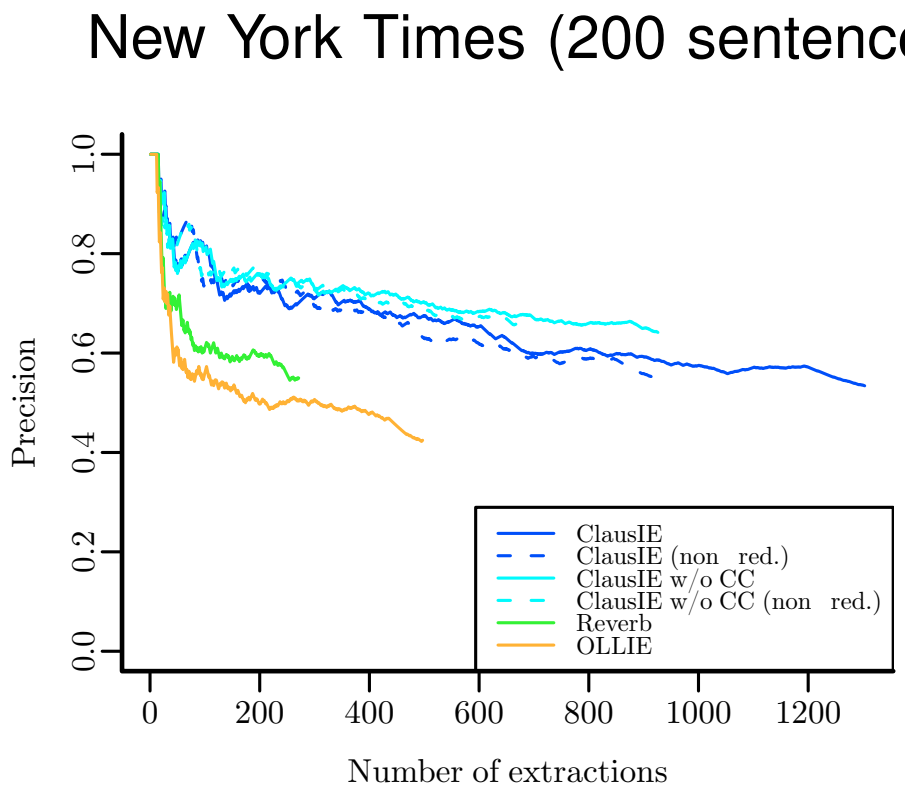
- 3 datasets
  - Reverb: Web, very noisy (500 sentences)
  - New York Times: Complex, written by experts (200 sentences)
  - Wikipedia: Simple, written by non-experts (200 sentences)
- 2 labelers, pessimistic approach.
- Agreement 57%-68%.
- High precision, high recall.

Results I: Reverb Sentences



Results II: Wikipedia and New York Times





Outline

Conclusions and Future Directions

Conclusions

- ClausIE is a principled approach for OIE
- Separates identification and representation
- No training needed
- DP based
- Publicly available `http://www.mpi-inf.mpg.de/departments/d5/software/clausie/`

Conclusions and Future Directions

Conclusions

- ClausIE is a principled approach for OIE
- Separates identification and representation
- No training needed
- DP based
- Publicly available `http://www.mpi-inf.mpg.de/departments/d5/software/clausie/`

Future Directions

- Build dictionaries
- Incorporate context analysis
- Post processing of arguments
- Input to other tasks: discourse processing, SRL, targeted IE, ontology learning, QA, ...

Conclusions and Future Directions

Conclusions

- ClausIE is a principled approach for OIE
- Separates identification and representation
- No training needed
- DP based
- Publicly available `http://www.mpi-inf.mpg.de/departments/d5/software/clausie/`

Future Directions

- Build dictionaries
- Incorporate context analysis
- Post processing of arguments
- Input to other tasks: discourse processing, SRL, targeted IE, ontology learning, QA, ...

Thank You!