Iterative joint extraction of entities, relationships and coreferences from text sources

Slavko Žitnik & Marko Bajec

University of Ljubljana
Faculty for computer and information science

15th May 2015

Agenda

About 293,000 results (0.19 seconds)

Ivan Cankar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Ivan_Cankar** ▾
Ivan Cankar was born in the Carniolan town of Vrhnika near Ljubljana. He was one of the many children of a poor artisan who emigrated to Bosnia shortly after ...
Biography - Work - Personality and world view - Influence

Ivan Cankar - Wikipedija, prosta enciklopedija
sl.wikipedia.org/wiki/**Ivan_Cankar** ▾ Translate this page
Ivan Cankar je za svoje pisateljsko delo uporabljal številne šifre in psevdonime. Ti so značilni predvsem za zgodnja leta njegovega ustvarjanja. Izmišljena ...
Življenje - Delo - Psevdonimi - Bibliografija

Ivan Cankar - Wikisource, the free online library
en.wikisource.org/wiki/Author:**Ivan_Cankar** ▾
Mar 19, 2014 - Author:**Ivan Cankar**. From Wikisource. Jump to: navigation, search.
←Author Index: Ca, **Ivan Cankar** (1876–1918) ...

Ivan Cankar – Wikivir
sl.wikisource.org/wiki/**Ivan_Cankar** ▾ Translate this page
May 28, 2014 - **Ivan Cankar**. Iz Wikivira, proste knjižnice besedil v javni lasti. Skoči na: navigacija, iskanje. **Ivan Cankar** (1876–1918). Glej tudi življenjepis ...

Ivan Cankar – največji mojster slovenske besede - Veliki ...
www.kam.si › Novice › Veliki Slovenci ▾ Translate this page
Ivan Cankar se je rodil na Vrhniki (na Klancu) 10. maja 1876 kot osmi otrok v propadajoči obrtniško – proletarski družini trškega krojača. Mladost je preživel na ...

[DOC] **Ivan Cankar** (1876 - 1918) - Dijaski.net
www.dijaski.net/.../slo_dob_**cankar_ivan**_hlapci_09__.. ▾ Translate this page
Ivan Cankar (1876 - 1918). Največji mojster slovenske besede je osrednja postava v moderni književnosti izvira iz revne družine z Vrhnike. Na Klancu ...

[PDF] ŽIVLJENJEPIS **Ivan Cankar** - Dijaski.net
www.dijaski.net/get/slo_rfk_**cankar_ivan**_07.pdf ▾ Translate this page
Ivan Cankar se je rodil 10. maja leta 1876 v kmečko družino na. Vrhniki. V družini je bilo osem otrok. Zapustil je družino. Ker je bil zelo nadarjen učenec in je ...

Web Images Books Videos More ▾ Search tools

About 293,000 results (0.28 seconds)

Ivan Cankar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Ivan_Cankar** ▾
Ivan Cankar was born in the Carniolan town of Vrhnika near Ljubljana. He was one of the many children of a poor artisan who emigrated to Bosnia shortly after ...
Biography - Work - Personality and world view - Influence

Ivan Cankar - Wikipedija, prosta enciklopedija
sl.wikipedia.org/wiki/**Ivan_Cankar** ▾ Translate this page
Ivan Cankar je za svoje pisateljsko delo uporabljal številne šifre in psevdonime. Ti so značilni predvsem za zgodnja leta njegovega ustvarjanja. Izmišljena ...
Življenje - Delo - Psevdonimi - Bibliografija

Ivan Cankar - Wikisource, the free online library
en.wikisource.org/wiki/Author:**Ivan_Cankar** ▾
Mar 19, 2014 - Author:**Ivan Cankar**. From Wikisource. Jump to: navigation, search. ←Author Index: Ca, **Ivan Cankar** (1876–1918) ...

Ivan Cankar – Wikivir
sl.wikisource.org/wiki/**Ivan_Cankar** ▾ Translate this page
May 28, 2014 - **Ivan Cankar**. Iz Wikivira, proste knjižnice besedil v javni lasti. Skoči na: navigacija, iskanje. **Ivan Cankar** (1876–1918). Glej tudi življenjepis ...

Ivan Cankar – največji mojster slovenske besede - Veliki ...
www.kam.si › Novice › Veliki Slovenci ▾ Translate this page
Ivan Cankar se je rodil na Vrhniki (na Klancu) 10. maja 1876 kot osmi otrok v propadajoči obrtniško – proletarski družini trškega krojača. Mladost je preživel na ...
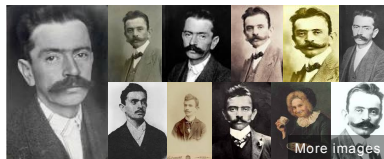
[DOC] **Ivan Cankar** (1876 - 1918) - Dijaski.net
www.dijaski.net/.../slo_dob_**cankar_ivan**_hlapci_09__... ▾ Translate this page
Ivan Cankar (1876 - 1918). Največji mojster slovenske besede in osrednja postava v moderni književnosti izvira iz revne družine z Vrhnike. Na Klancu ...

[PDF] ŽIVLJENJEPIS **Ivan Cankar** - Dijaski.net
www.dijaski.net/get/slo_rfk_**cankar_ivan**_07.pdf ▾ Translate this page
Ivan Cankar se je rodil 10. maja 1876 v kmečko družino na. Vrhniki. V družini je bilo osem otrok. Zapustil je družino. Ker je bil zelo nadarjen učenec in je ...

Ivan Cankar

Writer

Ivan Cankar was a Slovene writer, playwright, essayist, poet and political activist. Together with Oton Župančič, Dragotin Kette, and Josip Murn, he is considered as the beginner of modernism in Slovene literature. Wikipedia

Born: May 10, 1876, Vrhnika
Died: December 11, 1918, Ljubljana
Education: University of Vienna

People also search for   View 15+ more

France Prešeren Oton Župančič Dragotin Kette Srečko Kosovel Josip Murn

Feedback

Ivan Cankar

From Wikipedia, the free encyclopedia

Ivan Cankar (🔊 pronunciation (help·info)) (10 May 1876 – 11 December 1918) was a Slovene writer, playwright, essayist, poet and political activist. Together with Oton Župančič, Dragotin Kette, and Josip Murn, he is considered as the beginner of modernism in Slovene literature. He is regarded as the greatest writer in the Slovene language, and has sometimes been compared to Franz Kafka and James Joyce.[1]

Contents [hide]
1 Biography
2 Work
3 Personality and world view
4 Influence
5 Legacy
6 Bibliography
7 Notes
8 References
9 Further reading
10 External links

Biography [edit]

Cankar's birth house in Vrhnika

Ivan Cankar was born in the Carniolan town of Vrhnika near Ljubljana. He was one of the many children of a poor artisan who emigrated to Bosnia shortly after Ivan's birth.[2] He was raised by his mother, Neža Cankar née Pivk, with whom he established a close, but ambivalent relationship.[3][4] The figure of a self-sacrificing and submissively repressive mother would later become one of the most recognizable features of Cankar's prose.[5][6] After finishing grammar school in his hometown, he studied at the Technical High School (*Realka*) in Ljubljana.

During this period, he started writing literature, mostly poetry, under the influence of Romantic and post-Romantic poets such as France Prešeren, Heinrich Heine, Simon Jenko and Simon Gregorčič.[7] In 1893, he discovered the epic poetry of Anton Aškerc, which had a huge influence on the development of his style and ideals. Under Aškerc's influence, Cankar rejected the sentimental post-Romantic poetry and embraced literary realism and national liberalism.[8]

In 1896, he enrolled at the University of Vienna, where he studied engineering, but later switched to Slavic philology.[2] In Vienna, he soon started to lead a bohemian lifestyle. He came under the influence of contemporary European literature, especially decadentism, symbolism and naturalism. He became friends with Fran Govekar, a young Slovene writer and intellectual living in Vienna, who introduced him to positivism and naturalism.[9] Between 1897 and 1899, Cankar's core

	Ivan Cankar
Born	10 May 1876 Vrhnika, Duchy of Carniola, Austria-Hungary (now in Slovenia)
Died	11 December 1918 (aged 42) Ljubljana, Kingdom of Serbs, Croats and Slovenes (now the capital of Slovenia)
Occupation	Writer, essayist, playwright, poet, political activist
Genres	plays, short stories, short novels, essays
Literary movement	Symbolism, Modernism

WIKIPEDIA
The Free Encyclopedia

Ivan Cankar

DBPedia entry

dbpprop:birthDate	• 1876-05-10 (xsd:date)
dbpprop:birthPlace	• Vrhnika, Duchy of Carniola, Austria-Hungary
dbpprop:dateOfBirth	• 10 (xsd:integer)
dbpprop:dateOfDeath	• 11 (xsd:integer)
dbpprop:deathDate	• 1918-12-11 (xsd:date)
dbpprop:deathPlace	• dbpedia:Kingdom_of_Yugoslavia • dbpedia:Ljubljana
dbpprop:genre	• plays, short stories, short novels, essays
dbpprop:id	• Ivan_Cankar_
dbpprop:influenced	• most of modern Slovenian literature, Fulvio Tomizza
dbpprop:influences	• dbpedia:Friedrich_Nietzsche • dbpedia:Leo_Tolstoy • dbpedia:Ralph_Waldo_Emerson • dbpedia:Fyodor_Dostoevsky
dbpprop:movement	• dbpedia:Modernism • dbpedia:Symbolism_(arts)

Ivan Cankar

orn	10 May 1876 Vrhnika, Duchy of Carniola, Austria-Hungary (now in Slovenia)
ied	11 December 1918 (aged 42) Ljubljana, Kingdom of Serbs, Croats and Slovenes (now the capital of Slovenia)
ccupation	Writer, essayist, playwright, poet, political activist
Genres	plays, short stories, short novels, essays
Literary movement	Symbolism, Modernism

the most recognizable features of Cankar's prose.[5][6] After finishing grammar school in his hometown, he studied at the Technical High School (*Realka*) in Ljubljana.

Cankar's birth house in Vrhnika

During this period, he started writing literature, mostly poetry, under the influence of Romantic and post-Romantic poets such as France Prešeren, Heinrich Heine, Simon Jenko and Simon Gregorčič.[7] In 1893, he discovered the epic poetry of Anton Aškerc, which had a huge influence on the development of his style and ideals. Under Aškerc's influence, Cankar rejected the sentimental post-Romantic poetry and embraced literary realism and national liberalism.[8]

In 1896, he enrolled at the University of Vienna, where he studied engineering, but later switched to Slavic philology.[2] In Vienna, he soon started to lead a bohemian lifestyle. He came under the influence of contemporary European literature, especially decadentism, symbolism and naturalism. He became friends with Fran Govekar, a young Slovene writer and intellectual living in Vienna, who introduced him to positivism and naturalism.[9] Between 1897 and 1899, Cankar's core

WIKIPEDIA
The Free Encyclopedia

Ivan Cankar

DBPedia entry

dbpprop:birthDate	• 1876-05-10 (xsd:date)
dbpprop:birthPlace	• Vrhnika, Duchy of Carniola, Austria-Hungary
dbpprop:dateOfBirth	• 10 (xsd:integer)
dbpprop:dateOfDeath	• 11 (xsd:integer)
dbpprop:deathDate	• 1918-12-11 (xsd:date)
dbpprop:deathPlace	dbpedia:Kingdom_of_Yugoslavia
	dbpedia:Ljubljana
dbpprop:genre	• plays, short stories, short novels, essays
dbpprop:id	• Ivan_Cankar_
dbpprop:influenced	• most of modern Slovenian literature, Fulvio Tomizza
dbpprop:influences	dbpedia:Friedrich_Nietzsche

Ivan Cankar

Text excerpt

become one of the most recognizable features of Cankar's prose.
[5][6] After finishing grammar school in his hometown, he studied at
the Technical High School (*Realka*) in Ljubljana.

orn   10 May 1876
           Vrhnika, Duchy of Carniola,
           (now in Slovenia)
           8 (aged 42)
           n of Serbs,
           es (now the
)
           aywright, poet,
           , short novels,
           nism

influence of Romantic and post-Romantic poets such as France Prešeren,
Heinrich Heine, Simon Jenko and Simon Gregorčič.[7] In 1893, he discovered the epic poetry of Anton Aškerc, which had a
huge influence on the development of his style and ideals. Under Aškerc's influence, Cankar rejected the sentimental post-Romantic poetry and embraced literary
realism and national liberalism.[8]

In 1896, he enrolled at the University of Vienna, where he studied engineering, but later switched to Slavic philology.[2] In Vienna, he soon started to lead a
bohemian lifestyle. He came under the influence of contemporary European literature, especially decadentism, symbolism and naturalism. He became friends with
Fran Govekar, a young Slovene writer and intellectual living in Vienna, who introduced him to positivism and naturalism.[9] Between 1897 and 1899, Cankar's core

Zoogle - Traditional

http://tradicionalni-iskalnik.si

Ivan Cankar | Išči |

666 najdenih zadetkov:

Ivan Cankar - Wikipedia, prosta enciklopedija ★★★★★
Ivan Cankar se je rodil v hiši Na klancu 141, kot eden od dvanajstih otrok
obrtniško-proletarske družine. Leta 1882 se je vpisal v osnovno …

Ivan Cankar – največji mojster slovenske besede ★★★★☆
Ivan Cankar se je rodil na Vrhniki (na Klancu) 10. maja 1876 kot osmi
otrok v propadajoči obrtniško – proletarski družini trškega krojača …

Cankarjeva smrt je bila političen umor ★★★★☆
Cankarjeva smrt je bila politčen umor. Pisatelj, politik in ljudski tribun
Ivan Cankar Ivan Cankar velja za največjega slovenskega pisatelja in …

[PDF] Ivan Cankar (1876 - 1918) ★★★☆☆
Ivan Cankar (1876 - 1918). Največji mojster slovenske besede in
osrednja postava v moderni književnosti izvira iz revne družine …

ŽIVLJENJEPIS Ivan Cankar ★★☆☆☆
Ivan Cankar se je rodil 10. maja leta 1876 v kmečko družino na. Vrhniki.
V družini je bilo osem otrok. Zapustil je družino. Ker je bil zelo …

Ivan Cankar memorial house ★☆☆☆☆
Ivan Cankar memorial house. **Ivan Cankar** (1876 – 1918) is considered
to be Slovenia's most important writer. The original house …
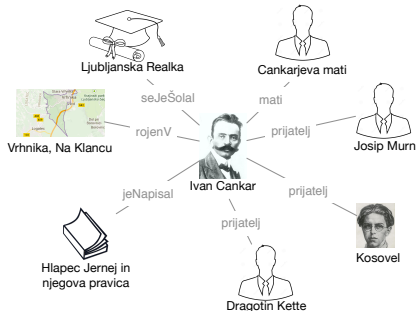
1 | 2 | 3 | 4 | 5 | … | Zadnja

Zoogle - Semantic

http://semantični-iskalnik.si

Ivan Cankar | Išči |

Informacije ekstrahirane iz 25 zadetkov, najdeni 666:

Ljubljanska Realka

Cankarjeva mati

seJeŠolal   mati

Vrhnika, Na Klancu

rojenV   prijatelj

Josip Murn

Ivan Cankar

jeNapisal   prijatelj

Kosovel

Hlapec Jernej in
njegova pravica

prijatelj

Dragotin Kette

Ivan Cankar - Wikipedia, prosta enciklopedija ★★★★★
Ivan Cankar se je rodil v hiši Na klancu 141, kot eden od dvanajstih otrok
obrtniško-proletarske družine. Leta 1882 se je vpisal v osnovno …

Ivan Cankar – največji mojster slovenske besede ★★★★☆
Ivan Cankar se je rodil na Vrhniki (na Klancu) 10. maja 1876 kot osmi
otrok v propadajoči obrtniško – proletarski družini trškega krojača …
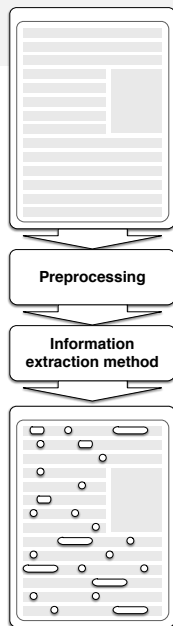
Agenda

Definition

Information extraction

- **type of** information retrieval
- **goal** to automatically extract structured data from (half-)structured data sources
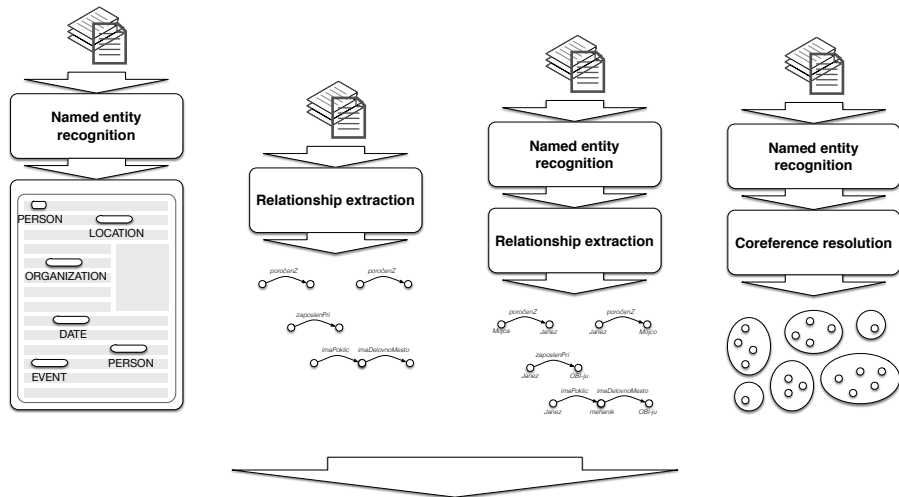
Subtasks

- named entity recognition
- relationship extraction
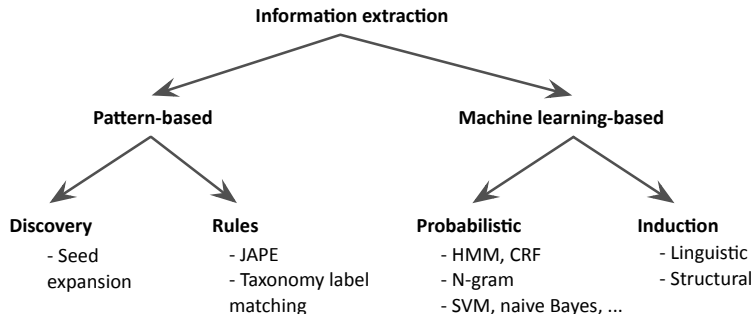- coreference resolution

Preprocessing

	John	is	married	to	Jena	.	They	work	at	OBI	.
Sentence detection	John	is	married	to	Jena	.	They	work	at	OBI	.
Tokenization	John	is	married	to	Jena	.	They	work	at	OBI	.
Lemmatization	John	be	marry	to	Jena	.	They	work	at	OBI	.
Part-of-speech tagging	NNP	VBZ	VBN	TO	NNP	.	PRP	VBP	IN	NNP	.
Dependency parsing	John	is	married	to	Jena	.	They	work	at	OBI	.

Proposed systems

Iterative and joint information extraction using an ontology

General approaches

Information extraction

Pattern-based

Machine learning-based

Discovery
- Seed
expansion

Rules
- JAPE
- Taxonomy label
matching

Probabilistic
- HMM, CRF
- N-gram
- SVM, naive Bayes, ...

Induction
- Linguistic
- Structural

Algorithm selection
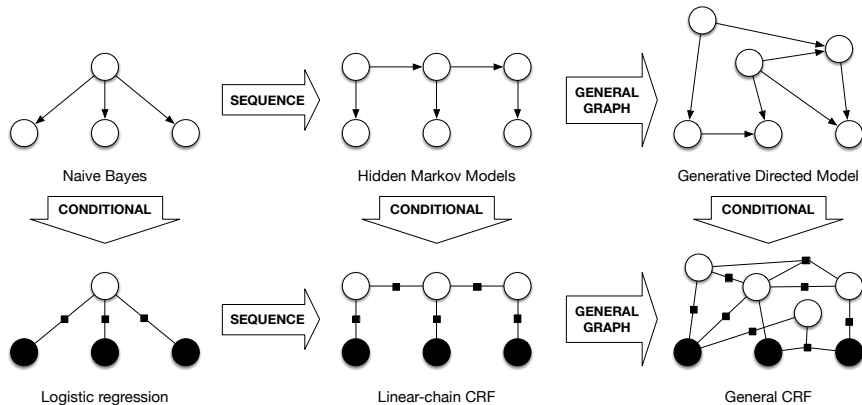
Conditional random fields (CRF)

$$P(\overline{y}|\overline{x}) = \frac{1}{Z(\overline{x})} \prod_{i=1}^{n} \left(\exp \left(\sum_{j=1}^{m_1} \lambda_j f_j(y_i, x_i) \right) \exp \left(\sum_{j=1}^{m_2} \lambda_j f_j(y_i, y_{i-1}, x_i) \right) \right)$$

$$f_1(x_i, y_i, y_{i-1}) = \begin{cases} 1, & \text{if } x_{i-1} = \text{Mr.}, y_i = PER, y_{i-1} = O \\ 0, & \text{otherwise.} \end{cases}$$

- discriminative model
- probabilistic label distributions
- complex interdependent sequences
- lots of features

Agenda

About coreference resolution

John is married to Jena . He is a mechanic at OBI and she also works there .

It is a DIY market .

About coreference resolution

John is married to **Jena**. **He** is a mechanic at **OBI** and **she** also works **there**.

It is a **DIY market**.

About coreference resolution

John is married to **Jena** . **He** is a mechanic at **OBI** and **she** also works **there** .

It is a **DIY market** .

About coreference resolution

Coreference resolution approaches:

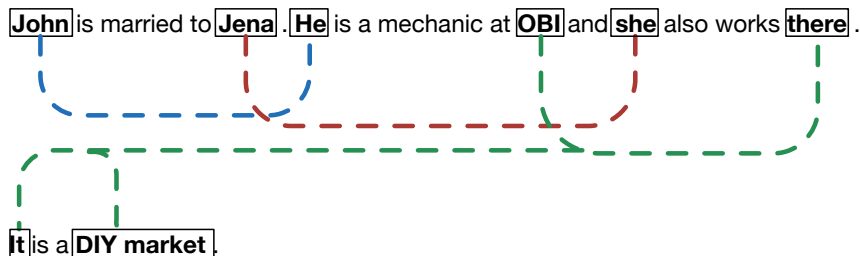|	UNSUPERVISED	SUPERVISED
MENTION SEQUENCES	[3], [4], [6]	**SkipCor**
MENTION PAIRS	[11], [16], [8]	[7], [15], [5], [20], [13], [17], [9], [14], [18]

SkipCor – coreference resolution method

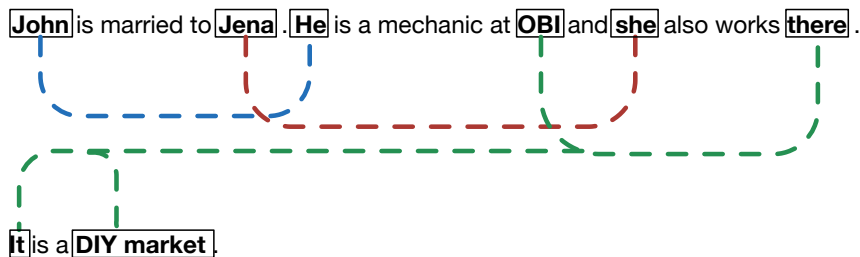**John** is married to **Jena** . **He** is a mechanic at **OBI** and **she** also works **there** .

It is a **DIY market** .

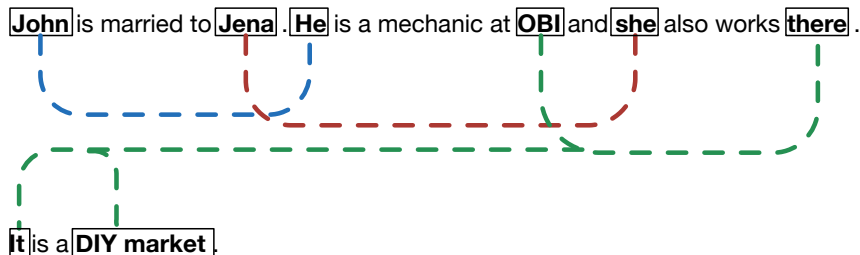$$\overline{x} = [\text{JOHN}, \text{JENA}, \text{HE}, \text{OBI}, \text{SHE}, \text{THERE}, \text{IT}, \text{DIY MARKET}].$$

O	O	O	O	O	O	C	C
John	Jena	He	OBI	she	there	It	DIY Market
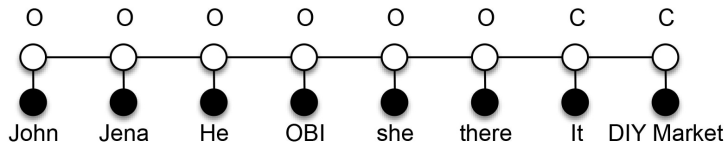
SkipCor – coreference resolution method

Distribution of distances between two consecutive coreferent mentions – SemEval-2010 data set

Mention distance between two consecutive coreferent mentions

SkipCor – coreference resolution method

$\overline{x} = [\text{John}, \text{Jena}, \text{He}, \text{OBI}, \text{she}, \text{there}, \text{It}, \text{DIY market}].$

SkipCor – coreference resolution method

$\overline{x} = [\text{JOHN}, \text{JENA}, \text{HE}, \text{OBI}, \text{SHE}, \text{THERE}, \text{IT}, \text{DIY MARKET}].$
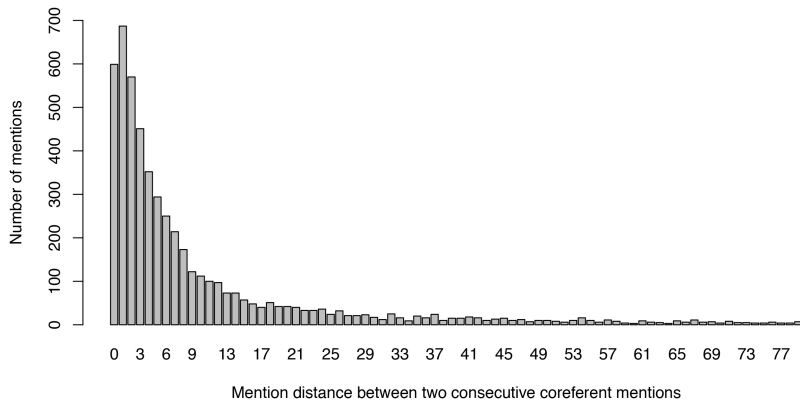
SkipCor – coreference resolution method

$\overline{x} = [\text{JOHN}, \text{JENA}, \text{HE}, \text{OBI}, \text{SHE}, \text{THERE}, \text{IT}, \text{DIY MARKET}].$

SkipCor – coreference resolution method

Input documents

Entity clustering

Document mention detection

Skip-mention sequences

Model training

Mention labeling

Model 0

Model 1

Model 2

...

Model s

0

1

2

s ...
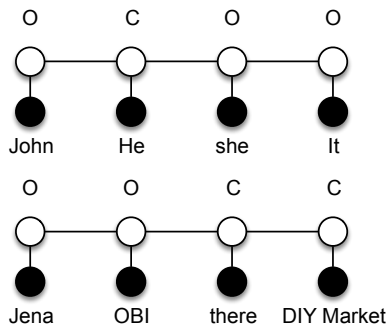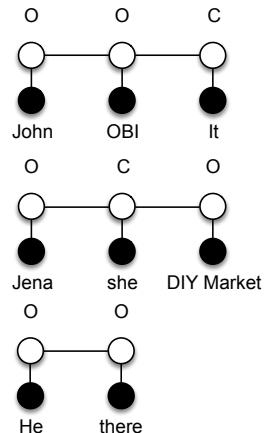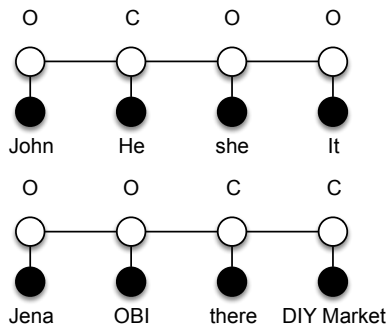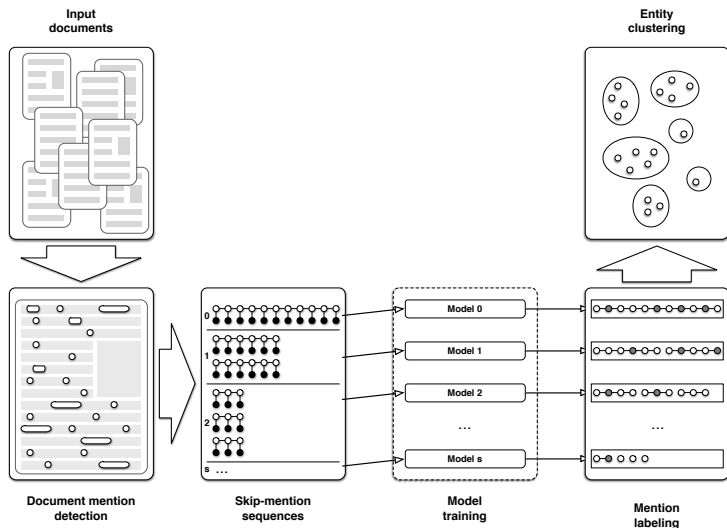
SkipCor – results

	MUC			BCubed			CEAF		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Sistem	**SemEval2010**								
SkipCor	68.8	30.1	41.8	94.8	**80.8**	**87.3**	74.0	78.5	**76.2**
SkipCorZero	67.0	3.6	6.8	**99.6**	75.1	85.7	73.0	73.1	73.1
SkipCorPair	**76.7**	35.6	48.7	97.1	79.0	87.1	72.7	**79.4**	75.9
RelaxCor [19]	72.4	21.9	33.7	97.0	74.8	84.5	**75.6**	75.6	75.6
SUCRE [10]	54.9	**68.1**	**60.8**	78.5	86.7	82.4	74.3	74.3	74.3
TANL-1 [1]	24.4	23.7	24.0	72.1	74.6	73.4	61.4	75.0	67.6
UBIU [22]	25.5	17.2	20.5	83.5	67.8	74.8	68.2	63.4	65.3

Results of the proposed SkipCor system, baseline approaches and other systems against SemEval-2010 data set. Metrics are MUC [21], BCubed [2] in CEAF [12].

SkipCor – results examples, CoNLL 2012

Target entities	Identified entities
Entity(Milosevic 's successor, Vojislav Kostunica, a critic of the War Crimes Tribunal at The Hague, He, Kostunica),	Entity(Milosevic 's, Milosevic 's successor, Vojislav Kostunica, a critic of the War Crimes Tribunal at The Hague, Milosevic, He, Milosevic, Kostunica, former President Slobodan Milosevic),
Entity(Carla Ponte, The chief U.N. war crimes prosecutor),	Entity(Carla Ponte),
Entity(Milosevic 's, Milosevic, Milosevic, former President Slobodan Milosevic)	Entity(The chief U.N. war crimes prosecutor)

SkipCor – results examples, CoNLL 2012

Target entities	Identified entities
Entity(Israel and the Palestinians, The two sides),	Entity(Israel and the Palestinians),
Entity(Israel, Israel, Israel)	Entity(The two sides), Entity(Israel, Israel, Israel)

SkipCor – results examples, CoNLL 2012

Target entities	Identified entities
Entity(shot, shot, an assassination attempt),	Entity(an assassination attempt),
Entity(Belgrade, Belgrade),	Entity(Belgrade, Belgrade),
Entity(The Serbian Prime Minister, Zoran Djindjic, Zoran Djindjic, the Prime Minister, his, he, his, he)	Entity(shot, The Serbian Prime Minister, Zoran Djindjic, Zoran Djindjic, the Prime Minister, his, shot, he, his, he)

SkipCor – results examples, CoNLL 2012

Target entities	Identified entities
Entity(Northern Ireland, Northern Ireland),	Entity(Northern Ireland, Northern Ireland),
Entity(President Clinton, he, Mr. Clinton 's, Mr. Clinton, He, he)	Entity(President Clinton, he, Mr. Clinton 's, Mr. Clinton, He, he)

Error types at coreference resolution - CoNLL2012

Tabela: Sums of specific error types for SkipCor on CoNLL2012-BN-Test data set.

Error	Occurences
Span error	3
Missing entity	124
Extra entity	0
Missing mention	255
Extra mention	3
Divided entity	399
Conflated entities	568

Agenda

About relationship extraction

relationship

subject   *object*

About relationship extraction

John is married to **Jena** . **He** is a **mechanic** at **OBI** and **she** also works **there** .

It is a **DIY market** .

Related approaches

- relationship mention extraction
- binary classification
- unsupervised extraction

About relationship extraction

CRF-based relationship extraction method

Relationship extraction – BioNLP 2013

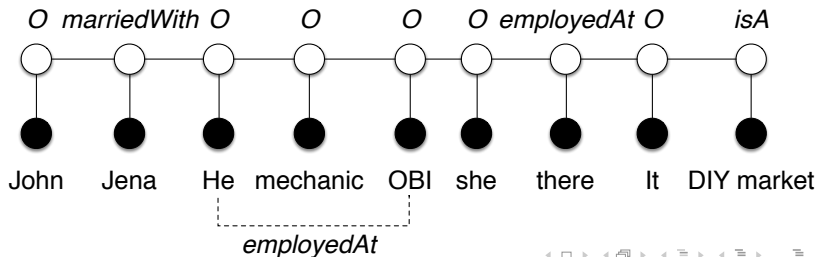| Participant	S	D	I	M	SER
U. of Ljubljana	**8**	50	6	30	**0.73**
K. U. Leuven	15	53	5	20	0.83
TEES-2.1	9	59	8	20	0.86
IRISA-TexMex	27	**25**	28	**36**	0.91
EVEX	10	67	**4**	11	0.92

BioNLP 2013 GRN challenge official results. The table shows the number of substitutions (S), deletions (D), insertions (I), matches (M) and "slot error rate" (SER) score.

- usage of manual rules – regular expressions
- fine tuned to 0.67 SER (feature functions, additional processing)

Relationship extraction – BioNLP 2013 result

Agenda

Slavko Žitnik & Marko Bajec (FRI)   Information Extraction   15th May 2015   30 / 47

About named entity recognition

Person   Person   Position   Organization
John is married to **Jena** . **He** is a **mechanic** at **OBI** and **she** also works **there** .

Organizacija
It is a **DIY market** .

- named entities
- also "entity extraction"
- mostly sequence labeling, also multinomial/binomial classification
- IOB notation (e.g.: B-ORG I-ORG)

About named entity recognition

Person   Person   Position   Organization
John is married to **Jena** . **He** is a **mechanic** at **OBI** and **she** also works **there** .
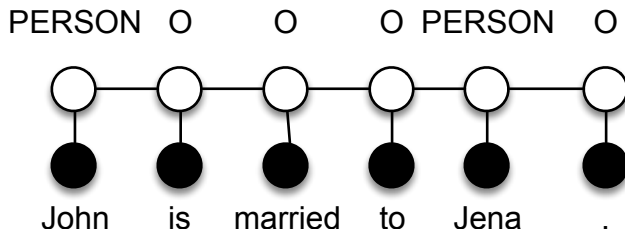
Organizacija
It is a **DIY market** .

PERSON   O   O   O   PERSON   O

John   is   married   to   Jena   .

Named entity recognition – CHEMDNER 2013
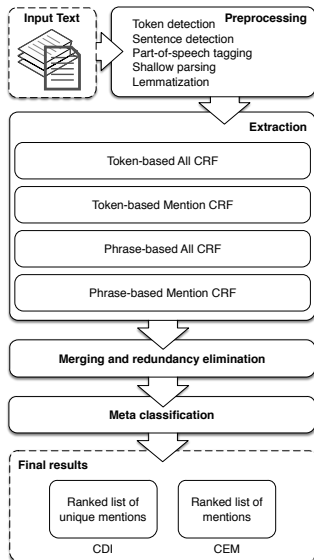
		Micro-average		
Model	Pred.	P	R	F
CDI Results				
TS (run1)	12381	**0.83**	0.75	**0.79**
TM (run2)	12783	0.81	0.76	0.78
TSM (run3)	**13172**	0.80	**0.77**	0.79
CEM Results				
TS (run1)	20438	**0.87**	0.70	0.77
TM (run2)	21109	0.85	0.71	0.77
TSM (run3)	**21562**	0.85	**0.72**	**0.78**

Official CHEMDNER 2013 results. The table shows the number of extracted chemical entities and drugs (Pred.), precision (P), recall (R) and F_1 score. The models were trained against training data set and development data set and evaluated against test set. We used token-based (TS), token-based mention (TM) and a combination of both (TSM).

- we later improved results to 84.6 F_1 (CDI) and 83.2 F_1 (CEM)
- 4% lower than the best performer

Agenda

1. Motivation

2. Information extraction
   - Definition
   - Related work
   - Systems classification
   - Conditional random fields (CRF)

3. Information extraction
   - Coreference resolution
   - Relationship extraction
   - Named entity recognition
   - Iterative and joint information extraction

4. Further work

Slavko Žitnik & Marko Bajec (FRI)   Information Extraction   15th May 2015   34 / 47

About iterative and joint information extraction
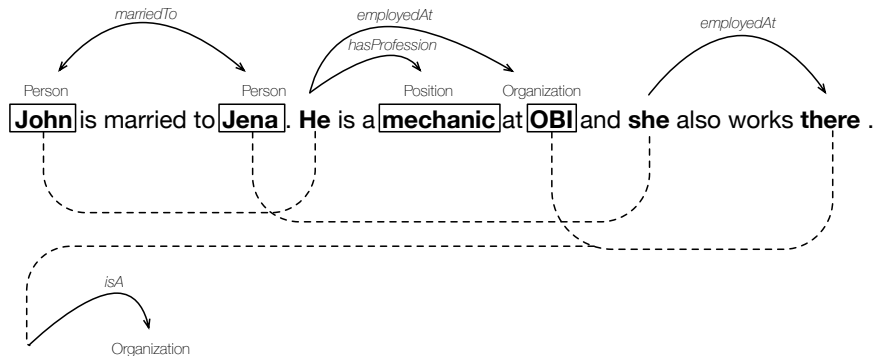
It is a **DIY market** .

- linear-chain CRF for all tasks
- ontology as a schema, rules and as a lexicon
- improvement of specific IE tasks because of the influence of others
- additional feature functions

Iterative information extractions system – IOBIE
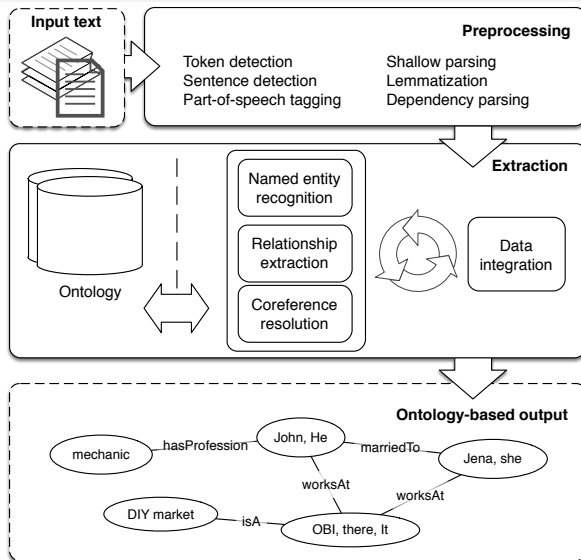
Results

Named entity recognition

Model	Error (%)	CA	MaP	MaR	MaF	MiF
Independent	–	97.0	54.0	30.4	38.9	90.8
Second iteration	10.3	97.2	55.0	33.2	41.4	91.6
Third iteration	**15.0**	**97.8**	**55.2**	**33.5**	**41.7**	**92.2**
Fourth iteration	**15.0**	**97.8**	**55.2**	**33.5**	**41.7**	**92.2**
Fifth iteration	**15.0**	**97.8**	**55.2**	**33.5**	**41.7**	**92.2**

Tabela: Named entity recognition results. Shown measures are error reduction in % (Error), classification accuracy (CA), macro-averaged precision (MaP), macro-averaged recall (MaR), macro-averaged F-score (MaF) and micro-averaged F-score (MiF).

Results

Relationship extraction

Model	Error (%)	P	R	F
Independent	–	54.3	55.2	54.7
Second iteration	0.8	**55.1**	55.6	**55.3**
Third iteration	**2.4**	54.8	55.6	55.2
Fourth iteration	**2.4**	55.0	55.4	55.2
Fifth iteration	2.0	54.2	**55.7**	54.9

Tabela: Relationship extraction results. Shown measures are error in % (Error), precision (P), recall (R) and F score.

Results

Coreference resolution

Model	MUC	BCubed	CEAF
Independent	73.2	73.9	49.8
Second iteration	73.8	73.5	50.0
Third iteration	74.0	**74.1**	**52.9**
Fourth iteration	**74.3**	73.8	52.8
Fifth iteration	**74.3**	73.8	52.8

Tabela: Coreference resolution results. Used measures are MUC [21], BCubed [2] and CEAF [12].

Further work

- Interdependencies between tasks.
- "End-to-end" information extraction systems evaluation.
- Models weighting with respect to skip mention number.

Thanks!

@szitnik
slavko.zitnik@fri.uni-lj.si

Literatura I

Giuseppe Attardi, Stefano Dei Rossi, and Maria Simi.
TANL-1: coreference resolution by parse analysis and similarity clustering.
In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 108—-111, Pennsylvania, 2010.
Association for Computational Linguistics.

Amit Bagga and Breck Baldwin.
Algorithms for scoring coreference chains.
In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 1–7, 1998.

Cosmin Bejan, Matthew Titsworth, Andrew Hickl, and Sanda Harabagiu.
Nonparametric bayesian models for unsupervised event coreference resolution.
In *Advances in Neural Information Processing Systems*, pages 73—-81, 2009.

Cosmin Adrian Bejan and Sanda Harabagiu.
Unsupervised event coreference resolution with rich linguistic features.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412—-1422, Stroudsburg, 2010. Association for Computational Linguistics.

Eric Bengtson and Dan Roth.
Understanding the value of features for coreference resolution.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294—-303, Pennsylvania, 2008. Association for Computational Linguistics.

Eugene Charniak.
Unsupervised learning of name structure from coreference data.
In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–7, Stroudsburg, 2001. Association for Computational Linguistics.

Literatura II

Aron Culotta, Michael Wick, Robert Hall, and Andrew McCallum.
First-order probabilistic models for coreference resolution.
In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 81—88, 2007.

Aria Haghighi and Dan Klein.
Simple coreference resolution with rich syntactic and semantic features.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 3, pages 1152—1161, Pennsylvania, 2009. Association for Computational Linguistics.

Shujian Huang, Yabing Zhang, Junsheng Zhou, and Jiajun Chen.
Coreference resolution using markov logic networks.
Advances in Computational Linguistics, 41:157–168, 2009.

Hamidreza Kobdani and Hinrich Schütze.
SUCRE: a modular system for coreference resolution.
In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 92—95, 2010.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihail Surdeanu, and Dan Jurafsky.
Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task.
In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28—34, Pennsylvania, 2011. Association for Computational Linguistics.

Xiaoqiang Luo.
On coreference resolution performance metrics.
In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25—32, Pennsylvania, 2005. Association for Computational Linguistics.

Literatura III

Xiaoqiang Luo, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla, and Salim Roukos.
A mention-synchronous coreference resolution algorithm based on the bell tree.
In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 136–143, Pennsylvania, 2004. Association for Computational Linguistics.

Andrew McCallum and Ben Wellner.
Conditional models of identity uncertainty with application to noun coreference.
In *Neural Information Processing Systems*, pages 1–8, 2004.

Vincent Ng and Claire Cardie.
Improving machine learning approaches to coreference resolution.
In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104—111, Philadelphia, 2002. Association for Computational Linguistics.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning.
A multi-pass sieve for coreference resolution.
In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492—501, Pennsylvania, 2010. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng.
Supervised models for coreference resolution.
In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 2, pages 968—977, 2009.

Literatura IV

R. Vijay Sundar Ram and Sobha Lalitha Devi.
Coreference resolution using tree CRFs.
In *Proceedings of Conference on Intelligent Text Processing and Computational Linguistics*, pages 285—-296, Heidelberg, 2012. Springer.

Emili Sapena, Lluís Padró, and Jordi Turmo.
RelaxCor: a global relaxation labeling approach to coreference resolution.
In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 88—-91, 2010.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim.
A machine learning approach to coreference resolution of noun phrases.
Computational linguistics, 27(4):521—-544, 2001.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman.
A model-theoretic coreference scoring scheme.
In *Proceedings of the 6th conference on Message understanding*, pages 45—-52, Pennsylvania, 1995. Association for Computational Linguistics.

Desislava Zhekova and Sandra Kübler.
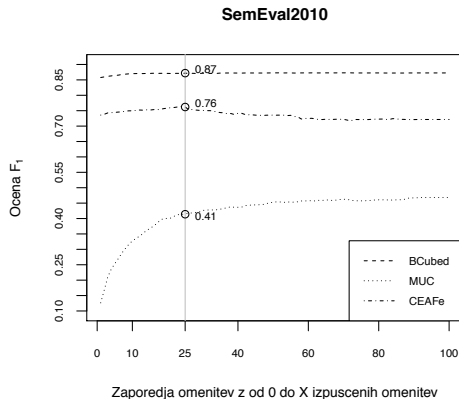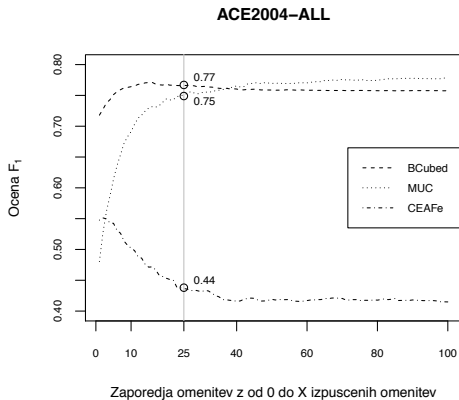UBIU: a language-independent system for coreference resolution.
In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96—-99, 2010.

SkipCor – expected results

	Model				
Podatki	**A-BN**	**A-NW**	**C-BN**	**C-NW**	**SE**
A-BN	*74, 78, 54*	**72**, 77, 39	65, 70, 28	64, 69, 29	42, **71**, 49
A-NW	**72**, 73, 42	*73, 75, 58*	60, 64, 27	59, 69, 29	42, 67, **50**
C-BN	33, 56, 37	40, 58, 39	*68, 70, 43*	**65**, 70, 27	**57**, 64, 31
C-NW	39, 57, 39	41, 59, 41	**67**, 66, 28	*68, 70, 48*	56, 64, 32
SE	19, **82**, **70**	23, **85**, **74**	39, 76, 40	39, **77**, **33**	*42, 87, 76*

Tabela: Coreference resolution results comparison on ACE2004 (i.e., A), CoNLL2012 (i.e., C) and SemEval2010 newswire (i.e., NW) and broadcast news (i.e., BN) datasets. Each column represents a model trained on a specific dataset, while each row represents a dataset. Values represent F-scores of MUC [21], BCubed [2] and CEAF [12], respectively.

SkipCor – coreference resolution method

ACE2004–ALL

SemEval2010

Zaporedja omenitev z od 0 do X izpuscenih omenitev

Zaporedja omenitev z od 0 do X izpuscenih omenitev

Coreference resolution – error type classification

	Operation(s)	Error	System	Gold
i)	Alter Span	Span error	*Gorbachev*	*Soviet leader Gorbachev*
ii)	Multiple Introduces and Merges	Missing Entity	- -	*the pills* *the tranquilizing pills*
iii)	Multiple Splits and Removes	Extra Entity	*human rights* *Human Rights*	- -
iv)	Introduce and Merge	Missing Mention	*the Arab region* *the region* -	*the Arab region* *the region* *it*
v)	Split and Remove	Extra Mention	*her story* *this* *it*	*her story* *this* -
vi)	Merge	Divided Entity	*Iraq$_1$* *this nation$_2$* *the nation$_2$* *its$_1$*	*Iraq$_1$* *this nation$_1$* *the nation$_1$* *its$_1$*
vii)	Split	Conflated Entities	*Mohammed Rashid$_1$* *the Rashid case$_1$* *Rashid$_1$* *the case$_1$*	*Mohammed Rashid$_1$* *the Rashid case$_2$* *Rashid$_1$* *the case$_2$*

Coreference resolution scoring metrics

MUC The key idea in developing the MUC measure [21] was to give an intuitive explanation of the results for coreference resolution systems. It is a link-based metric (it focuses on pairs of mentions) and is the most widely used. MUC counts false positives by computing the minimum number of links that need to be added in order to connect all the mentions referring to an entity. Recall, on the other hand, measures how many of the links must be removed so that no two mentions referring to different entities are connected in the graph. Thus, the MUC metric gives better scores to systems having more mentions per entity, while it also ignores entities with only one mention (singleton entities).

Coreference resolution scoring metrics

BCubed The BCubed metric [2] tries to address the shortcomings of MUC by focusing on mentions, and measures the overlap of the predicted and true clusters by computing the values of recall and precision for each mention. If k is the key entity and r the response entity containing the mention m, the recall for mention m is calculated as $\frac{|k \cap r|}{|k|}$, and the precision for the same mention, as $\frac{|k \cap r|}{|r|}$. This score has the advantage of measuring the impact of singleton entities, and gives more weight to the splitting or merging of larger entities.

Coreference resolution scoring metrics

CEAF The goal of the CEAF metric [12] is to achieve better interpretability. The result therefore reflects the percentage of correctly recognized entities. We use entity-based metric (in contrast to a mention-based version) that tries to match the response entity with at most one key entity. For CEAF, the value of recall is $\frac{\text{total similarity}}{|k|}$, while precision is $\frac{\text{total similarity}}{|r|}$.