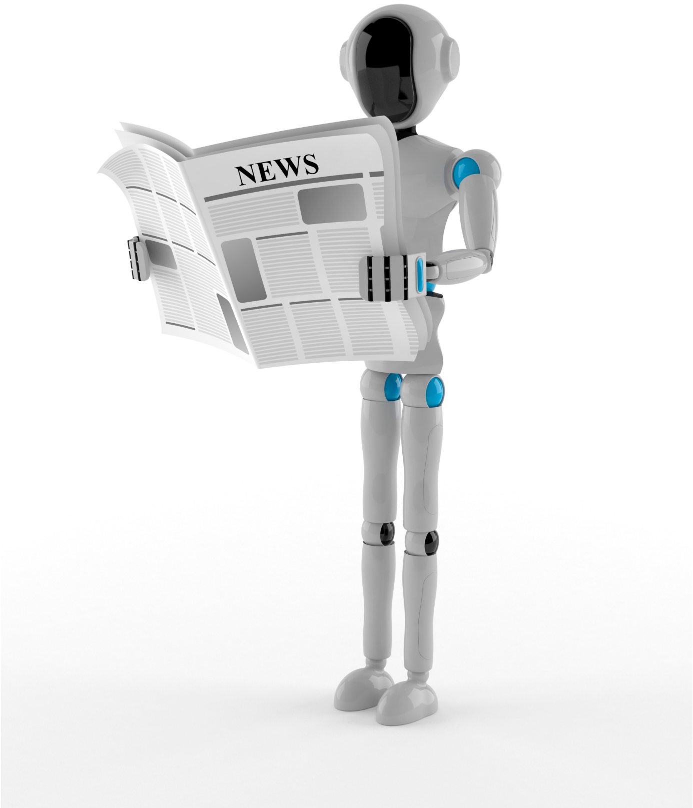
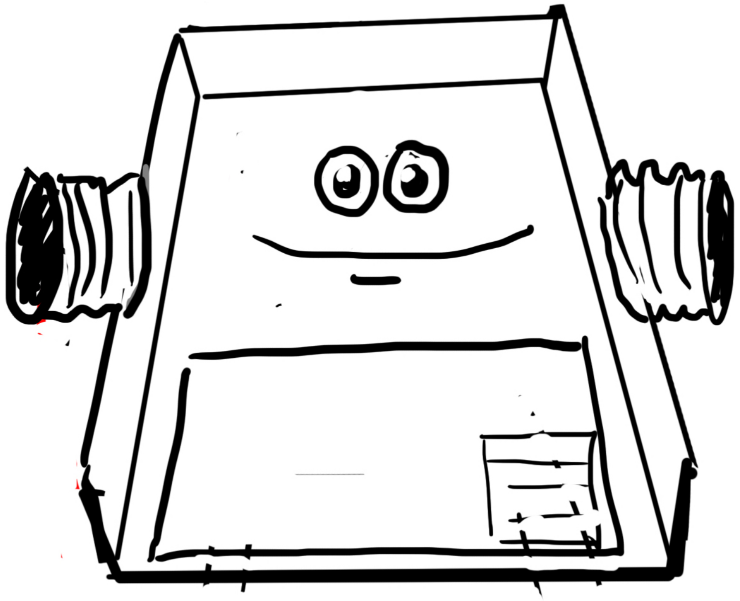
NewsReader
aggregating event-centric-knowledge graphs from massive streams of news



Piek Vossen
VU University Amsterdam

ICT 316404, FP7-ICT-2011-8
www.newsreader-project.eu



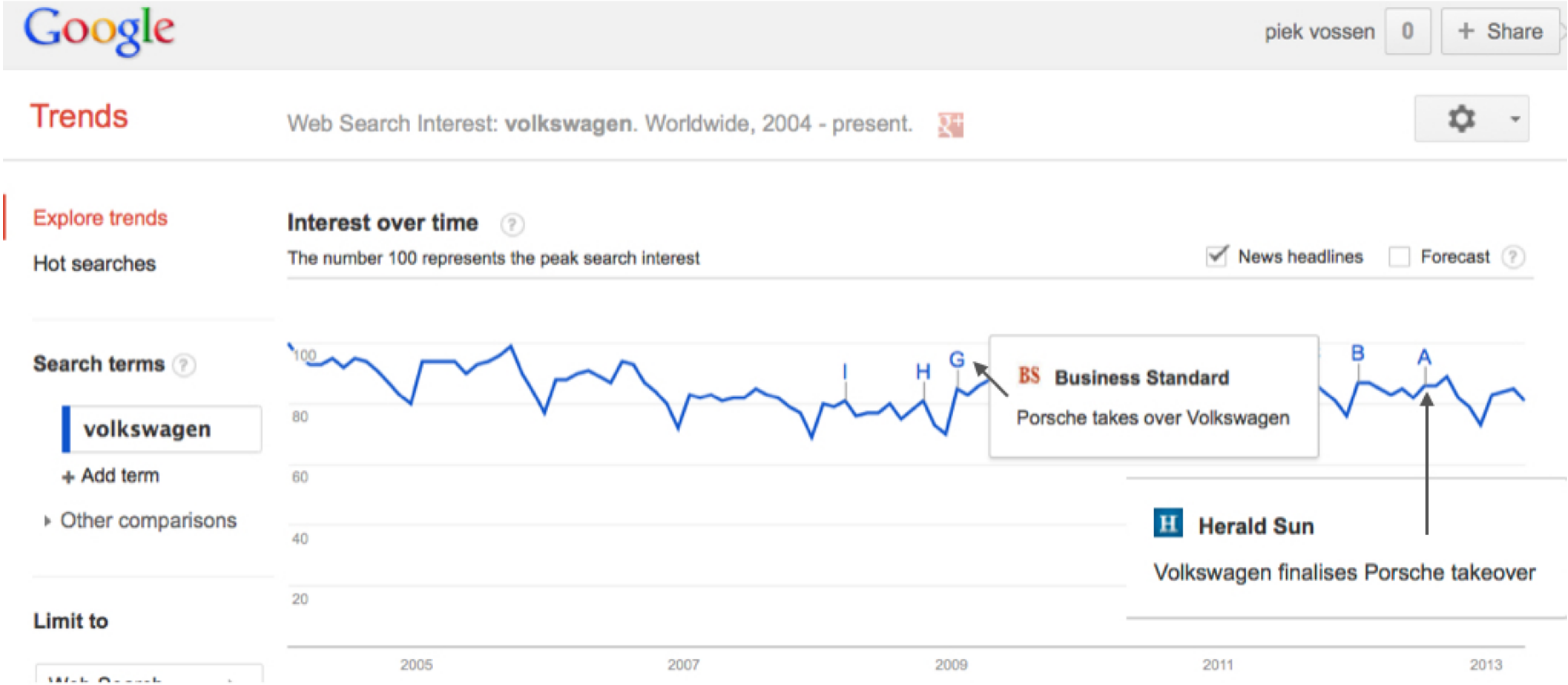
https://youtu.be/rYLaVN3oqLI

COOPERATION

Can we handle the news?

- Information broker LexisNexis:
  - 1.5 million articles on a single working day
  - 30,000 different sources

- How did the automotive industry change in the last 10 years?
  - read 6 million English news articles
  - Volkswagen takeover —> 2M Google hits



Google

piek vossen 0 + Share

Trends

Web Search Interest: **volkswagen**. Worldwide, 2004 - present.

Explore trends

Hot searches

Interest over time

The number 100 represents the peak search interest

☑ News headlines □ Forecast

Search terms

volkswagen

+ Add term

Other comparisons

Limit to

BS Business Standard

Porsche takes over Volkswagen

H Herald Sun

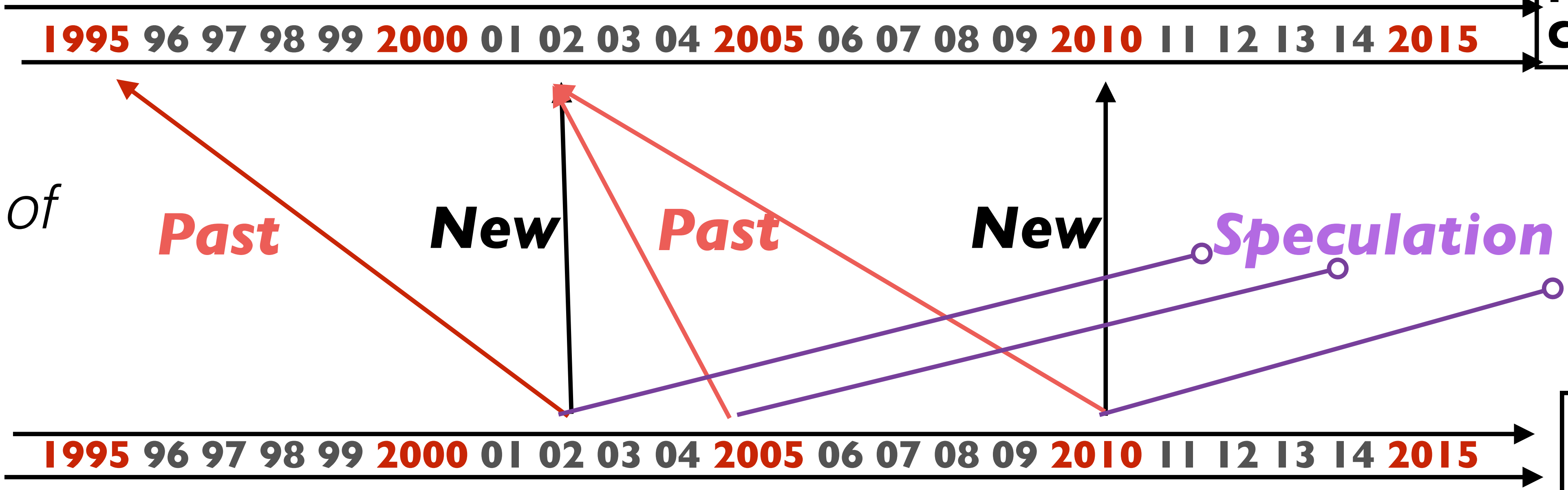
Volkswagen finalises Porsche takeover

VOLUME OF CHANGE

HOW MANY EVENTS, HOW MANY CHANGES?

2.2M entities

Volume of entities



50M mentions

2.3 MILLION ARTICLES

On 16 September 2008, Porsche ***increased its shares*** by another 4.89%, in effect ***taking control of*** the company, with more than 35% of the voting rights.

6 Jan 2009 – Porsche has been on *a **quest to takeover*** VW for more than two years.

NewsReader (ict316404)

- ***Reading Technology*** to process massive streams of news from many different sources in 4 languages (English, Dutch, Spanish and Italian):
    - ***What*** happened, ***where*** and ***when***, ***who*** was involved.
    - Recording the *changes* in the world as they are told in the media over long periods of time → ***history-recorder***.
    - ***KnowledgeStore*** to combine with background knowledge and to support reasoning
    - Who made what statement, where do sources agree and disagree: ***provenance*** and ***perspective***

2013-06-17

http://english.alarabiya.net

http://www.telegraph.co.uk

Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

2013-06-17

http://english.alarabiya.net

Qatar Holding sells 10% stake in Porsche to founding families

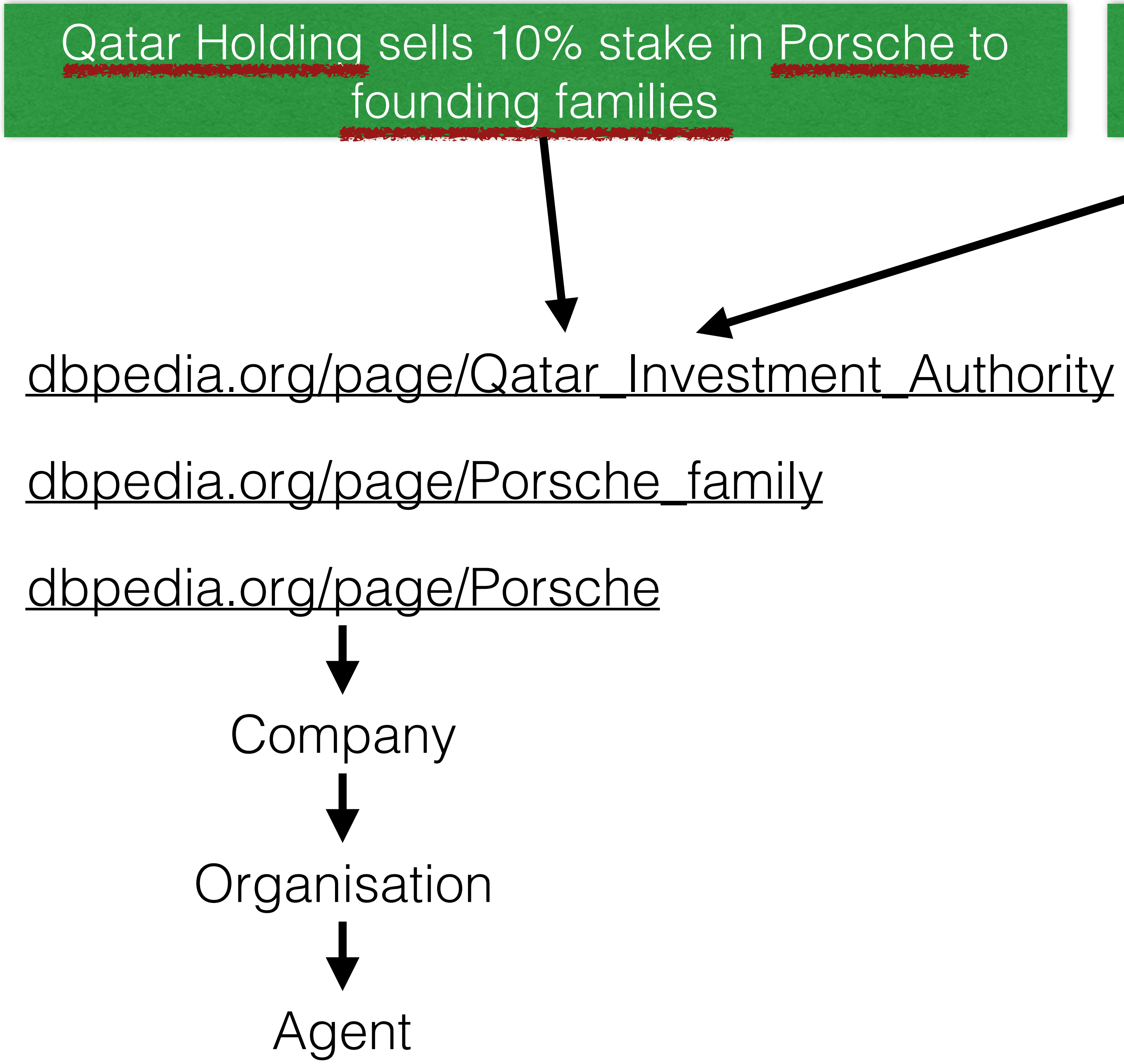
http://www.telegraph.co.uk

Porsche family buys back 10pc stake from Qatar

2013-06-17

http://english.alarabiya.net

http://www.telegraph.co.uk



Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

dbpedia.org/page/Qatar_Investment_Authority

dbpedia.org/page/Porsche_family

dbpedia.org/page/Porsche

Company

Organisation

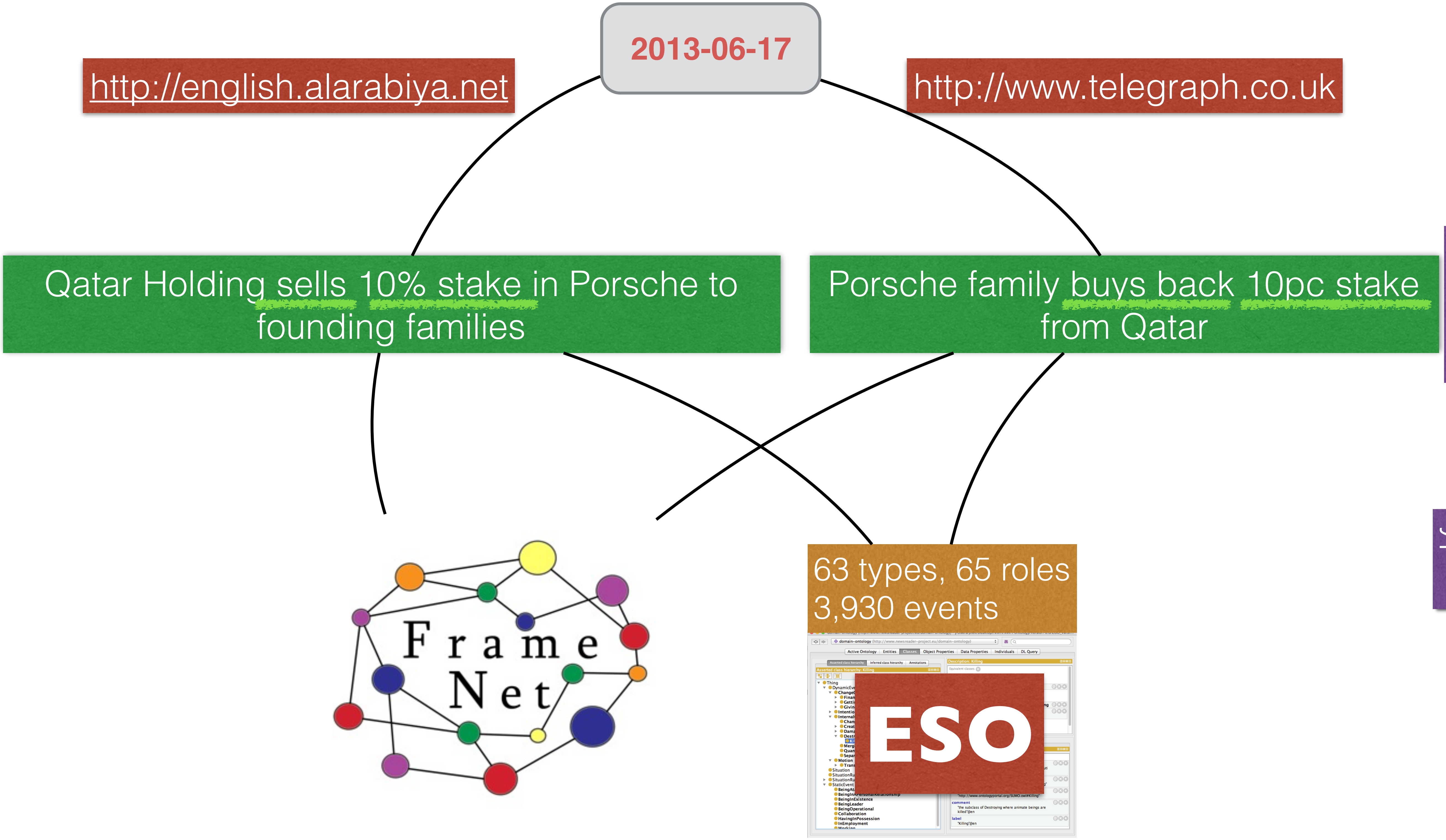
Agent

1,445,000 persons,
735,000 places,
241,000 organisations

DBpedia

mentions

instances



2013-06-17

http://english.alarabiya.net

http://www.telegraph.co.uk

Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

63 types, 65 roles
3,930 events

mentions

types

fn:Commerce_money_transfer

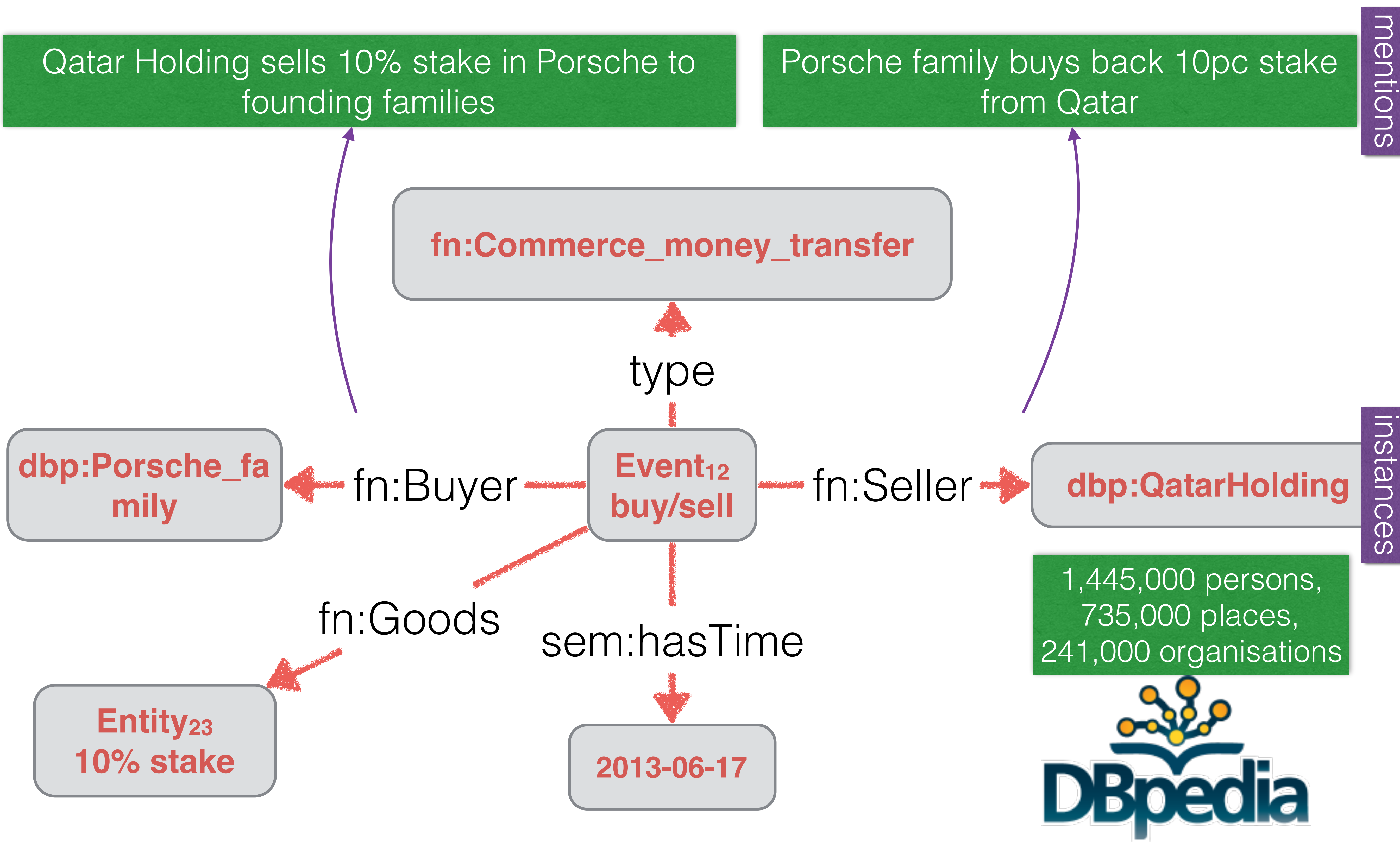
fn:Seller fn:Buyer fn:Goods fn:Money

http://english.alarabiya.net

2013-06-17

http://www.telegraph.co.uk

Event-centric-knowledge-graph (ECKG)



mentions

Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

fn:Commerce_money_transfer

type

dbp:Porsche_family

fn:Buyer

Event12 buy/sell

fn:Seller

dbp:QatarHolding

instances

fn:Goods

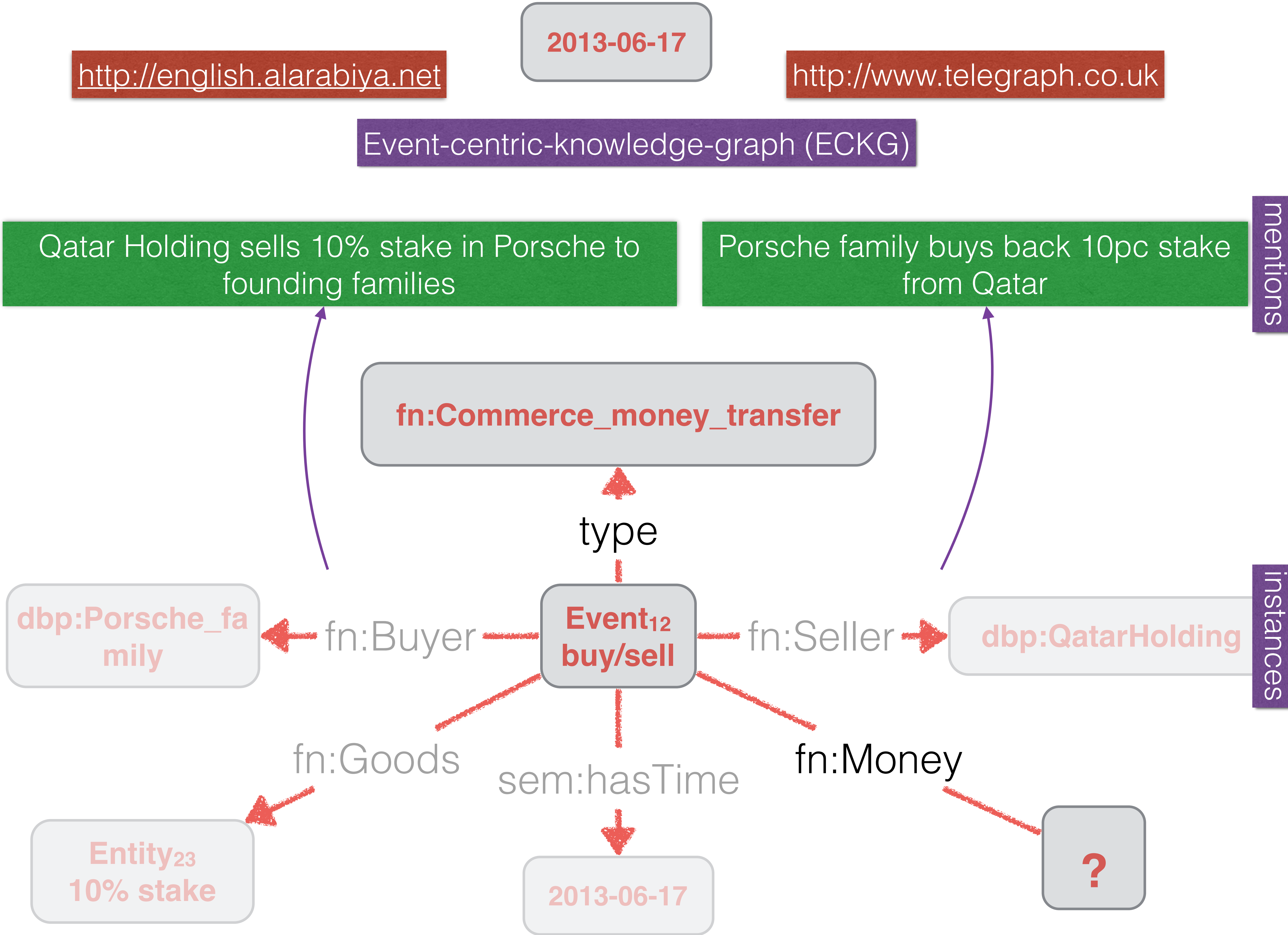
sem:hasTime

1,445,000 persons,
735,000 places,
241,000 organisations

Entity23 10% stake

2013-06-17

DBpedia



2013-06-17

http://english.alarabiya.net

http://www.telegraph.co.uk

Event-centric-knowledge-graph (ECKG)

Qatar Holding sells 10% stake in Porsche to founding families

Porsche family buys back 10pc stake from Qatar

mentions

fn:Commerce_money_transfer

type

dbp:Porsche_family

fn:Buyer

Event12 buy/sell

fn:Seller

dbp:QatarHolding

instances

fn:Goods

sem:hasTime

fn:Money

Entity23 10% stake

2013-06-17

?

Event-Centric Knowledge -Graphs

- Capture dynamic changes —> knowledge at points in time

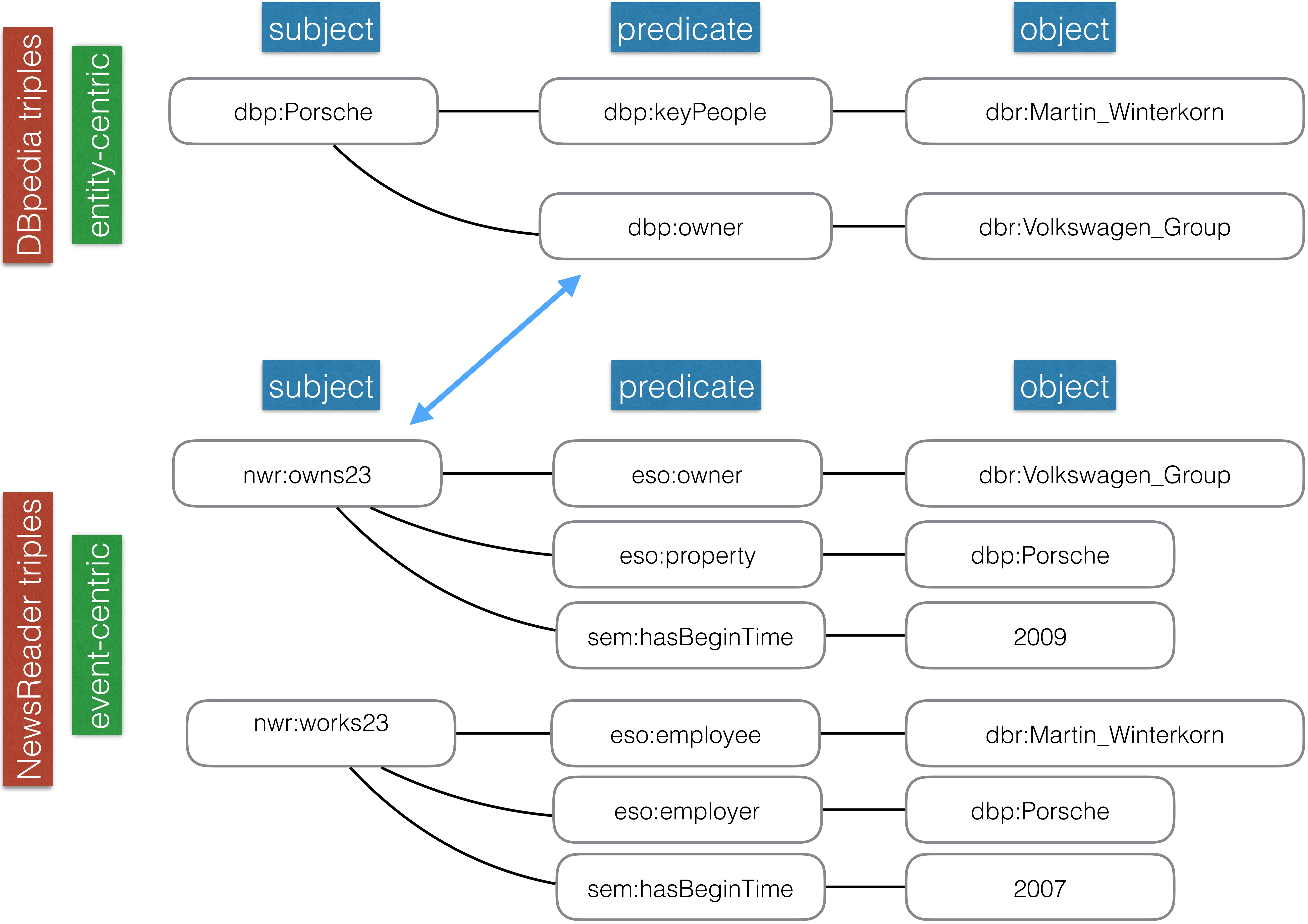
- Events represented once as instance objects —> deduplicates, no inconsistencies

- Events are subjects in triples:

    - subject:sell#24566, predicate:semActor, object: 10%stake#764334.

Entity-Centric Knowledge -Graphs

- DBpedia, Google knowledge graph
    - Give static biographies for entities with main events and facts
    - Duplicate information across entities which may lead to inconsistencies:
        - wikipedia:Porsche buys back 10% stake from Qatar
        - wikipedia:Qatar holds 17% stake in Porsche (sales is not mentioned and fact is out of date)
    - Events are properties in triples which do not represent instances and to which you cannot attach other properties such as begin and end time:
        - subject:Qatar, predicate:sell, object:10%stake



DBpedia triples

entity-centric

subject

predicate

object

dbp:Porsche

dbp:keyPeople

dbr:Martin_Winterkorn

dbp:owner

dbr:Volkswagen_Group

NewsReader triples

event-centric

subject

predicate

object

nwr:owns23

eso:owner

dbr:Volkswagen_Group

eso:property

dbp:Porsche

sem:hasBeginTime

2009

nwr:works23

eso:employee

dbr:Martin_Winterkorn

eso:employer

dbp:Porsche

sem:hasBeginTime

2007

IDAP method

- **Identification**: mentions of events are similar if their components are similar —> mentions to instances
- **Deduplication**: similar information is represented only once
- **Aggregation**: complementary information is combined in a single representation
- **Perspectivation**: differences and different view points are traceable through their sources and mentions in text

Event identity and reference

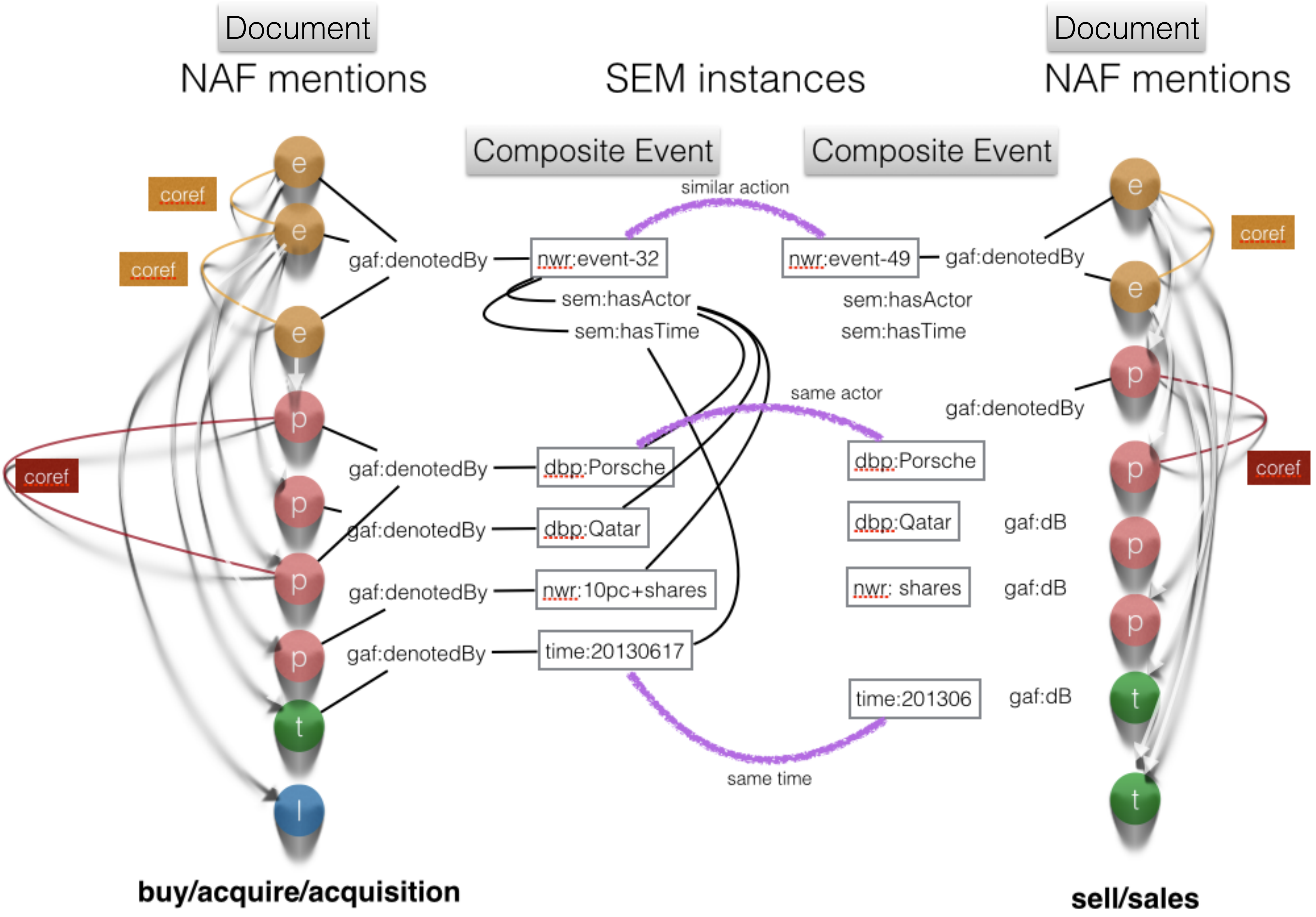
- **Composite events**: action **A**, participants **P**, location **L**, time **T** (Quine 1985)
    - genocide in Srebrenica, genocide in Rwanda, killings in Bosnia, Cafetaria bombing in Spain in 1974, train bombings in Madrid years ago
- **Components** spread over the complete document and not just within a single sentence
    - THOUSANDS of frightened residents flooded make shift refugee camps in **Indonesia 's West Papua** province today after **two** powerful **earthquakes** flattened buildings and killed at least one person ….As aid started to arrive , hundreds of aftershocks continued to rattle the coastal city which was hit by the **7.6** and **7.5** magnitude **quakes** early on **Sunday** , cutting power and prompting a brief tsunami warning .
    - The “American Pie” **actress** has **entered** **Promises** for undisclosed reasons. The **actress**, 33, reportedly **headed** to a **Malibu treatment facility** on **Tuesday**.

Two step approach

- Composite events: action + participants + location + time.
- Aggregate composite events from a single document from multiple sentences with coreferential event mentions (similarity): abstract event summary
- Compare composite events across documents:
  - Anchored to the same date (publication date and tense)
  - Similar actions (same word, WordNet similarity, word-embeddings)
  - Share sufficient participants and roles
- Exclude: source introducing events (***say, claim***), grammatical events (**stop, cause**), future events (**speculations**)

Cross document Event coreference

- Cybulska and Vossen 2015
  - Event mention identity ***I*** based on identity of components
    - $\boldsymbol{I}(Ei_e, Ei_f) = a.SIM(\mathbf{Am}_{i,j})\ p.SIM(\mathbf{Pm}_{p,q})\ l.SIM(\mathbf{Lm}_{l,m})\ t.SIM(\mathbf{Tm}_{n,o})$
    - (r, Am_i, Pm_p) & (r, Am_i, Lm_l) & (r, Am_i, Tm_n)
    - (r, Am_j, Pm_q) & (r, Am_j, Lm_m) & (r, Am_j, Tm_o)
    - a, p, l, t factors given the data collection
  - If ***I*** above threshold then merge all event components from two mentions into a single unique instance representation



Document

NAF mentions

SEM instances

Document

NAF mentions

Composite Event

Composite Event

similar action

coref

coref

gaf:denotedBy

nwr:event-32

nwr:event-49

gaf:denotedBy

coref

sem:hasActor

sem:hasTime

sem:hasActor

sem:hasTime

same actor

gaf:denotedBy

coref

gaf:denotedBy

dbp:Porsche

dbp:Porsche

coref

gaf:denotedBy

dbp:Qatar

dbp:Qatar

gaf:dB

gaf:denotedBy

nwr:10pc+shares

nwr: shares

gaf:dB

gaf:denotedBy

time:20130617

time:201306

gaf:dB

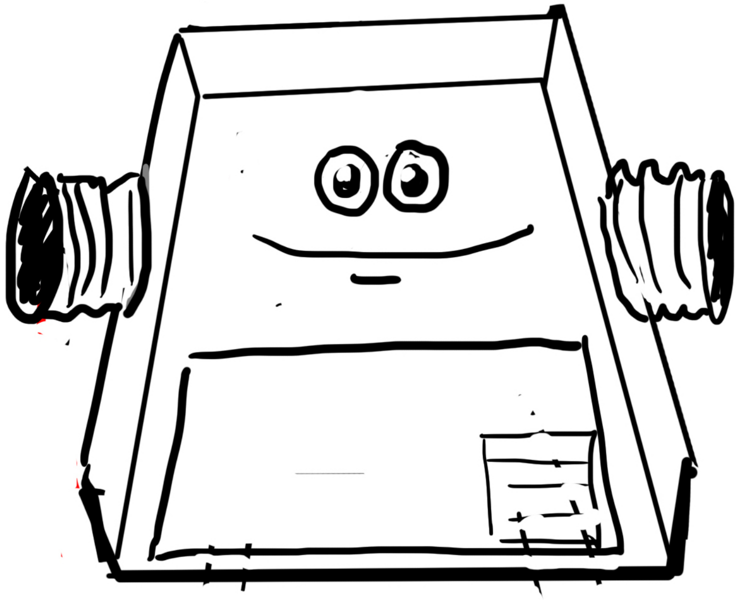
same time

buy/acquire/acquisition

sell/sales

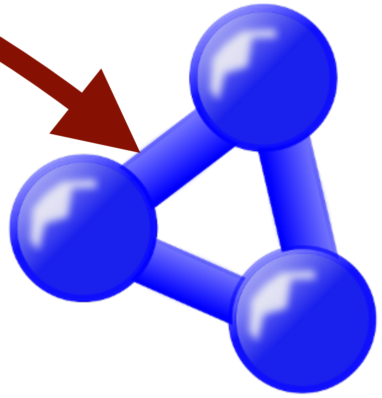
The Reading Machine

automotive industry
2003 - 2015



Semantic Web
RDF-TRIPLES

what -who - where - when



2.3 million articles
420 million event mentions
50 million entity mentions

1.2 billion statements
40 million event instances
2.2 million people, organisations, places

Mention → Instance

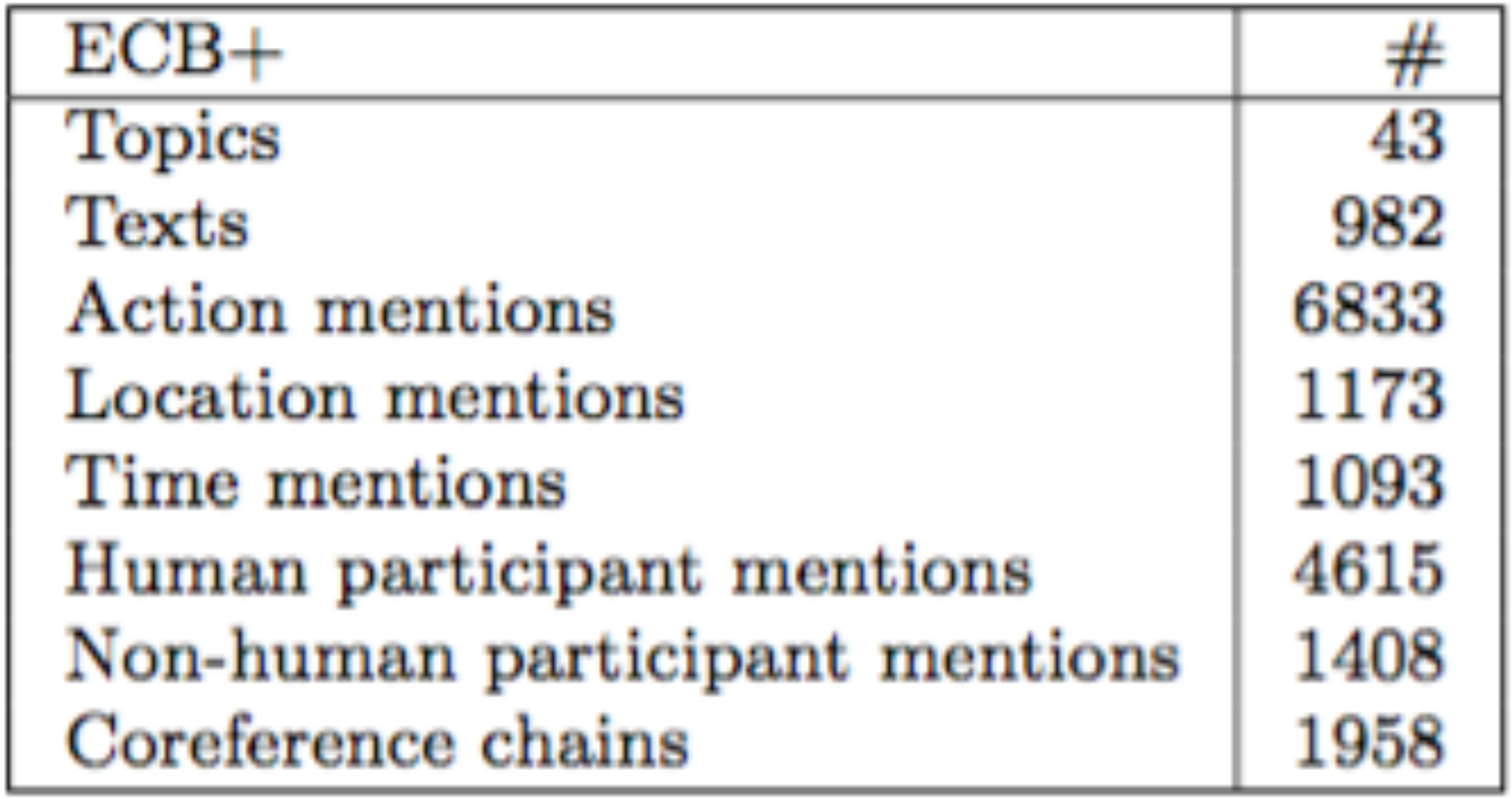
reduction by factor 10 - 20

Evaluation

- Cross-document event coreference —> IDAP
- RDF triple sample —> event-centric knowledge-graphs
- Timelines
- ESO reasoning

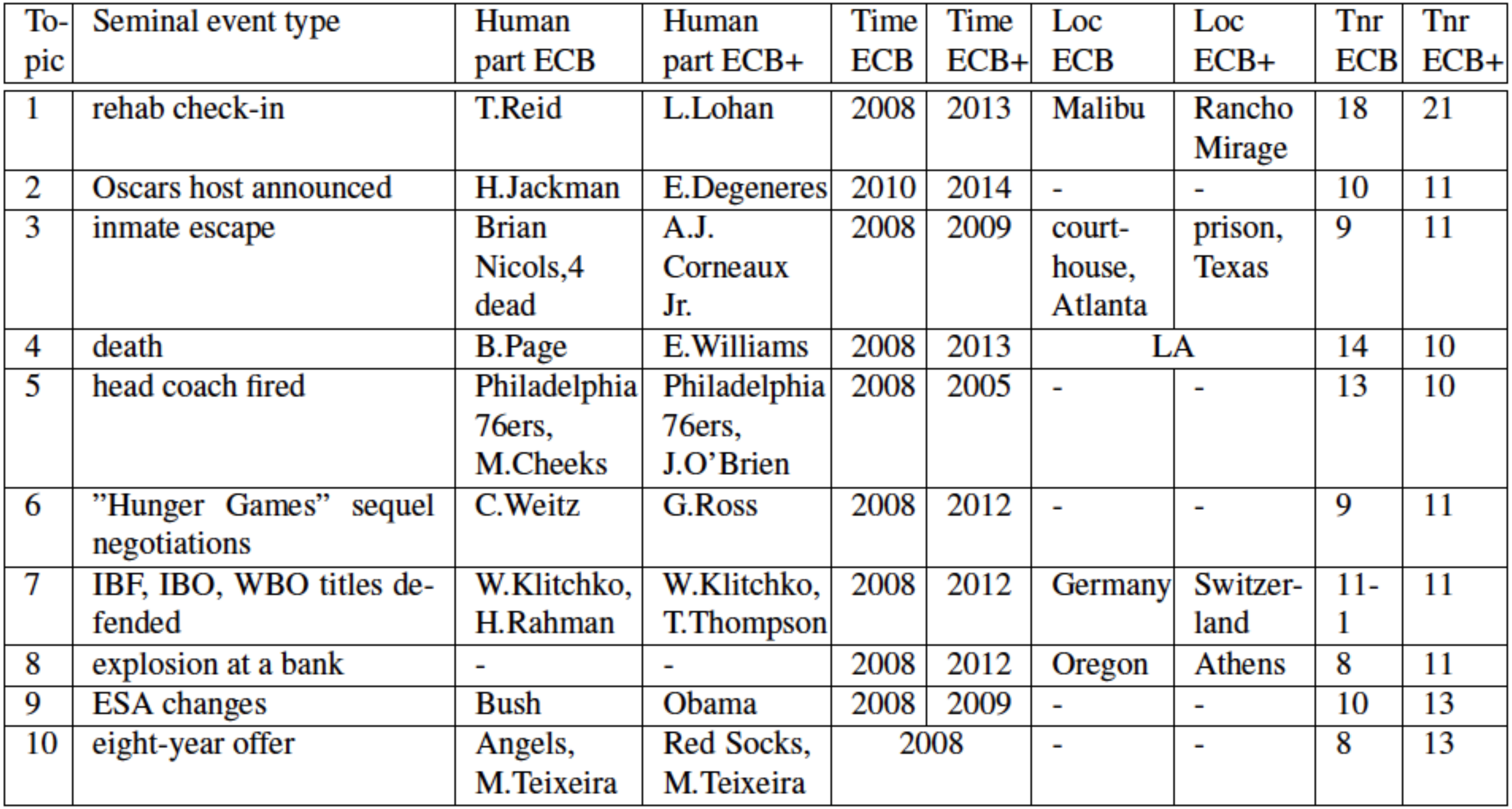
Cross-document event-coreference

- Event Coreference Bank (ECB, Bejan and Harabagiu 2010).
- Extended and re-annotated (Cybulska and Vossen (2014)



ECB+	#
Topics	43
Texts	982
Action mentions	6833
Location mentions	1173
Time mentions	1093
Human participant mentions	4615
Non-human participant mentions	1408
Coreference chains	1958

From Event Coreference Bank to ECB+ 1840 sentences annotated in 982 articles: 1.87 sentence/article.



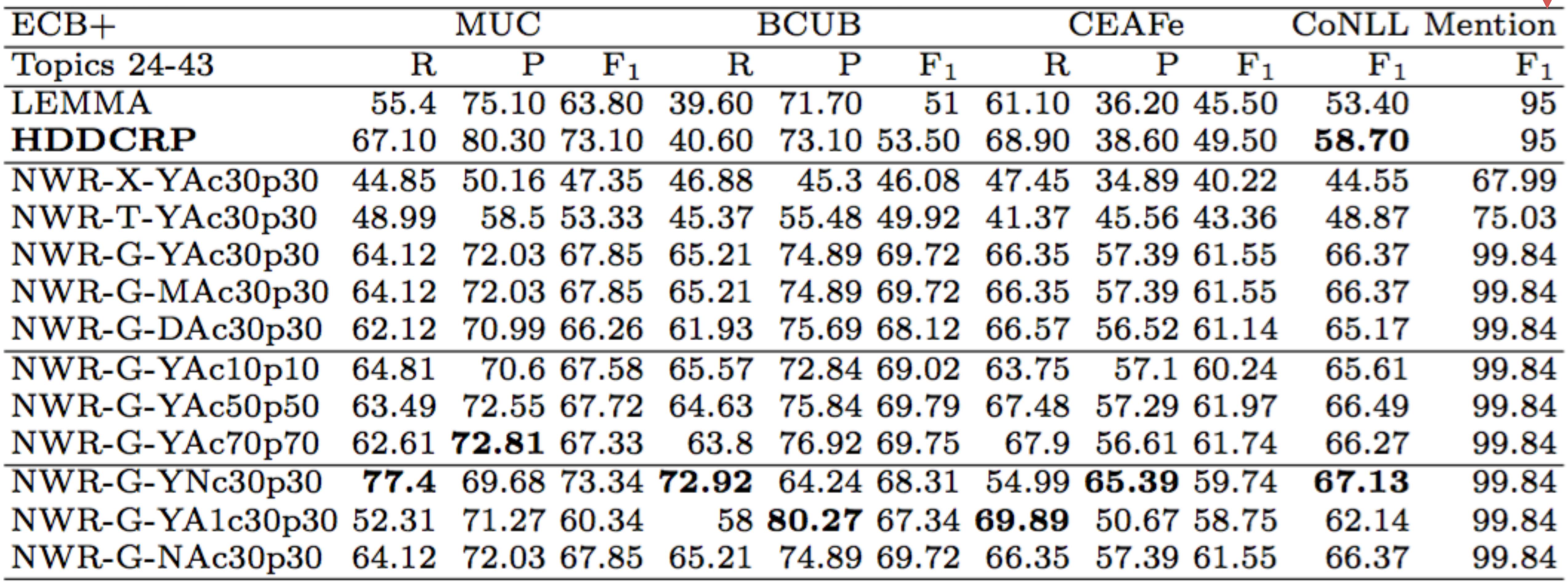
To-pic	Seminal event type	Human part ECB	Human part ECB+	Time ECB	Time ECB+	Loc ECB	Loc ECB+	Tnr ECB	Tnr ECB+
1	rehab check-in	T.Reid	L.Lohan	2008	2013	Malibu	Rancho Mirage	18	21
2	Oscars host announced	H.Jackman	E.Degeneres	2010	2014	-	-	10	11
3	inmate escape	Brian Nicols,4 dead	A.J. Corneaux Jr.	2008	2009	court-house, Atlanta	prison, Texas	9	11
4	death	B.Page	E.Williams	2008	2013	LA		14	10
5	head coach fired	Philadelphia 76ers, M.Cheeks	Philadelphia 76ers, J.O'Brien	2008	2005	-	-	13	10
6	"Hunger Games" sequel negotiations	C.Weitz	G.Ross	2008	2012	-	-	9	11
7	IBF, IBO, WBO titles defended	W.Klitchko, H.Rahman	W.Klitchko, T.Thompson	2008	2012	Germany	Switzer-land	11-1	11
8	explosion at a bank	-	-	2008	2012	Oregon	Athens	8	11
9	ESA changes	Bush	Obama	2008	2009	-	-	10	13
10	eight-year offer	Angels, M.Teixeira	Red Socks, M.Teixeira	2008		-	-	8	13

```
nwr:45_12ecbplus#ev10
        rdfs:label
                murder , kill , assassination , execution , Killing ,
        Shooting , slaying ;
        skos:prefLabel  murder ;
    gaf:denotedBy
                nwr:45_1ecbplus#char=1808,1815 , nwr:45_12ecbplus#char=109,115 ,
                nwr:45_5ecbplus#char=3281,3287 , nwr:45_6ecbplus#char=99,107 ,
                nwr:45_1ecbplus#char=1906,1913 , nwr:45_1ecbplus#char=5673,5686 ,
                etc... ;
        a
                ili:i28310 ,  ili:i28306 ,  ili:i28311 , ili:i34133 ,
        ili:i36562 , ili:i35417,  ili:i34134 , ili:i34139 ,
        ili:i34130
                fn:Killing , fn:Attack ,fn:Execution , , sem:Event , ;
        sem:hasActor
                dbp:Jerome_Flynn (Flynn , Herbert Flynn , his , Ka'Loni Flynn ,
                        Ka'Loni Flynn's , Ka'loni Flynn) ;
        sem:hasTime     nwr:45_6ecb#tmx2 (time:20121112 , Nov. 12) .

nwr:45_6ecbplus#ev16
        rdfs:label      charge , shooting , shoot ;
        skos:prefLabel  shoot ;
        gaf:denotedBy
                nwr:45_9ecbplus#char=640,644 , nwr:45_2ecbplus#char=633,637 ,
                nwr:45_4ecbplus#char=513,517 , nwr:45_7ecbplus#char=403,411 ,
                nwr:45_2ecbplus#char=359,366 , nwr:45_7ecbplus#char=69,77 ,
                etc... ;
        a
                ili:i106612 , ili:i25451 , ili:i25858 , ili:i25860 ,
        ili:i25976 , ili:i26598 , ili:i26600 , ili:i27206 ,
        ili:i27278 , ili:i27293 , ili:i27599 , ili:i29722 ,
        ili:i30898 , ili:i30954 , ili:i32022 , ili:i32053 ,
        ili:i33338 , ili:i34100 , ili:i34141 , ili:i35084 ,
                ili:i36049 , ili:i36050 , ili:i36591 , ili:i40503 ,
        ili:i70941 , ili:i27599 , ili:i32022 , ili:i26598 ,
        ili:i33338 , ili:i36049 , ili:i30898 , ili:i106612 ,
        ili:i27278 , ili:i26600 , ili:i25976 ,
                fn:Commerce_collect , fn:Motion , fn:Process_continue ,
                fn:Commerce_pay , fn:Killing , fn:Notification_of_charges ,
                fn:Hit_target , fn:Shoot_projectiles , fn:Use_firearm ;
        sem:hasActor
                dbp:Electoral_division_of_Flynn ,
                dbp:Jerome_Flynn (Flynn , Herbert Flynn , his , Ka'Loni Flynn ,
                Ka'Loni Flynn's , Ka'loni Flynn) ,
                dbp:Oklahoma (okla , Oklahoma , Okla , Okla - man) ,
        dbp:Robb_Flynn (Ka'loni Flynn , Flynn) ,
                dbp:Fort_Smith,_Arkansas ,
        nwr:entities/ChristopherKenyonSimpson ,
                dbp:Christopher_Simpson ,
                nwr:entities/Spiroman ,
                dbp:Arkansas ,
                dbp:O._J._Simpson (Purportedly Simpson , Simpson , his) ;
    sem:hasTime nwr:45_6ecb#tmx2 (time:2012 , 2012) .
```

Event detection

NewsReader extraction



ECB+	MUC			BCUB			CEAFe			CoNLL	Mention
Topics 24-43	R	P	F_1	R	P	F_1	R	P	F_1	F_1	F_1
LEMMA	55.4	75.10	63.80	39.60	71.70	51	61.10	36.20	45.50	53.40	95
HDDCRP	67.10	80.30	73.10	40.60	73.10	53.50	68.90	38.60	49.50	**58.70**	95
NWR-X-YAc30p30	44.85	50.16	47.35	46.88	45.3	46.08	47.45	34.89	40.22	44.55	67.99
NWR-T-YAc30p30	48.99	58.5	53.33	45.37	55.48	49.92	41.37	45.56	43.36	48.87	75.03
NWR-G-YAc30p30	64.12	72.03	67.85	65.21	74.89	69.72	66.35	57.39	61.55	66.37	99.84
NWR-G-MAc30p30	64.12	72.03	67.85	65.21	74.89	69.72	66.35	57.39	61.55	66.37	99.84
NWR-G-DAc30p30	62.12	70.99	66.26	61.93	75.69	68.12	66.57	56.52	61.14	65.17	99.84
NWR-G-YAc10p10	64.81	70.6	67.58	65.57	72.84	69.02	63.75	57.1	60.24	65.61	99.84
NWR-G-YAc50p50	63.49	72.55	67.72	64.63	75.84	69.79	67.48	57.29	61.97	66.49	99.84
NWR-G-YAc70p70	62.61	**72.81**	67.33	63.8	76.92	69.75	67.9	56.61	61.74	66.27	99.84
NWR-G-YNc30p30	**77.4**	69.68	73.34	**72.92**	64.24	68.31	54.99	**65.39**	59.74	**67.13**	99.84
NWR-G-YA1c30p30	52.31	71.27	60.34	58	**80.27**	67.34	**69.89**	50.67	58.75	62.14	99.84
NWR-G-NAc30p30	64.12	72.03	67.85	65.21	74.89	69.72	66.35	57.39	61.55	66.37	99.84

- LEMMA = baseline
- HDDCRP, hierarchical distance-dependent Chinese Restaurant Process Yang et al 2015
- NWR (Newsreader):
  - X = out-of-the-box, T = event detection using CRF trained on TimeEval2013 corpus, G = true mentions of events (gold data)
  - Y=year, M=Month, D= Day, N=none
  - A=participant in any role, A1=participant in PropBank, N=none
  - c10,30,50,70 = overlap of concepts for actions, p10,30,50,70 = overlap of surface forms for actions

Discussion

- Quality of entity coreference, action similarity, time detection and normalisation;

- Sparseness of data within sentence, and difficulty to collect data across sentences;

- ECB+ better than ECB but still limited variation and referential ambiguity —> too easy!!!

- 90% of event mentions in ECB+ not coreferential (95% in MEANTIME) —> annotators are very conservative

- Other relations included: subclass, subvert, topical relations

From NewsReader
to
NewsReasoner