MUC-7 EVALUATION OF IE TECHNOLOGY: Overview of Results

Elaine Marsh (NRL)
Dennis Perzanowski (NRL)

MUC-7
29 April 1998

MUC-7 Program Committee

Ralph Grishman (NYU), Co-Chair
Elaine Marsh (NRL), Co-Chair
Chinatsu Aone (SRA)
Lois Childs (Lockheed-Martin)
Nancy Chinchor (SAIC)
Jim Cowie (NMSU)
Rob Gaizauskas (Sheffield)
Megumi Kameyama (SRI)
Tom Keenan (DoD)
Boyan Onyshkevych (DoD)
Martha Palmer (Penn)
Beth Sundheim (SPAWARSYSCEN)
Marc Vilain (MITRE)
Ralph Weischedel (BBN)

MUC-7 Evaluation Members

Evaluation Participants:

American University in Cairo
BBN
FACILE
Isoquest
Kent Ridge Digital Laboratories
Lockheed-Martin
MITRE
National Taiwan University
New York University
NTT
Oki Electric Industry Co., Ltd.
SRA
TASC, Inc.
University of Durham
University of Edinburgh and Thomson
University of Manitoba
University of Pennsylvania
University of Sheffield

Evaluation Support:

Naval Research Laboratory
SAIC (San Diego)
MUC-7 Program Committee
DARPA/ITO and Tipster Program
Linguistic Data Consortium

Evaluation Participation by Task

ET:
ational Taiwan University
ational University of Singapore
ew York University
TT
ki Electric

amed Entity:
BN
ACILE
oquest
ITRE
ational Taiwan University
ational University of Singapore
ew York University
ki Electric
niversity of Durham
niversity of Edinburgh and Thomson
niversity of Manitoba
niversity of Sheffield

Template Element:
American University - Cairo
BBN
FACILE
Lockheed-Martin
New York University
Oki Electric
SRA
University of Durham
University of Sheffield

Template Relation:
American University - Cairo
BBN
Oki Electric
SRA
University of Sheffield

Scenario Template:
American University - Cairo
New York University
SRA
TASC
University of Sheffield

Coreference:
Oki Electric
University of Pennsylvania
University of Durham
University of Manitoba
University of Sheffield

IE Evaluation Tasks

- Named Entity Task [NE]: Insert SGML tags into the text to mark each string that represents a person, organization, or location name, or a date or time stamp, or a currency or percentage figure
- Multi-lingual Entity Task [MET]: NE task for Chinese and Japanese
- Template Element Task [TE]: Extract basic information related to organization, person, and artifact entities, drawing evidence from anywhere in the text

IE Evaluation Tasks

- Template Relation Task [TR]: Extract relational information on employee_of, manufacture_of, and location_of relations
- Scenario Template Task [ST]: Extract prespecified event information and relate the event information to particular organization, person, or artifact entities involved in the event.
- Coreference Task [CO]: Capture information on coreferring expressions: all mentions of a given entity, including those tagged in NE, TE tasks

Training and Data Sets

Corpus

New York Times News Service (supplied by Linguistic Data Consortium)

Evaluation Epoch: January 1 - September 11, 1996

Approximately 158,000 articles

- Training and test sets retrieved from corpus using Managing Gigabytes text retrieval system using domain relevant terms.

- 2 sets of 100 articles (aircraft accident domain) - preliminary training, including dryrun.

- 2 sets of 100 articles selected balanced for relevancy, type and source for formal run (launch event domain).

Training and Data Sets (con't)

Training Set

Training keys for NE, TE, TR available from preliminary set of 100 articles; CO from preliminary training set of 30 articles.

Formal training set of 100 articles and answer keys for ST task.

Test Set

100 Articles (and answer keys) for NE (Formal Training set)

100 articles (and answer keys) for TE, TR, ST

Subset of 30 articles (and answer keys) for CO task.

Test Procedure

Schedule

- 2-6 March: Formal Run Test for NE
- 9 March: Training set of articles available for electronic file transfer from SAIC (ST guidelines and keys).
- 31 March: Test set of articles available for electronic file transfer from SAIC.
- 6 April: Deadline for completing TE, TR, ST, and CO tests (via electronic file transfer of system outputs to SAIC)

Test Procedure (con't)

Notes on testing:

- Tests run by individual participating sites at their own facilities, following a written test procedure.
- Sites could conduct official "optional" tests in addition to the basic test.
- Adaptive systems were permitted.
- Walkthrough articles for:
  - NE
  - TR/TR/ST
  - CO

Example Text

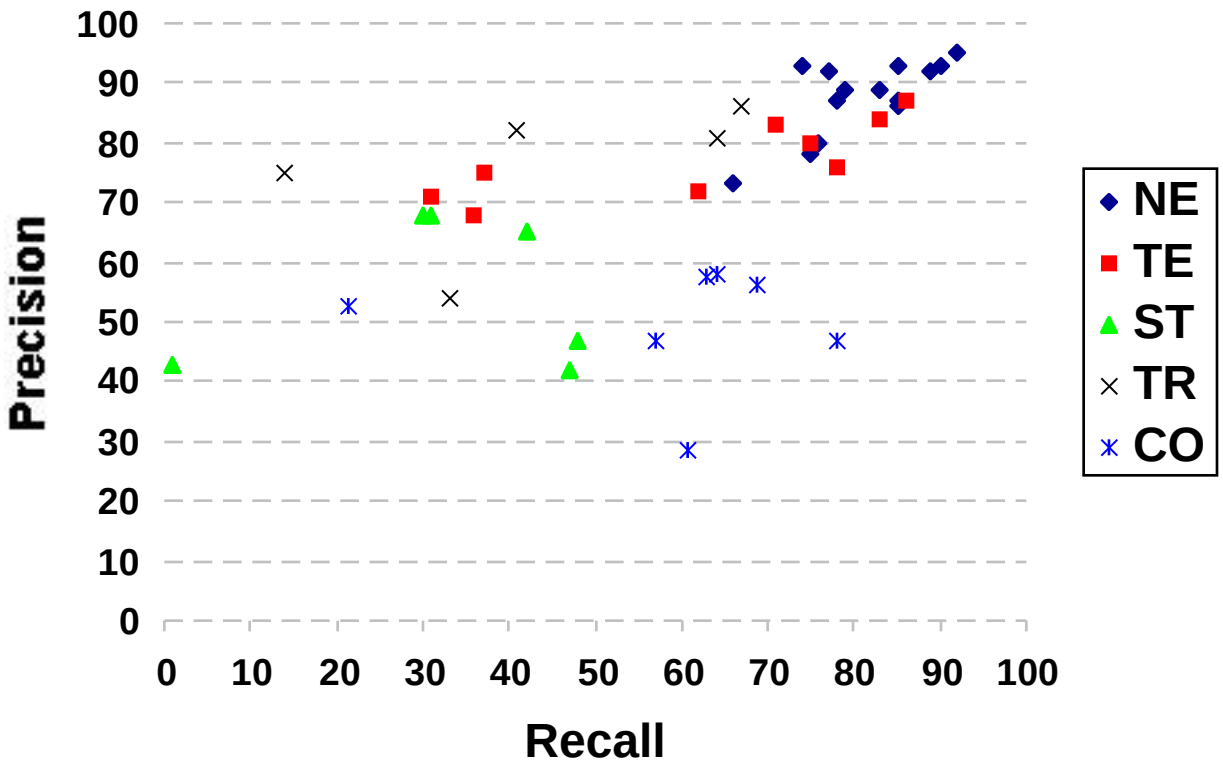
<DOC>
<DOCID> nyt960214.0704 </DOCID>
<STORYID cat=f pri=u> A4479 </STORYID>
<SLUG fv=taf-z> BC-MURDOCH-SATELLITE-NYT </SLUG>
<DATE> 02-14 </DATE>
<NWORDS> 0608 </NWORDS>
<PREAMBLE>
BC-MURDOCH-SATELLITE-NYT
MURDOCH SATELLITE FOR LATIN PROGRAMMING EXPLODES ON TAKEOFF
(kd)
By MARK LANDLER
c.1996 N.Y. Times News Service
</PREAMBLE>
<TEXT>
<p>

Example Text (con’t)

Chinese rocket carrying a television satellite exploded seconds after launch Wednesda
aling a potential blow to Rupert Murdoch's ambitions to offer satellite programming in
tin America.
>
Murdoch's News Corp. is one of four media companies in a partnership that had leased
ace on the Intelsat satellite to offer the Latin American service. The other partners are
le-Communications Inc., the nation's largest cable operator; Grupo Televisa SA, the
exican broadcaster and publisher, and the giant Brazilian media conglomerate Globo.
>

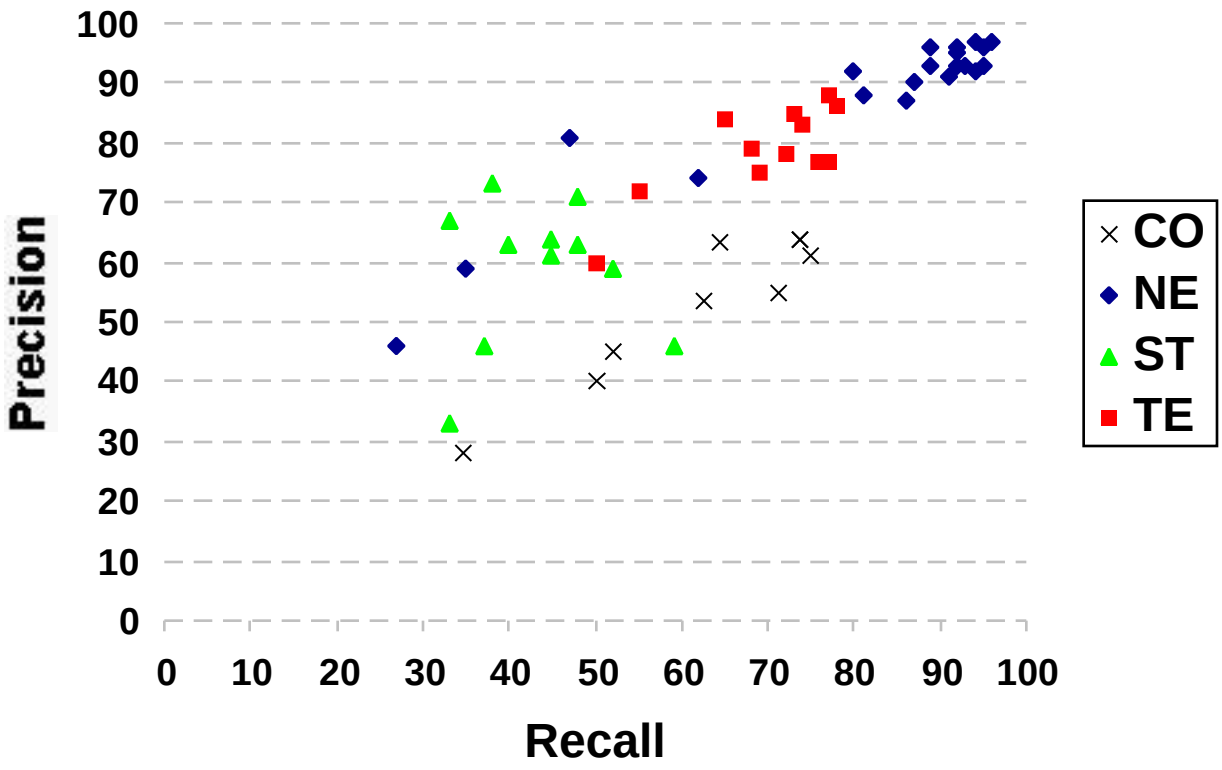
>
EXT>
RAILER>
T-02-14-96 2029EST
RAILER>
OC>

Composite Overall Results



MUC

Composite Overall Results



MUC

Named Entity Task (NE)

- NE mirrored Multilingual Entity Task
- SGML tagging in text stream from SLUG, DATE, PREAMBLE, TEXT, TRAILER
  - Elements: ENAMEX, NUMEX, TIMEX
  - Attributes: TYPE, STATUS (keys), MIN (keys)
- Markables
  - Names of organizations, persons, locations
  - Mentions of dates and times (relative and absolute)
  - Direct mentions of currency/percentage

Named Entity (NE) (con't)

- Non-markables
  - Artifacts (Wall Street Journal, MTV)
  - Common nouns used in anaphoric reference (the plane, the company,)
  - Names of groups of people and laws named after people (Republicans, Gramm-Rudman amendment, the Nobel prize)
  - Adjectival forms of location names (American, Japanese)
  - Miscellaneous uses of numbers which are not specifically currency or percentages (1 1/2 points, 1.5 times)
- Caveats: "newspaper" style, domain bias toward ST topic

NE Overall F-Measures

F-Measure	Error	Recall	Precision
93.39	11	92	95
91.60	14	90	93
90.44	15	89	92
88.80	18	85	93
86.37	22	85	87
85.83	22	83	89
85.31	23	85	86
84.05	26	77	92
83.70	26	79	89
82.61	29	74	93
81.91	28	78	87
77.74	33	76	80
76.43	34	75	78
69.67	44	66	73

Annotators:

97.60	4	98	98
96.95	5	96	98

MUC 7

NE Overall F-Measures

F-measure	Error	Recall	Precision
96.42	5	96	97
95.66	7	95	96
94.92	8	93	96
94.00	10	92	96
93.65	10	94	93
93.33	11	92	95
92.88	10	94	92
92.74	12	92	93
92.61	12	89	96
91.20	13	91	91
90.84	14	91	91
89.06	18	84	94
88.19	19	86	90
85.82	20	85	87
85.73	23	80	92
84.95	22	82	89

Annotators:

96.68	6	95	98
93.18	11	92	95

MUC

NE Scores by Document Section (ERR) sorted by F-Measure

F-Measure	Slug	Date	Preamble	Text
93.39	14	0	7	13
91.60	28	0	9	15
90.44	24	0	11	16
88.80	54	0	16	19
86.37	34	0	19	23
85.83	28	0	18	24
85.31	45	0	25	24
84.05	33	0	31	27
83.70	39	0	23	28
82.61	32	0	27	27
81.91	49	0	24	30
77.74	100	0	44	32
76.43	51	0	34	36
69.67	93	0	50	44

Annotators:

97.60	3	0	2	4
96.95	2	9	2	6

MUC

NE Scores by Document Section (ERR) sorted by F-Measure

F-Measure	Doc Date	Dateline	Headline	Text
96.42	0	0	8	5
95.66	0	0	7	7
94.92	0	0	8	8
94	0	0	20	9
93.65	0	2	16	10
93.33	0	4	38	9
92.88	0	0	18	10
92.74	0	0	22	11
92.61	100	0	18	9
91.2	0	0	30	13
90.84	3	11	19	14
89.06	3	4	28	18
88.19	0	0	22	20
85.82	0	6	18	21
85.73	0	44	53	21
84.95	0	0	50	21
Annotator:				
96.68	0	0	7	6

NE Subcategory Scores (ERR) sorted by F-measure

	enamex			timex		nume	
F-measure	**org**	**per**	**loc**	**date**	**time**	**money**	**p**
93.39	13	5	10	12	21	8	
91.60	21	7	10	12	19	11	
90.44	22	8	11	14	21	19	
88.80	25	12	16	15	22	23	
86.37	21	22	26	18	18	15	
85.83	27	19	24	16	20	20	
85.31	29	16	26	14	23	21	
84.05	44	22	17	14	19	10	
83.70	33	22	27	18	19	15	
82.61	25	10	12	58	100	17	
81.91	38	19	31	19	17	21	
77.74	40	24	32	27	27	26	
76.43	47	32	35	21	22	17	
69.67	60	47	44	26	22	25	
Annotators:							
97.60	3	1	1	5	5	1	
96.95	5	1	3	8	21	8	

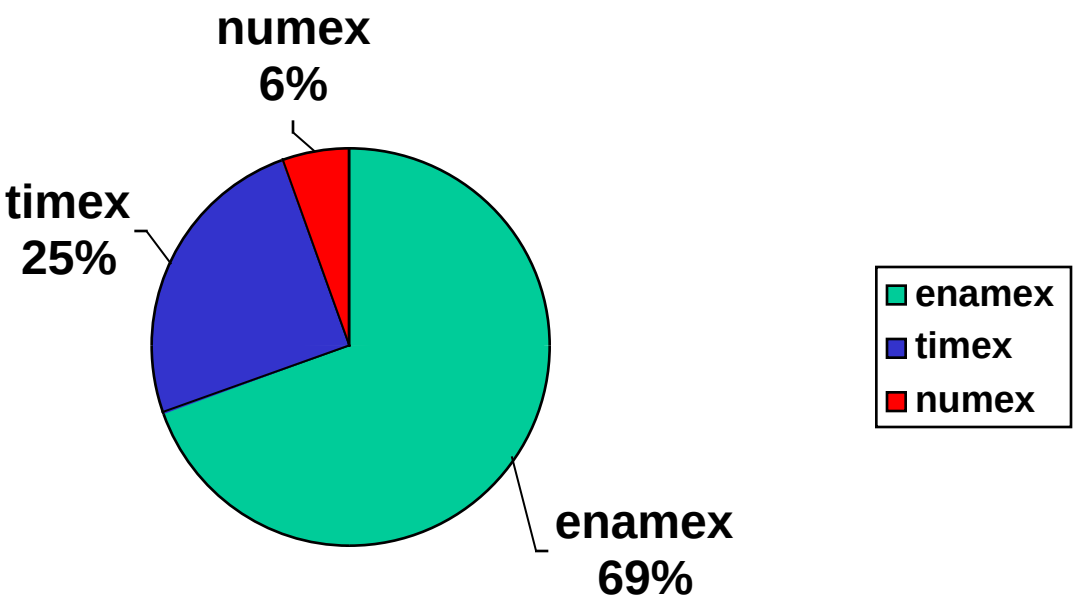
MU

NE Subcategory Scores (ERR) sorted by F-measure

	enamex			timex		numex	
F-measure	org	per	loc	date	time	money	percer
96.42	10	2	6	3	*	0	0
95.66	11	3	9	7	*	1	0
94.92	16	3	7	3	*	0	0
94.00	16	3	15	9	*	3	0
93.65	13	4	8	8	*	8	32
93.33	16	6	12	9	*	4	6
92.88	15	4	13	8	*	8	32
92.74	16	4	9	16	*	2	0
92.61	14	4	5	43	*	1	0
91.20	18	9	19	8	*	6	36
90.84	16	10	29	12	*	6	0
89.06	22	17	18	10	*	3	0
88.19	29	7	20	17	*	11	36
85.82	29	9	16	13	*	6	32
85.73	26	14	29	18	*	9	40
84.95	45	4	31	10	*	4	32
Annotator:							
96.68	6	1	4	8	*	0	0

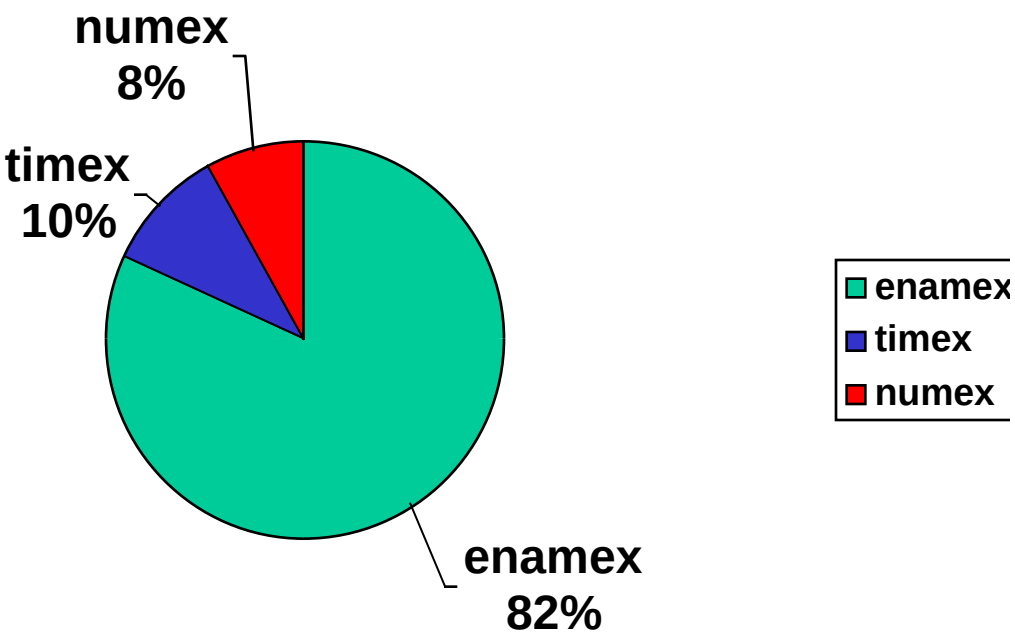
MUC

Distribution of NE tag elements



MUC

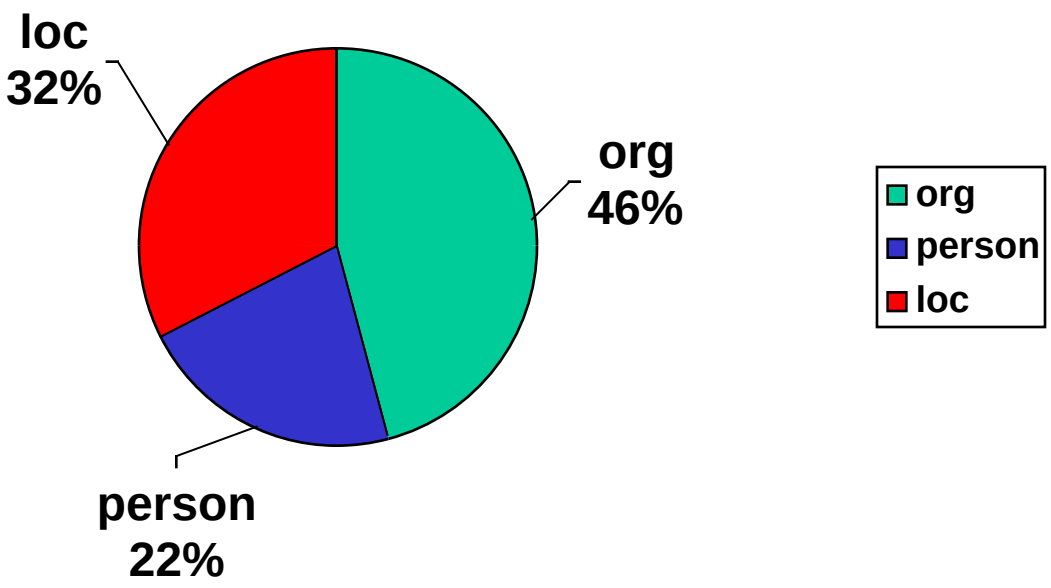
Distribution of NE tag elements



MUC

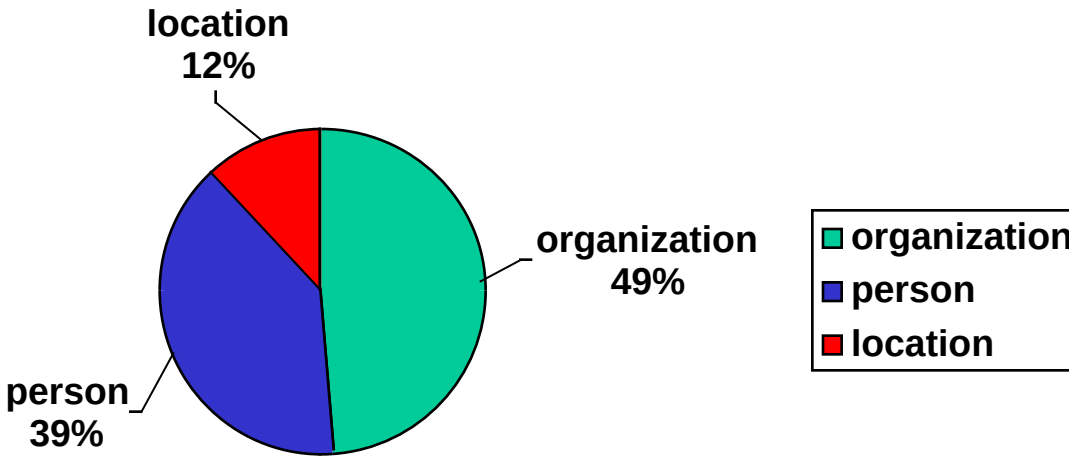
Distribution of NE

ENAMEX tag elements



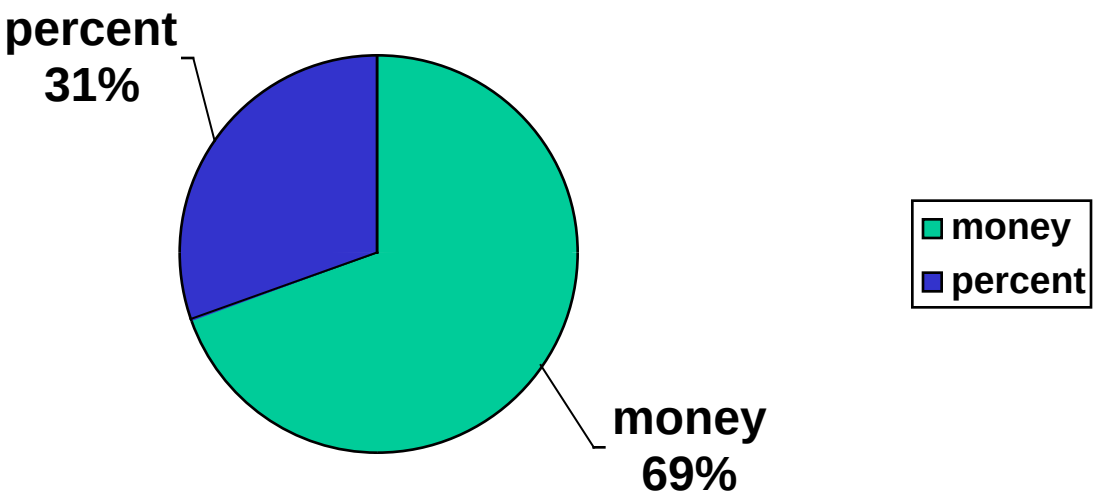
MUC

Distribution of NE tag elements
ENAMEX tag elements

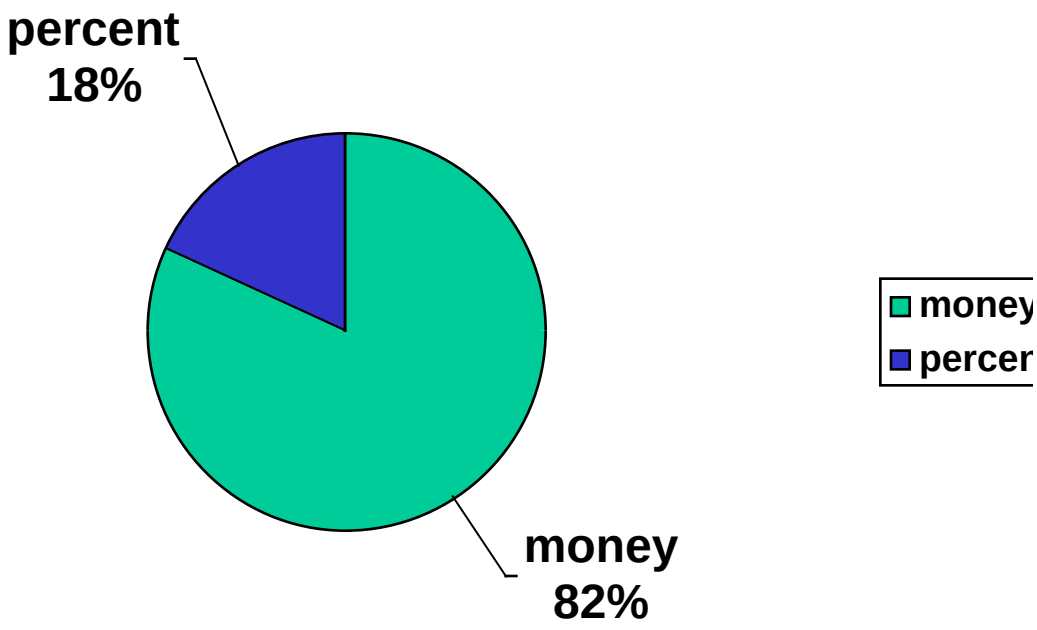


MUC

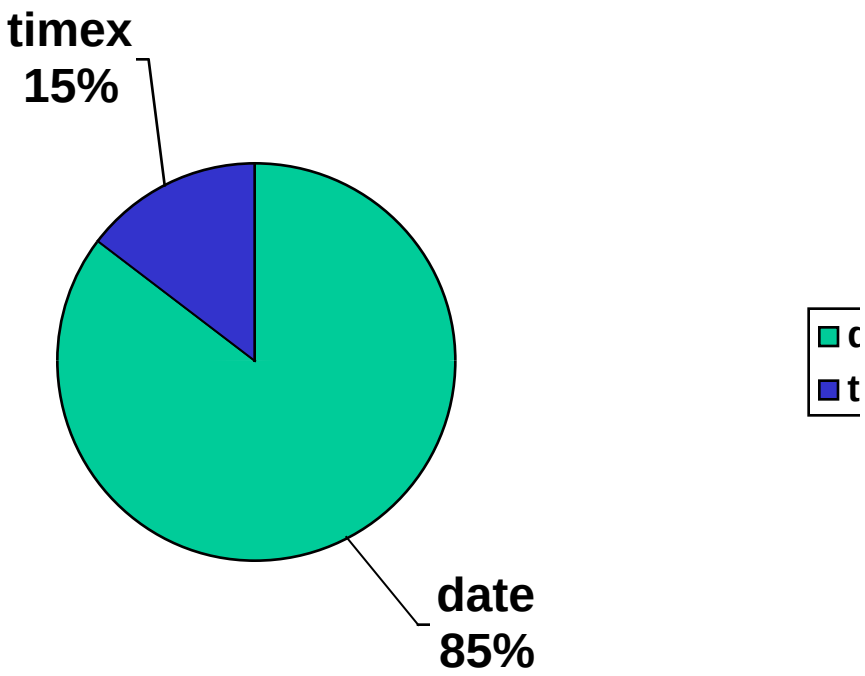
Distribution of NE NUMEX tag elements



Distribution of NE NUMEX tag elements

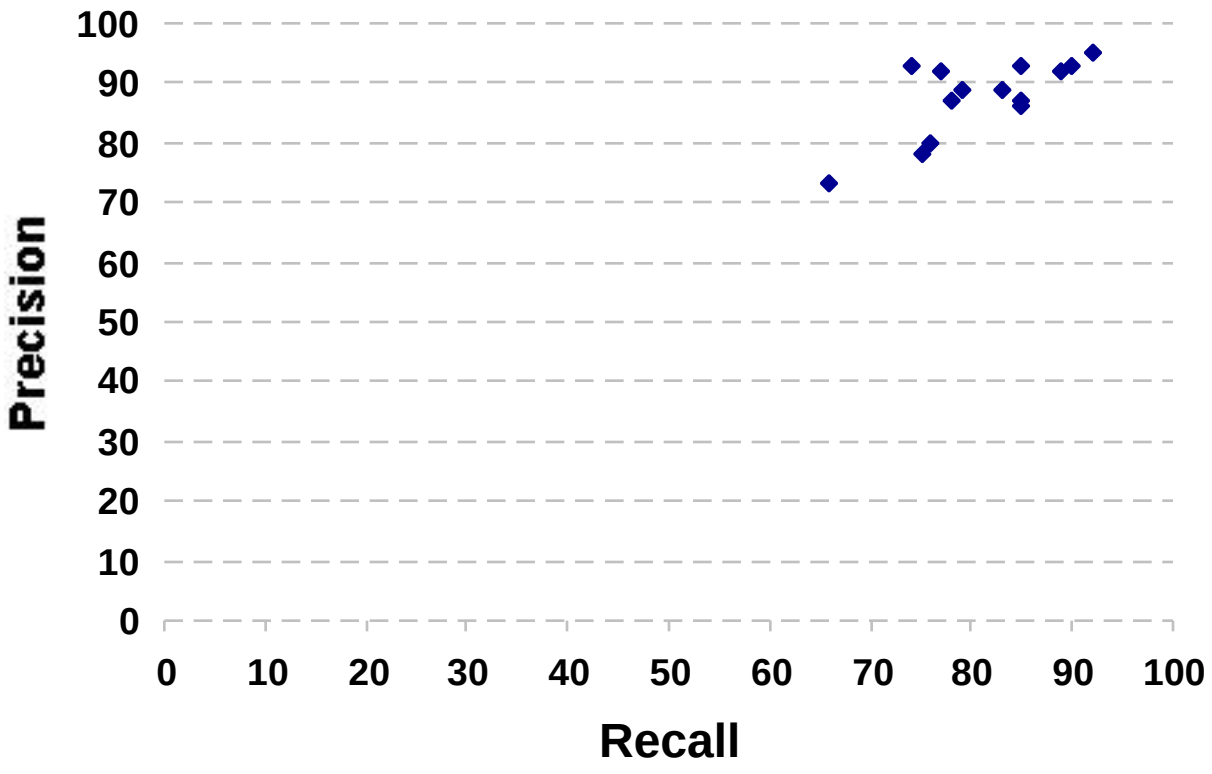


Distribution of NE TIMEX tag elements



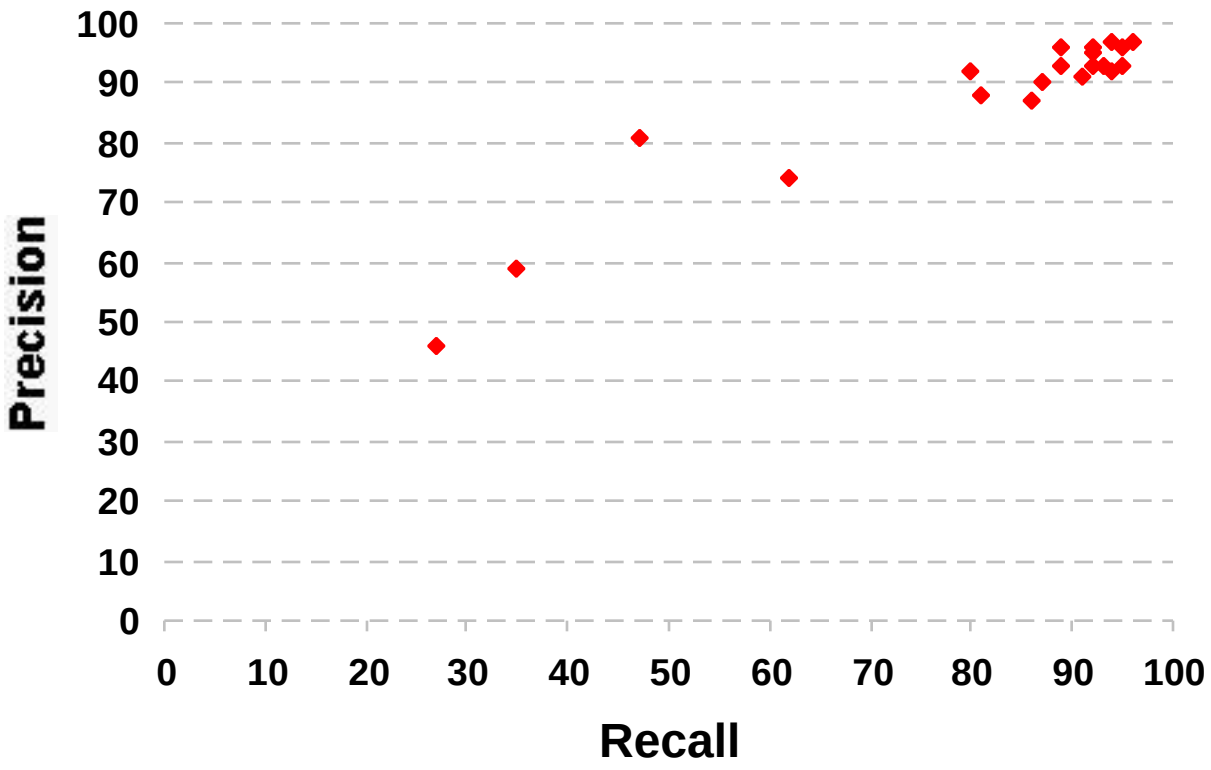
MUC

NE Results Overall



MUC 7

NE Overall Results

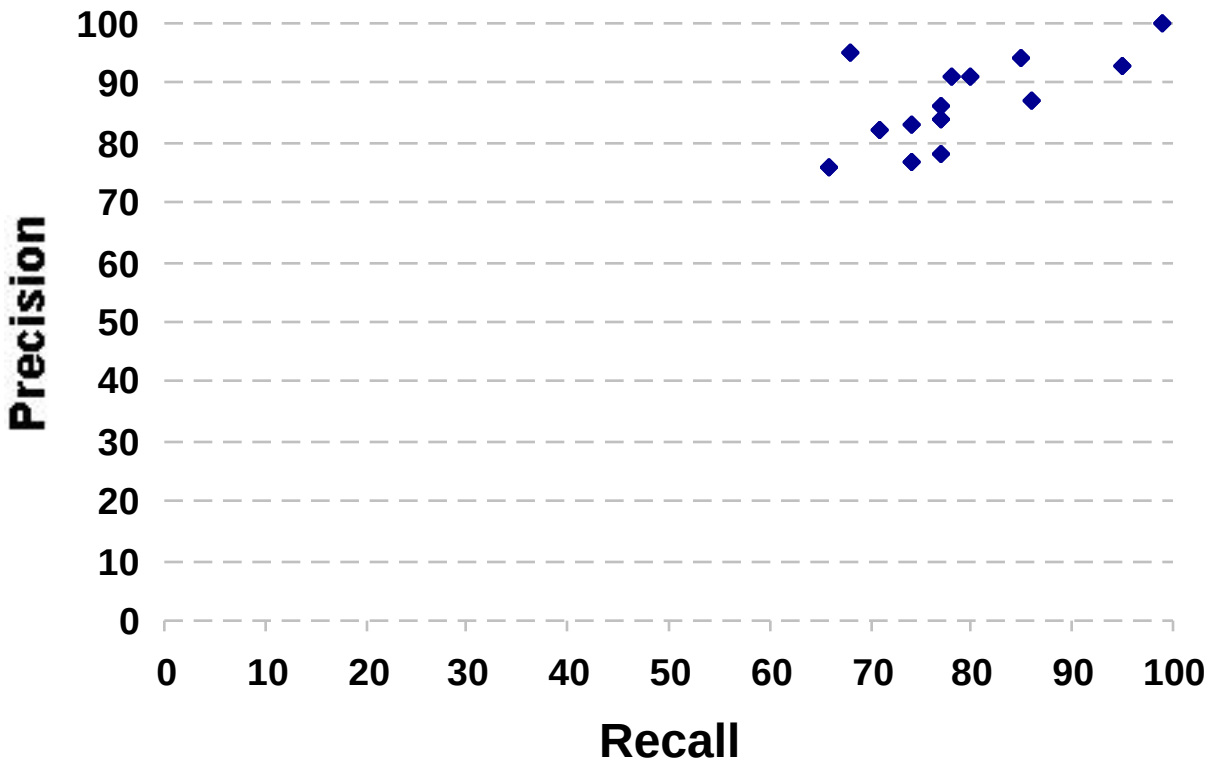


MUC

NE Results on Walkthrough

- **Number of tags in answer key:**
  - 52 Enamex
  - 1 Numex
  - 14 Timex
- **System scoring:**
  - Common mistakes on TIMEX: missed ***early Thursday morning***, ***within six months***
  - Common mistakes on ENAMEX: missed ***Globo, MURDOCH, Xichang; Long March as TIMEX, ENAMEX***
  - One site missed only one entity in whole document ***within six months***

NE Results on Walkthrough

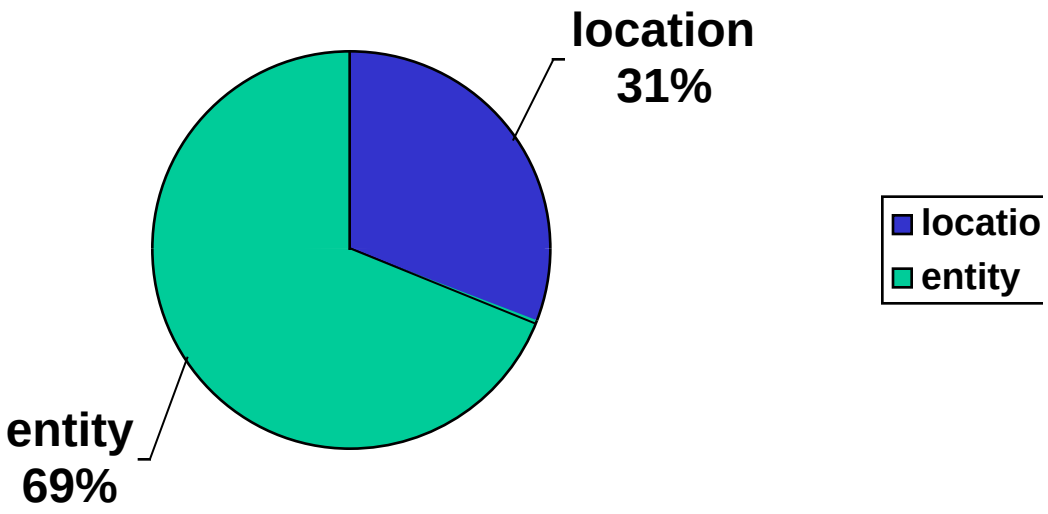


MUC

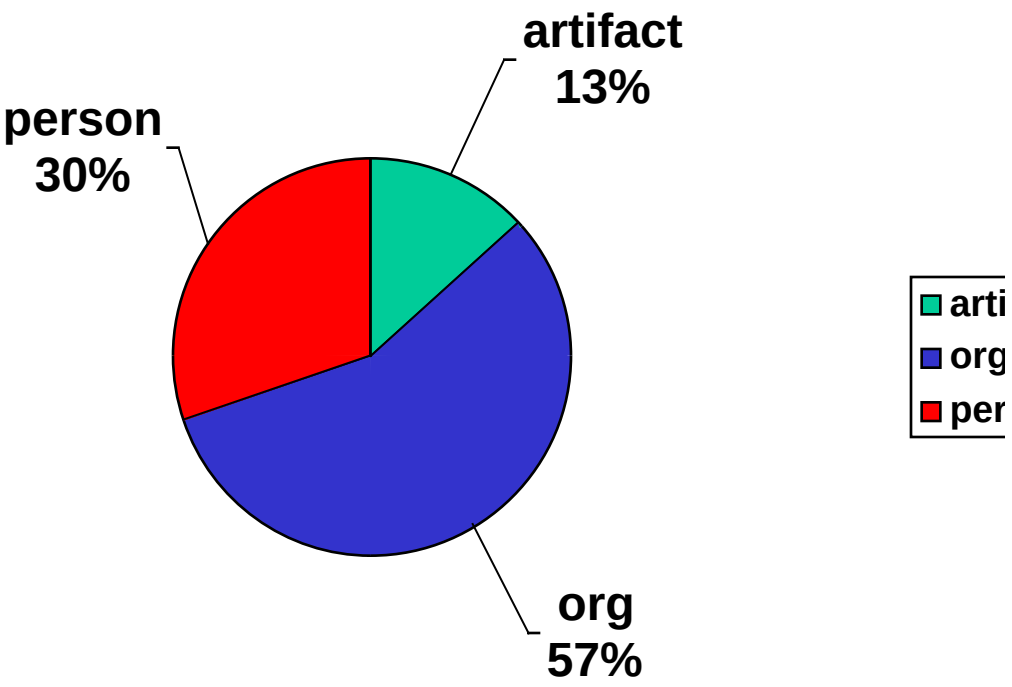
INFORMATION EXTRACTION: TEMPLATE ELEMENT (TE)

- **TEs are independent or neutral wrt scenario: generic objects and slots.**
- **Separates domain-independent from domain-dependent aspects of extraction.**
- **Consists of object types defined for a given scenario, but unconcerned with relevance.**
- **Answer key contains objects for all organizations, persons, and vehicle artifacts mentioned in the texts, whether relevant to scenario or not.**

TE Objects

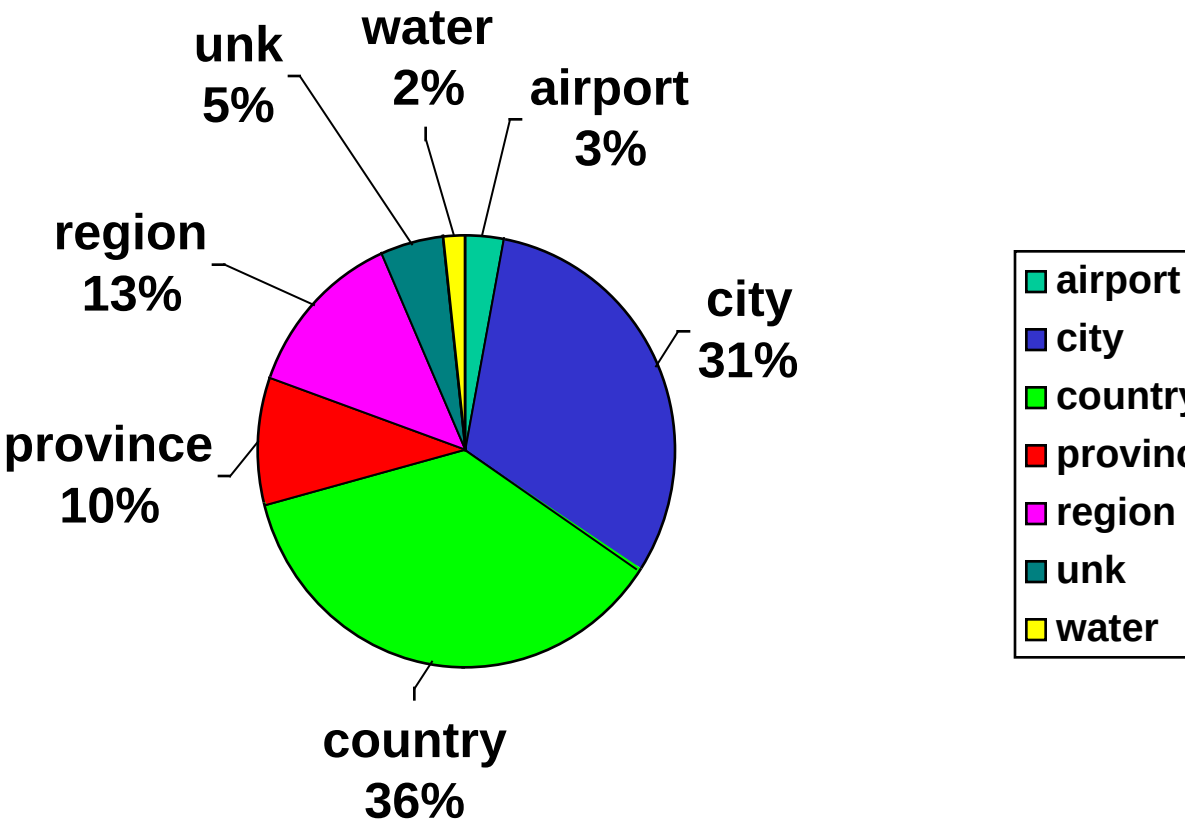


TE ENT_TYPE Distribution



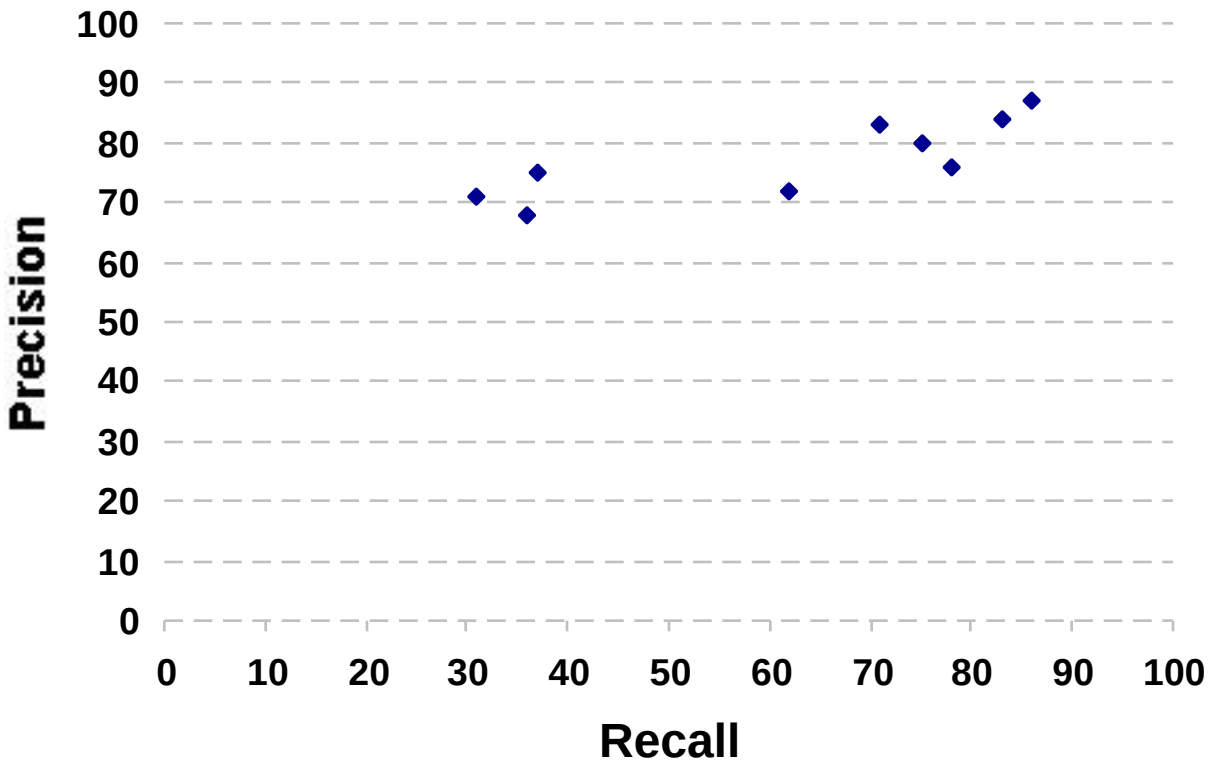
MUC

TE LOCALE_TYPE Distribution



MUC

TE Results Overall

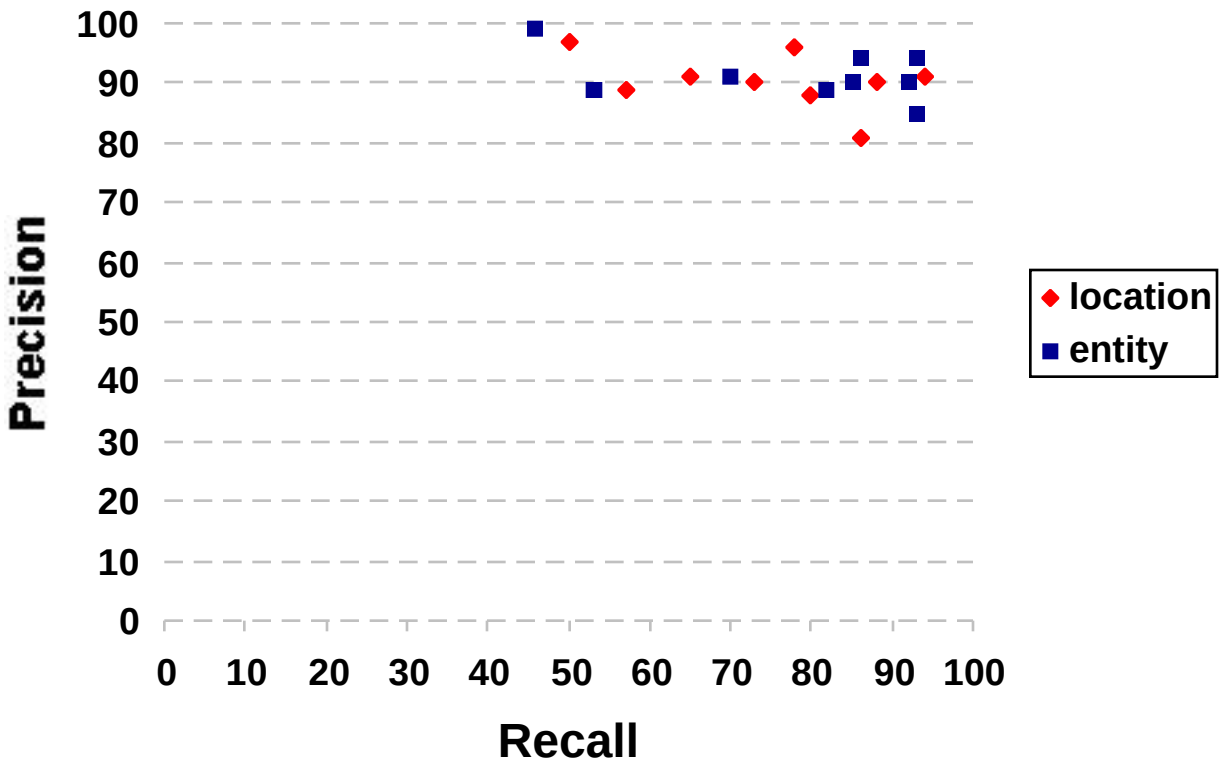


MUC

TE Overall Results

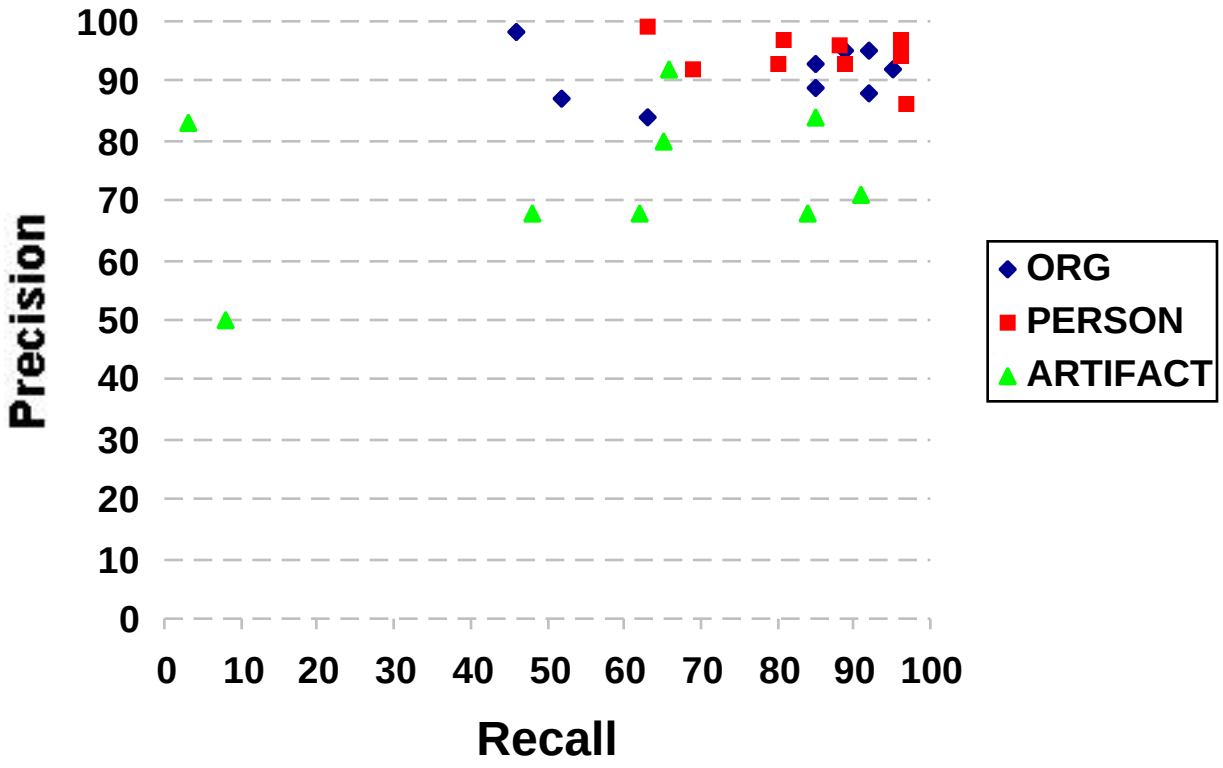
MUC

TE Results for TE Objects



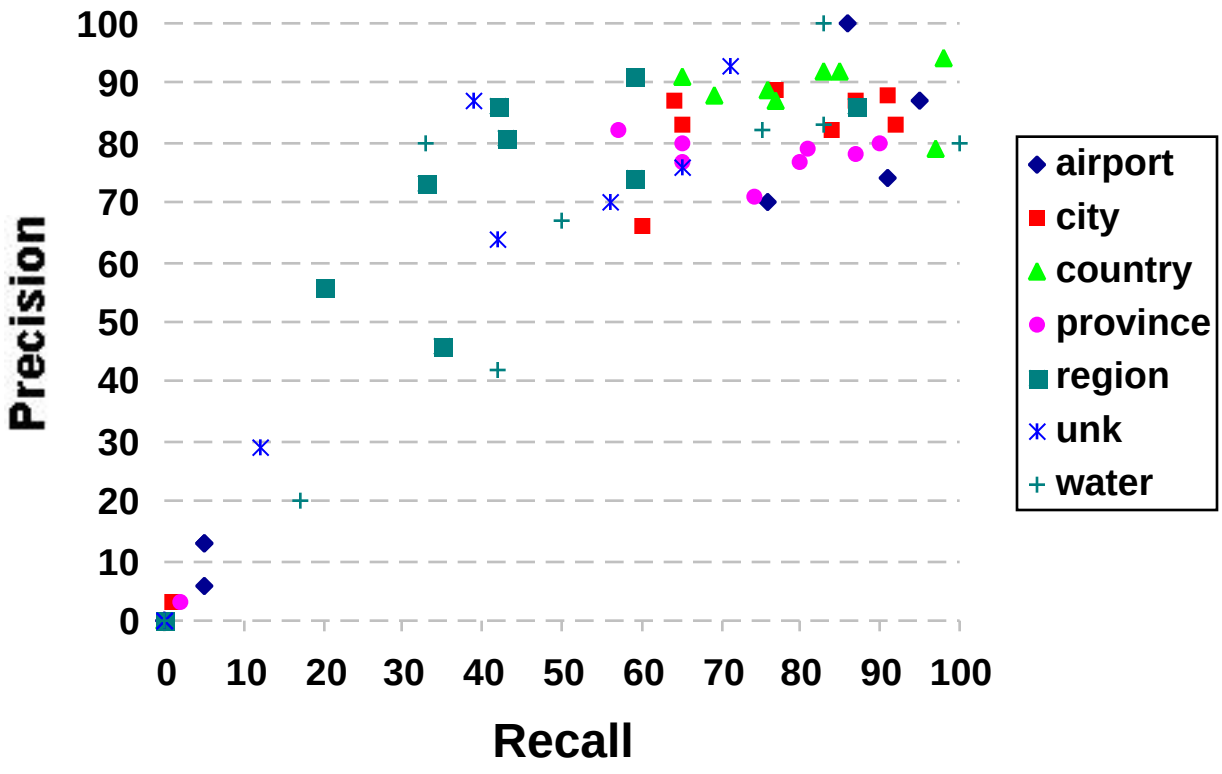
MUC

TE Results for ENT_TYPE



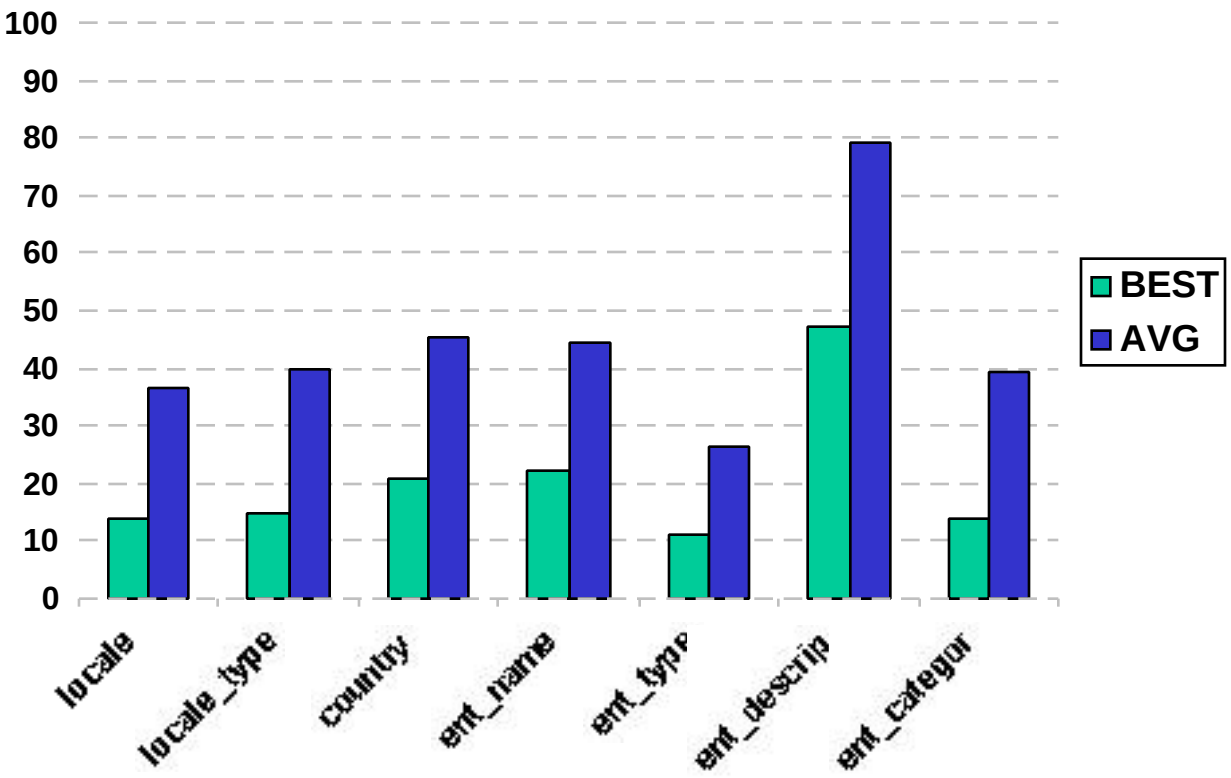
MUC

TE Results for LOCALE_TYPE



MUC

TE ERR Results by Slot

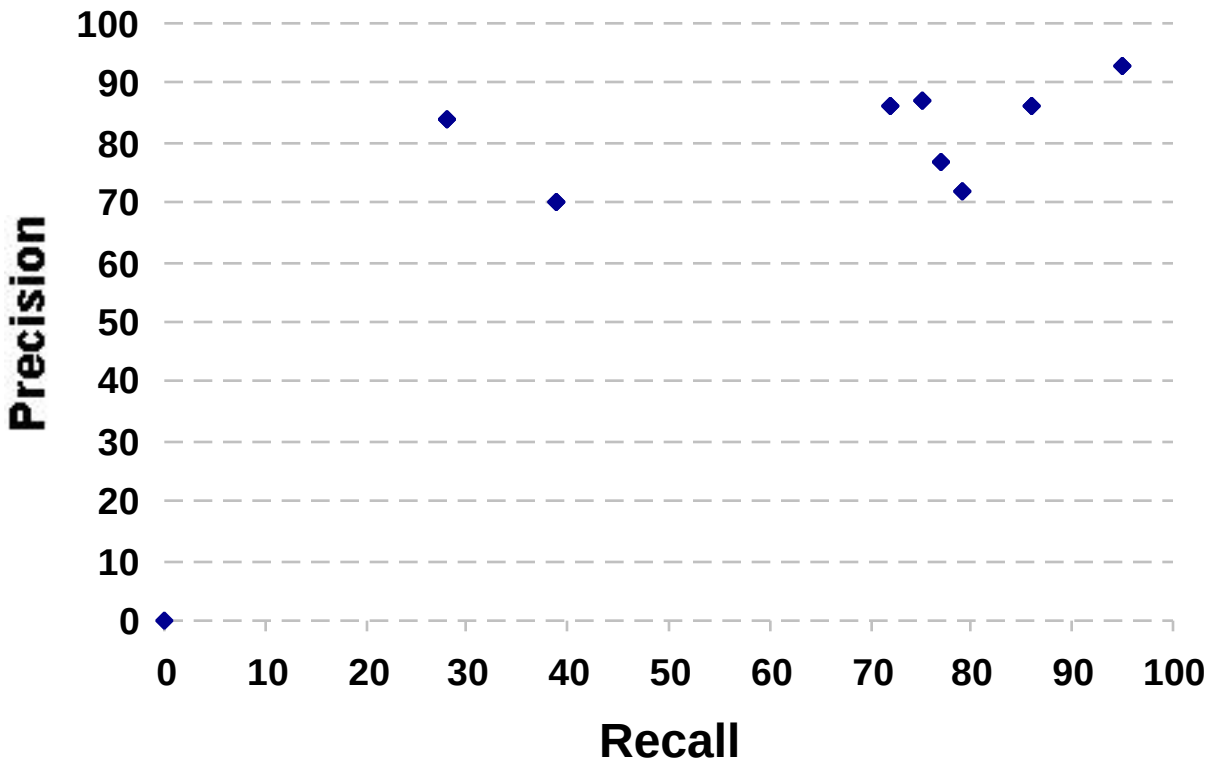


MUC

TE Results on Walkthrough Article

- **Omissions or errors in ENT_DESCRIPTOR (span of descriptor, descriptor itself)**
- **Omissions NAME slot: aliases missed (*China Great Wall, News Corp*.)**
- **LOCALE_TYPE (PROVINCE / COUNTRY / CITY)**
- **ENT_CATEGORY: ORG_OTHER vs. ORG_CO (*Space Transportation Association*)**
- **ORG as PERSON (Intelsat); PERSON as ORG (Murdoch)**

TE Results on Walkthrough

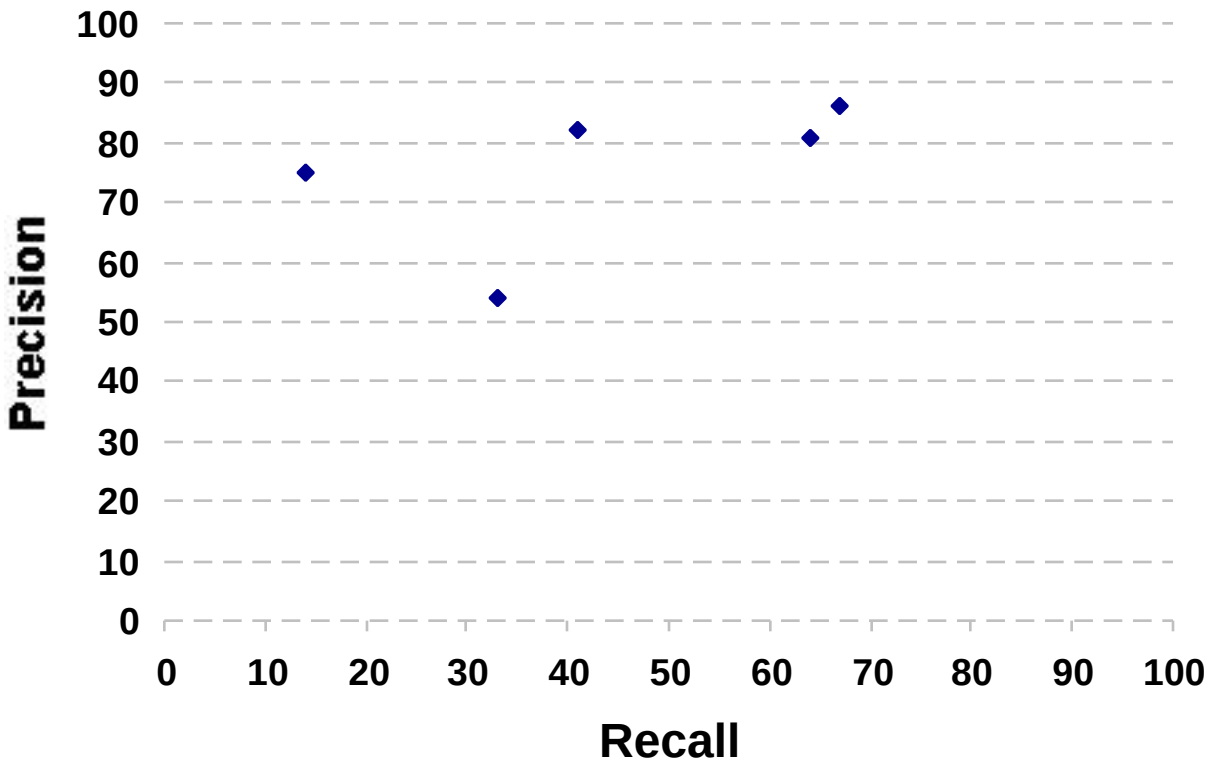


MUC

Template Relations Task (TR)

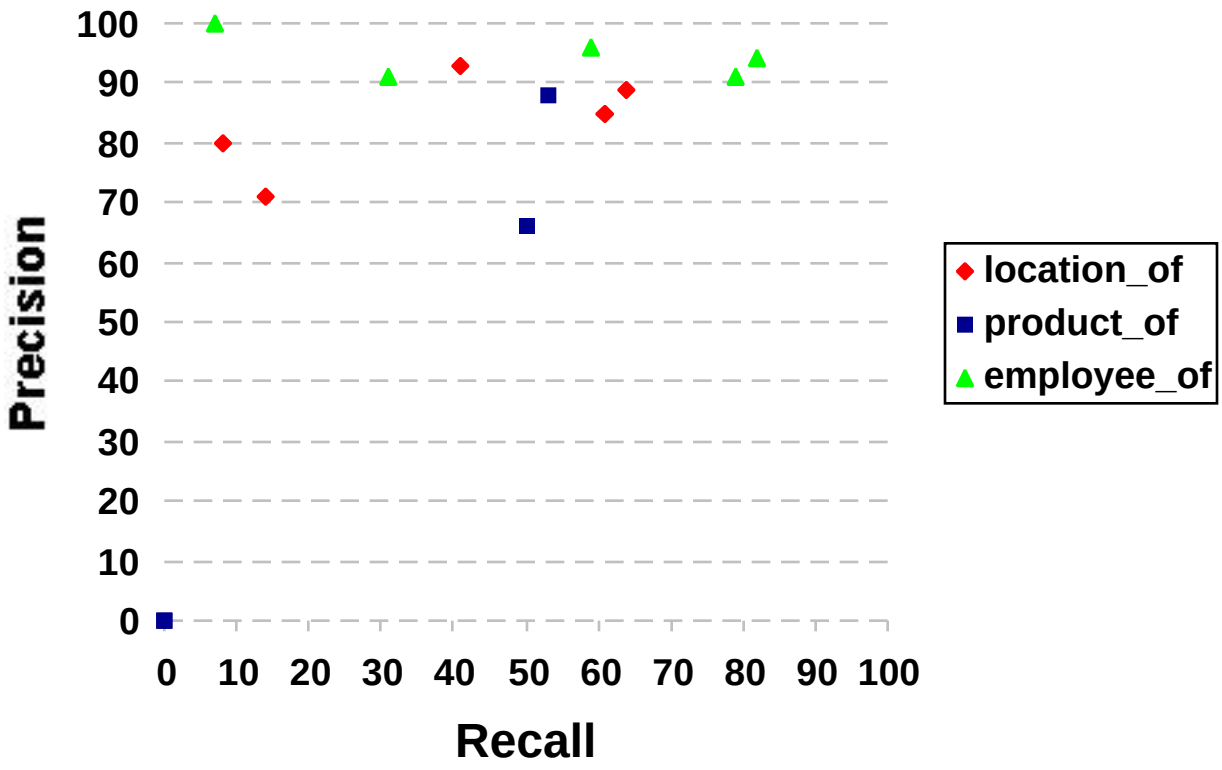
- New task for MUC-7.
- TRs express domain-independent relationships between entities, as compared with TEs which identify entities themselves.
- TR uses LOCATION_OF, EMPLOYEE_OF, and PRODUCT_OF relations.
- Answer key contains entities for all organizations, persons, and artifacts that enter into these relations, whether relevant to scenario or not.

TR Overall



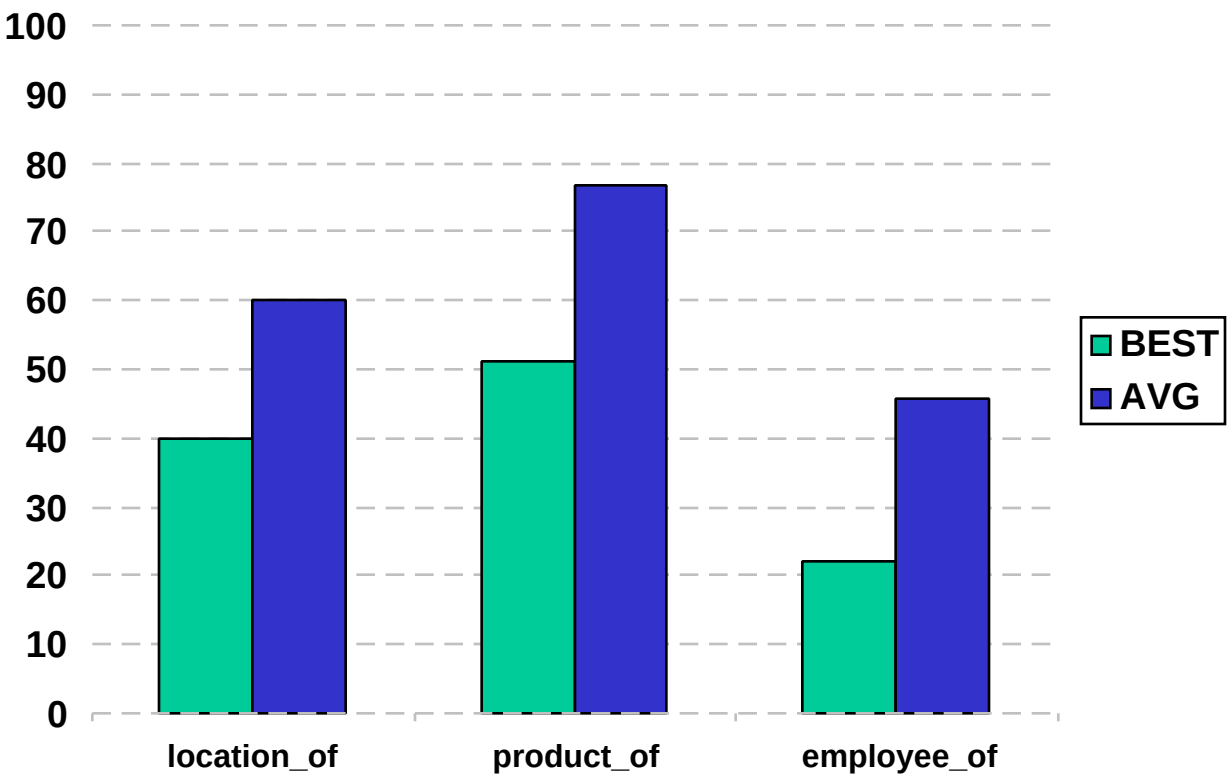
MUC

TR Results by Relation



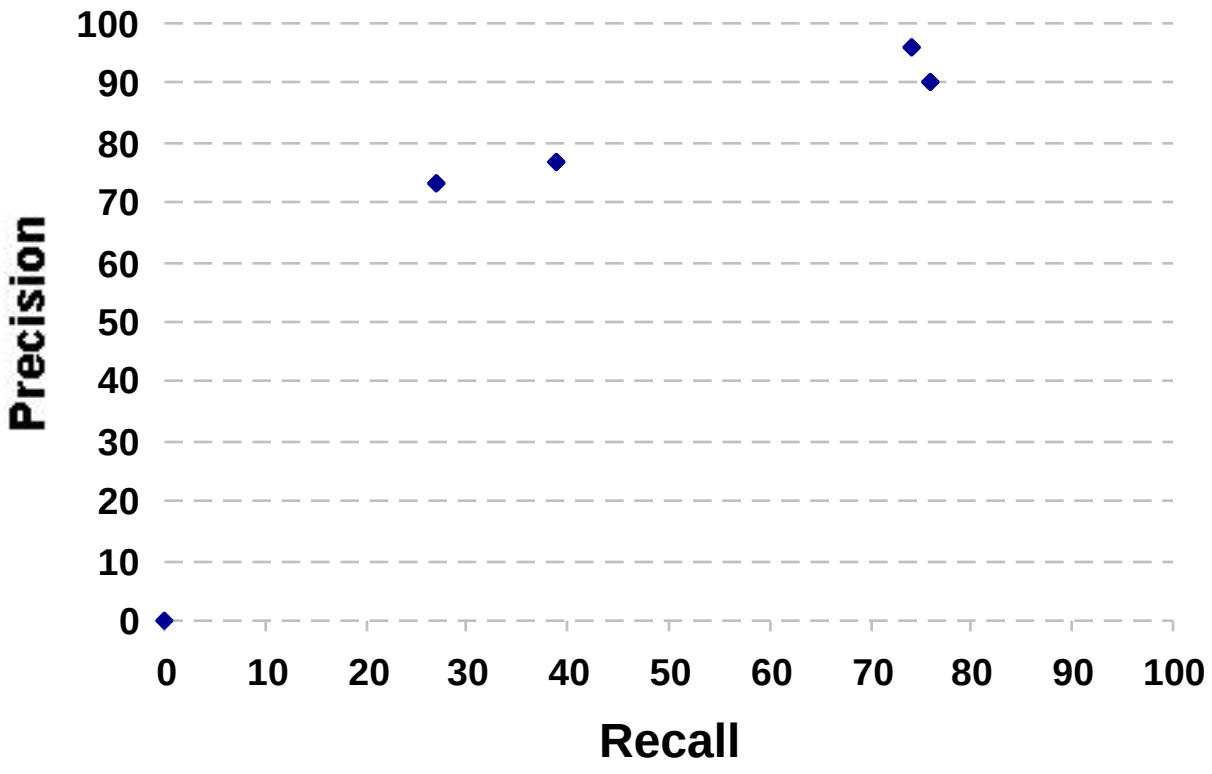
MUC

TR Error Scores



MUC

TR Results on Walkthrough



MUC

Scenario Template (ST)

- STs express domain and task-specific entities and relations. Similar to MUC-6 template.
- ST tests portability to new extraction problem; short time frame for system preparation (1 month)
- Scenario concerns vehicle launch events.
  - Template consists of one high-level event object (LAUNCH_EVENT) with 7 slots, including 2 relational objects (VEHICLE_INFO, PAYLOAD_INFO), 3 set fills (MISSION_TYPE, MISSION_FUNCTION, MISSION_STATUS), and 2 pointers to low-level objects (LAUNCH_SITE, LAUNCH_DATE)

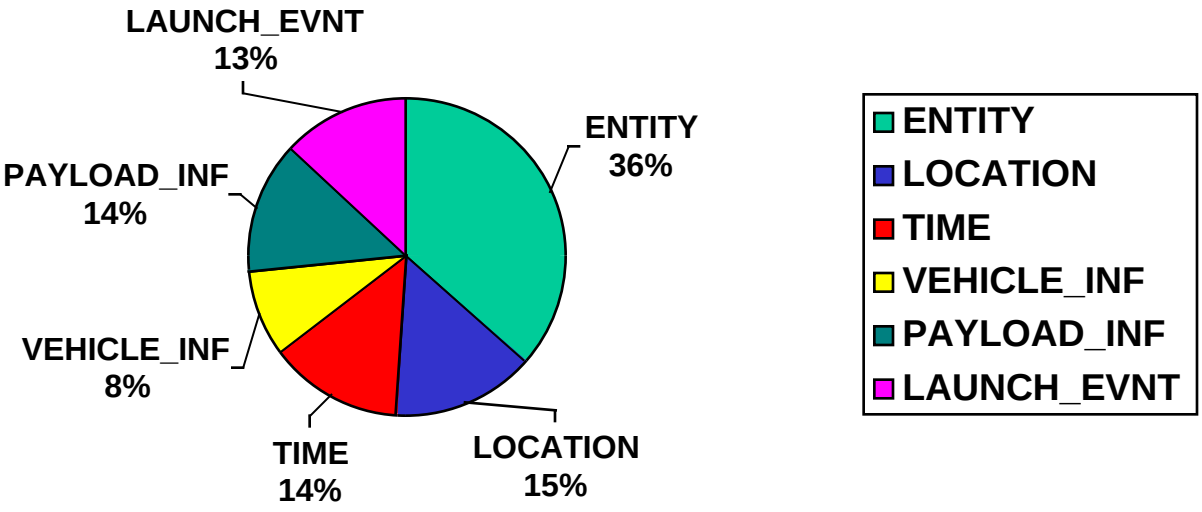
Scenario Template (con't)

  - Relational objects have pointers to Template Elements, set-fills.
  - Set fills require inferences from the text.
- **Test set statistics: 63/100 documents relevant to the scenario.**

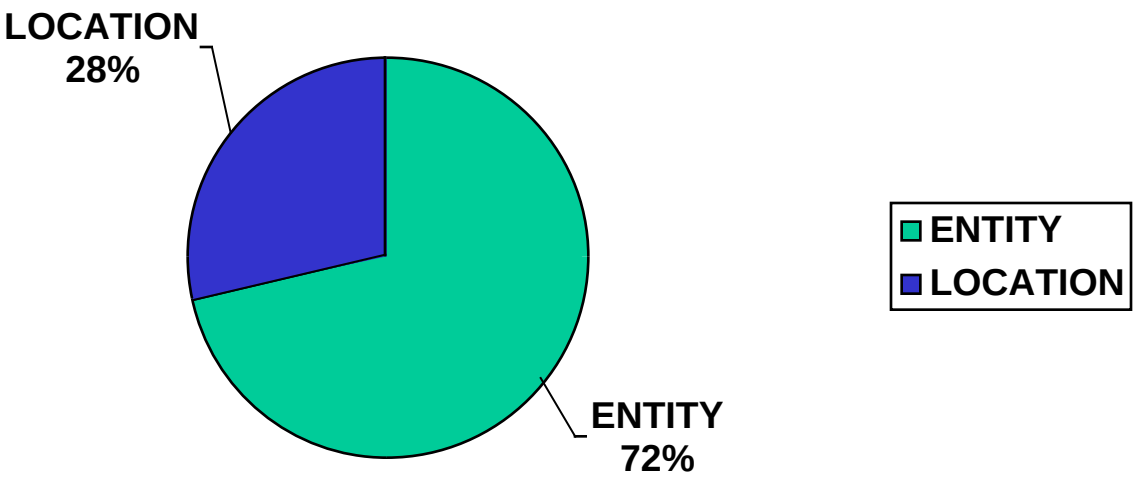
ST Overall Results

- **Systems scored points lower (F-measure) on ST than on TE.**
- **Interannotator variability (measured on all articles) was between 85.15 and 96.64 on the F-measures.**
- **Document-level relevance judgments (Text Filtering scores), were similar to those for MUC-6, although percentage of relevant articles in text set was greater.**

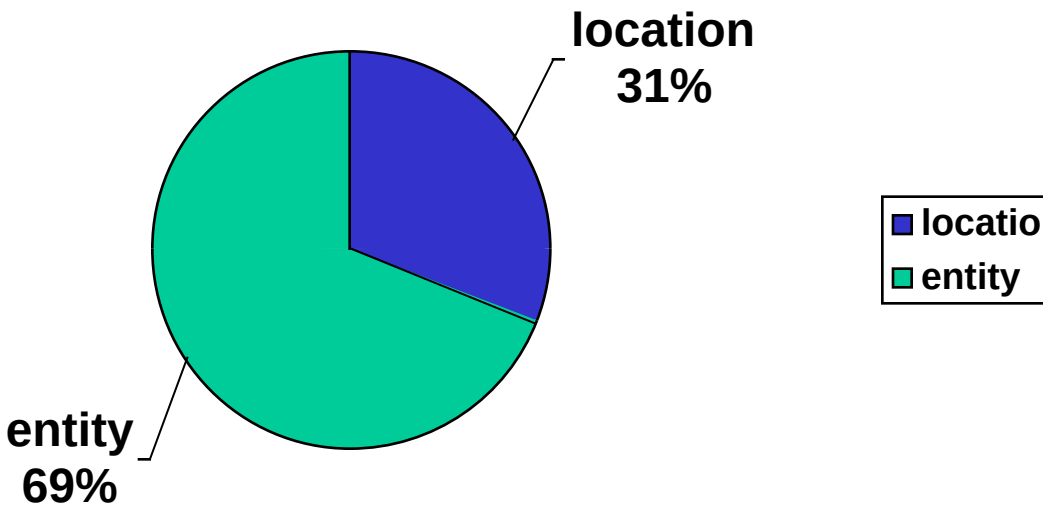
ST Slot Distribution



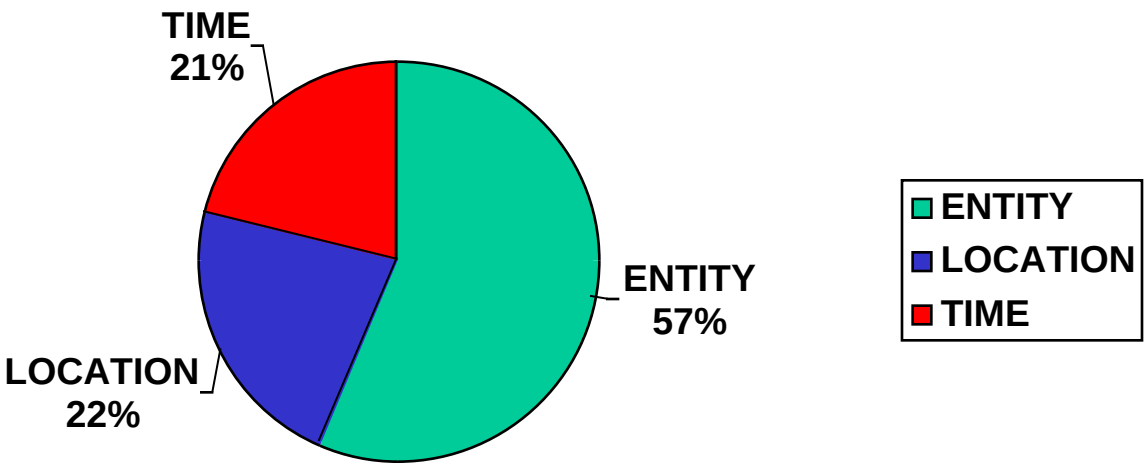
ST Template Elements



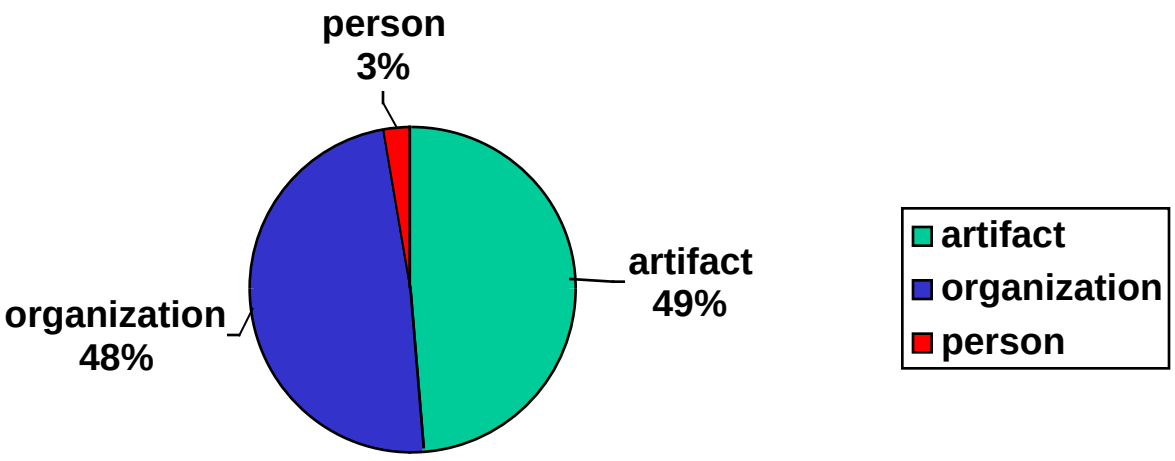
TE Objects



ST Template Elements

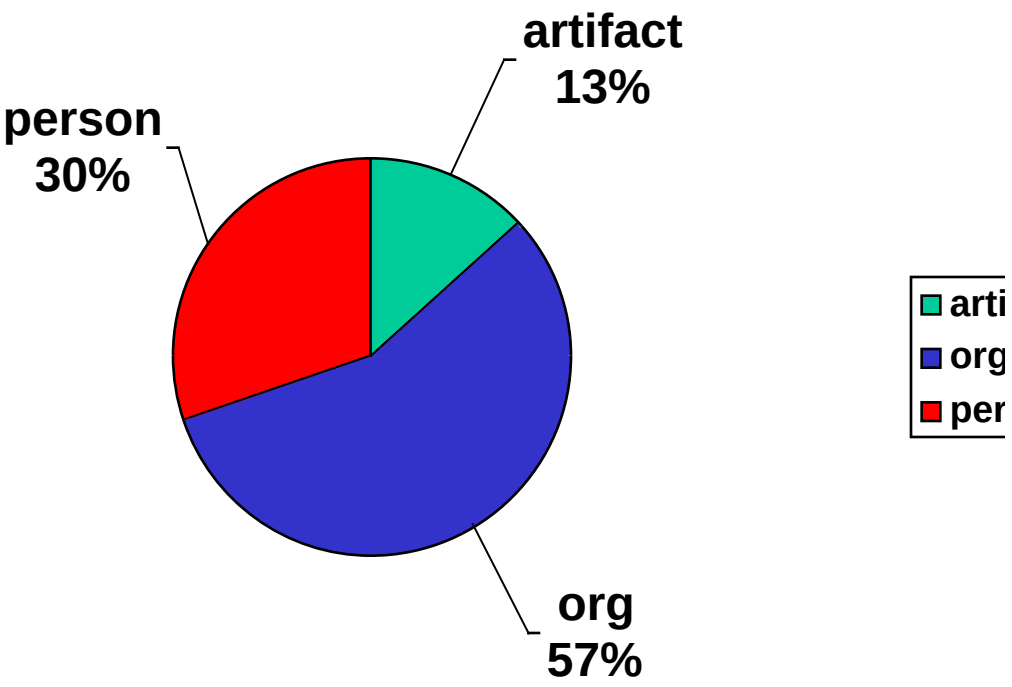


ST ENT_TYPE Distribution



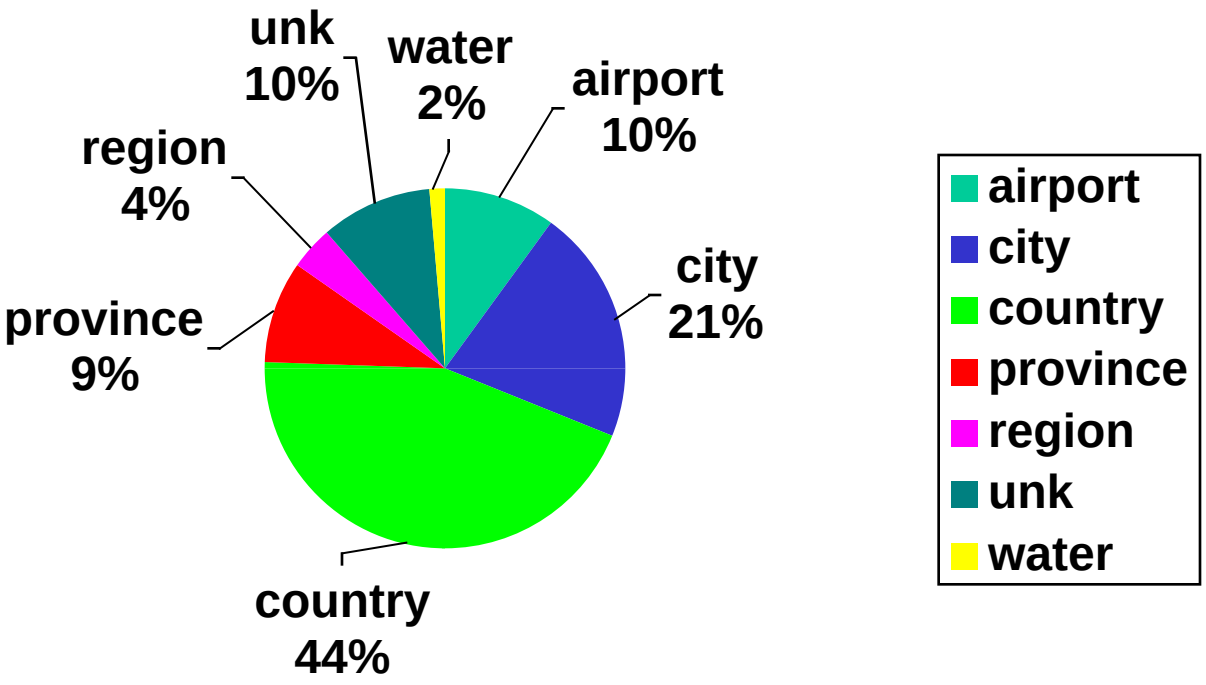
MUC

TE ENT_TYPE Distribution



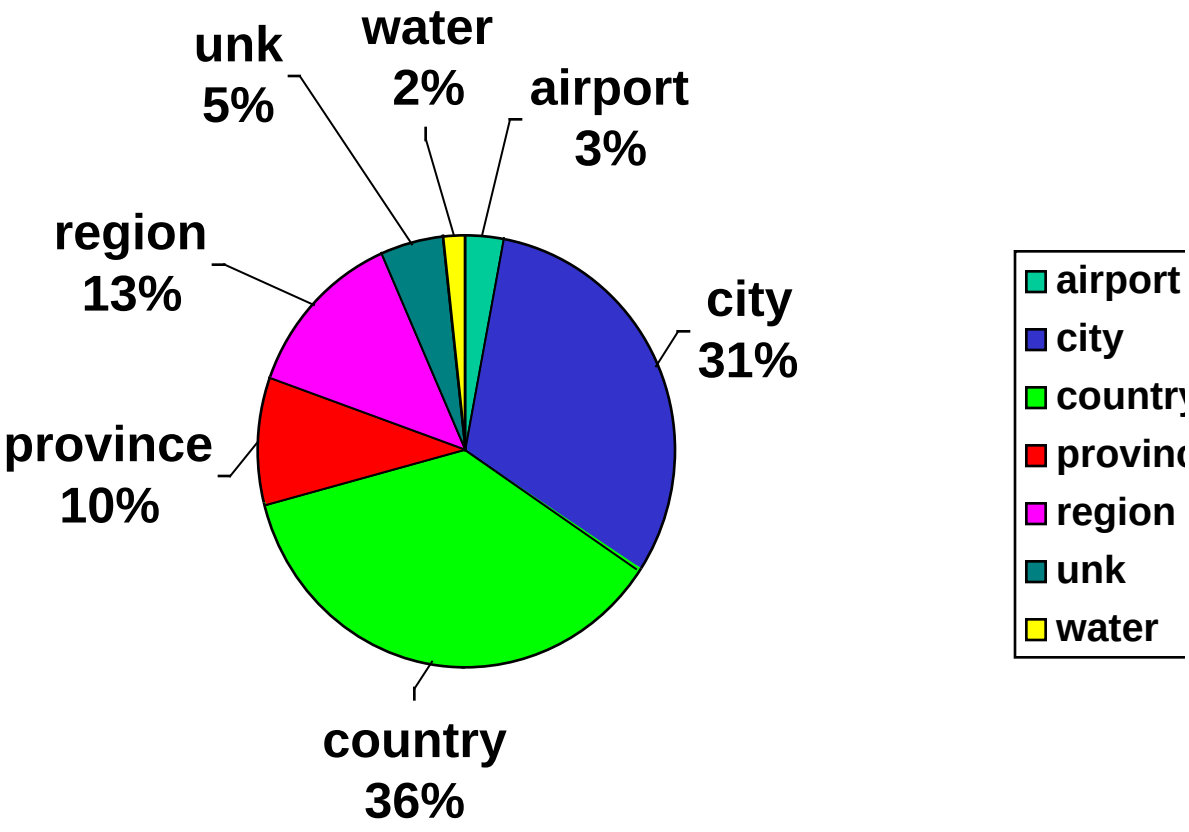
MUC

ST LOCALE_TYPE Distribution



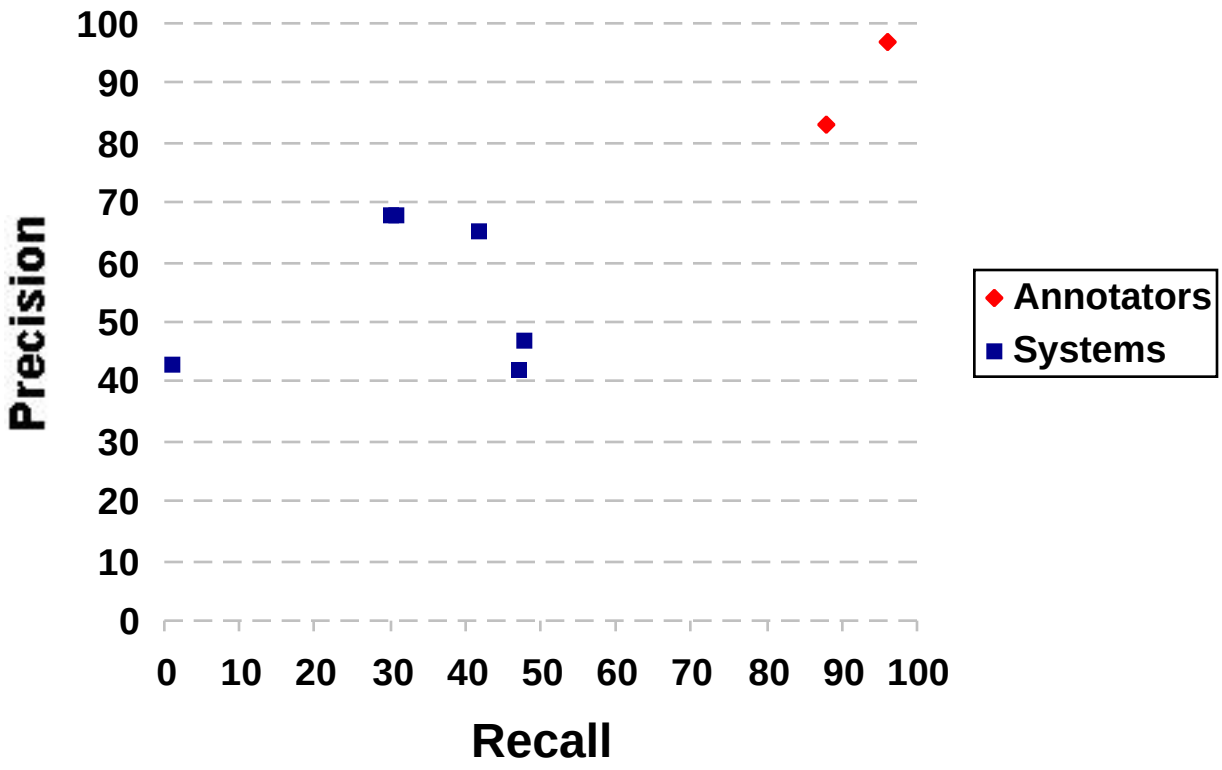
MUC

TE LOCALE_TYPE Distribution



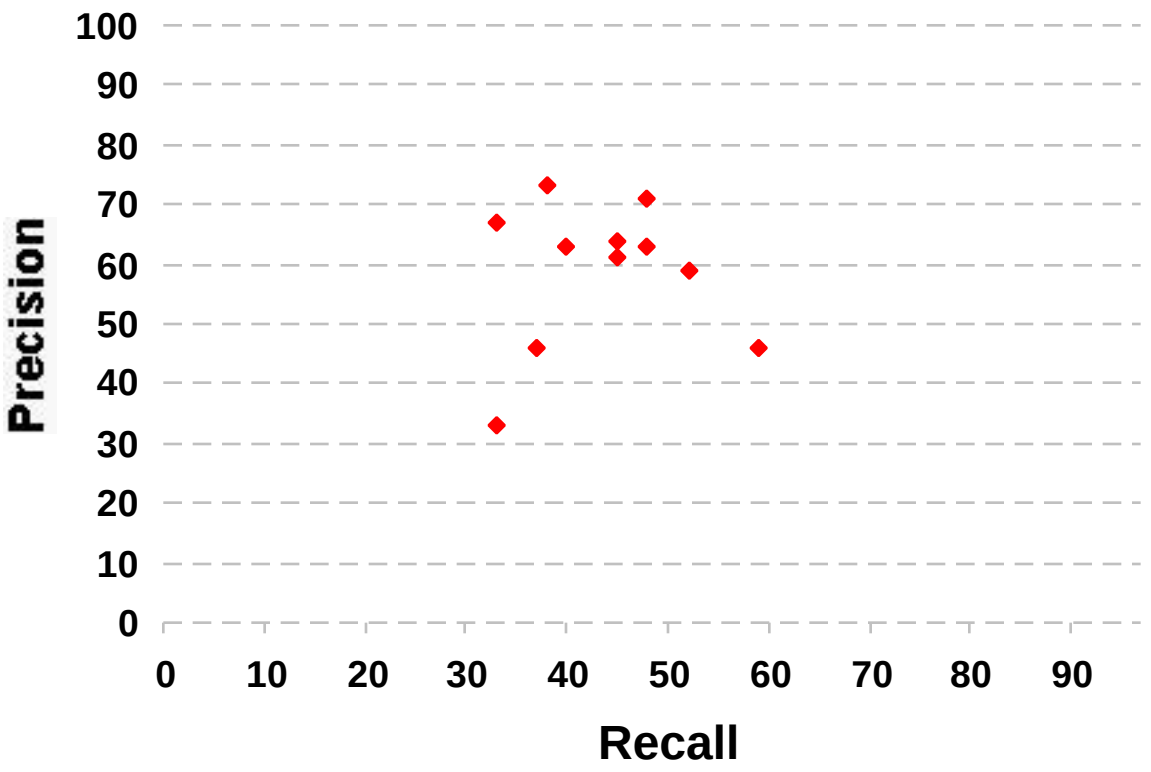
MUC

ST Results Overall



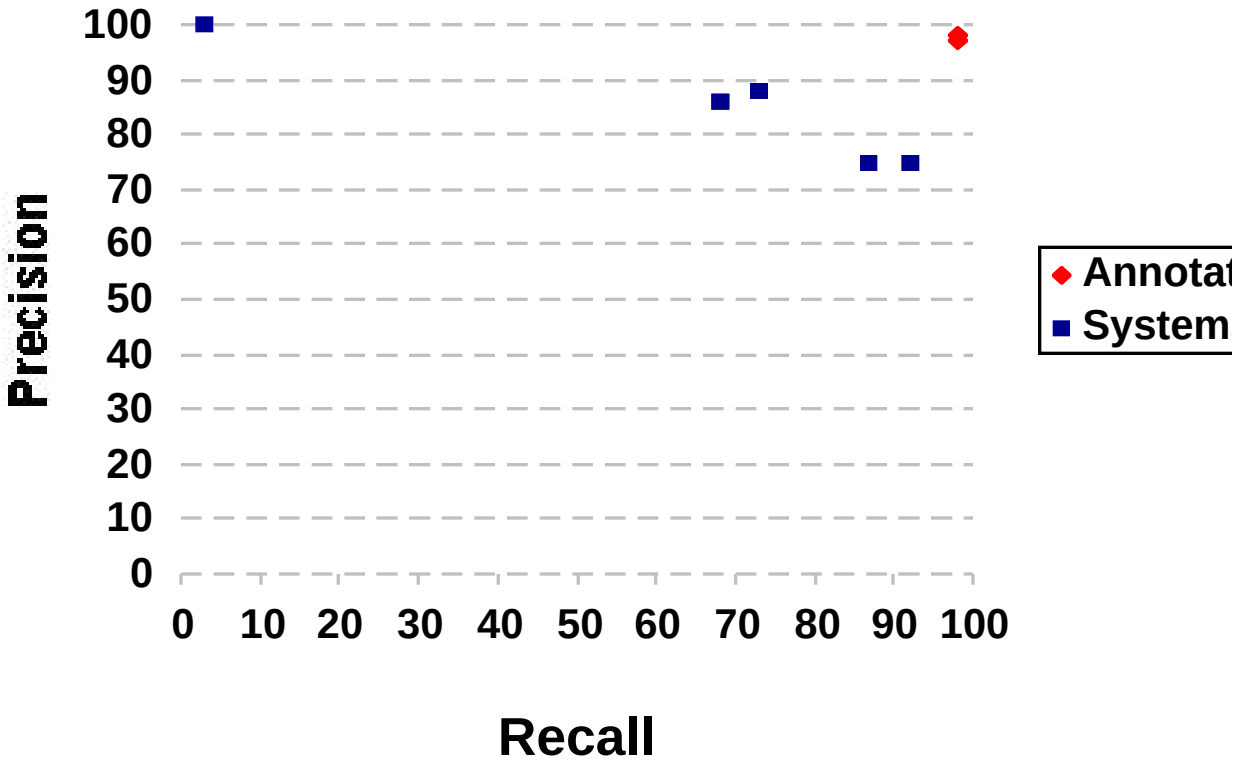
MUC

ST Overall Results



MUC

ST Results for Text Filtering



MUC

ST Results on Walkthrough

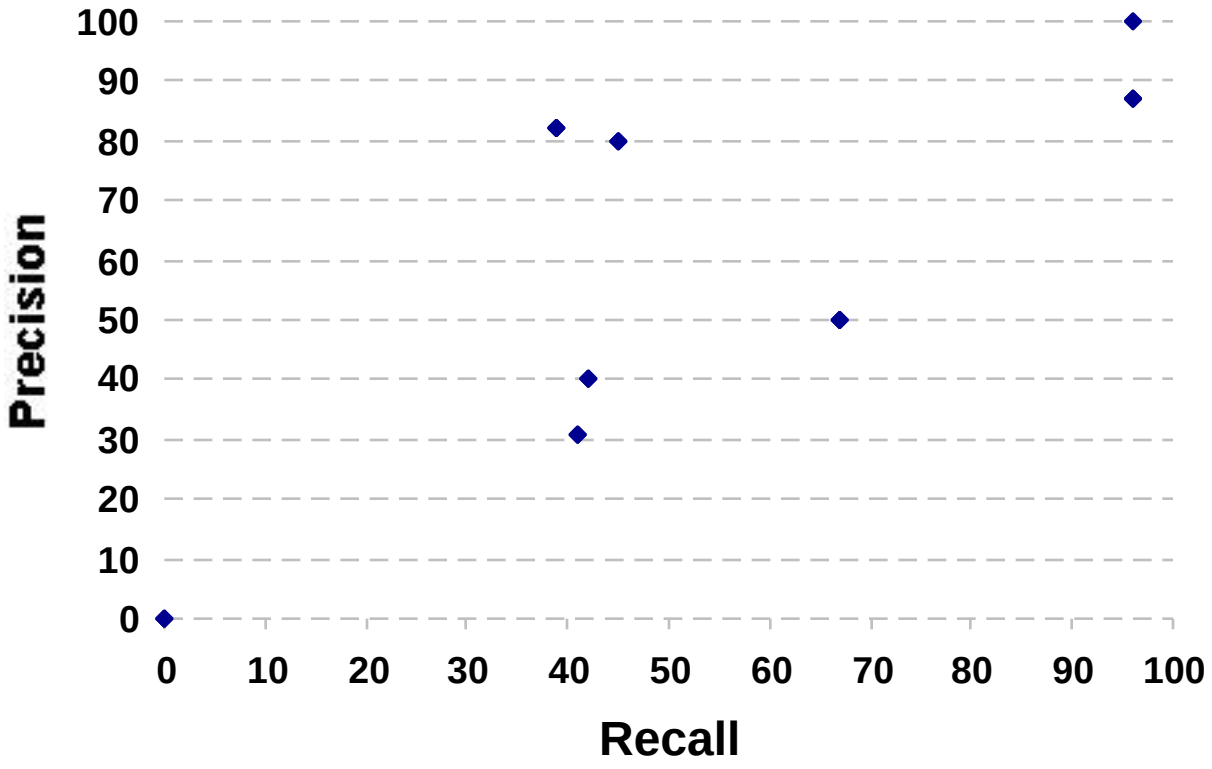
MUC-7

- F-measures for annotators: 98.13, 91.40
  ERR for annotators: 4%, 14%
- F-Measures for systems(all-1): 35.60-41.18
  ERR for systems (all-1): 56-75%

MUC-6:

- ERR for annotators: 8%
- ERR for systems: 30-89%

ST Results on Walkthrough

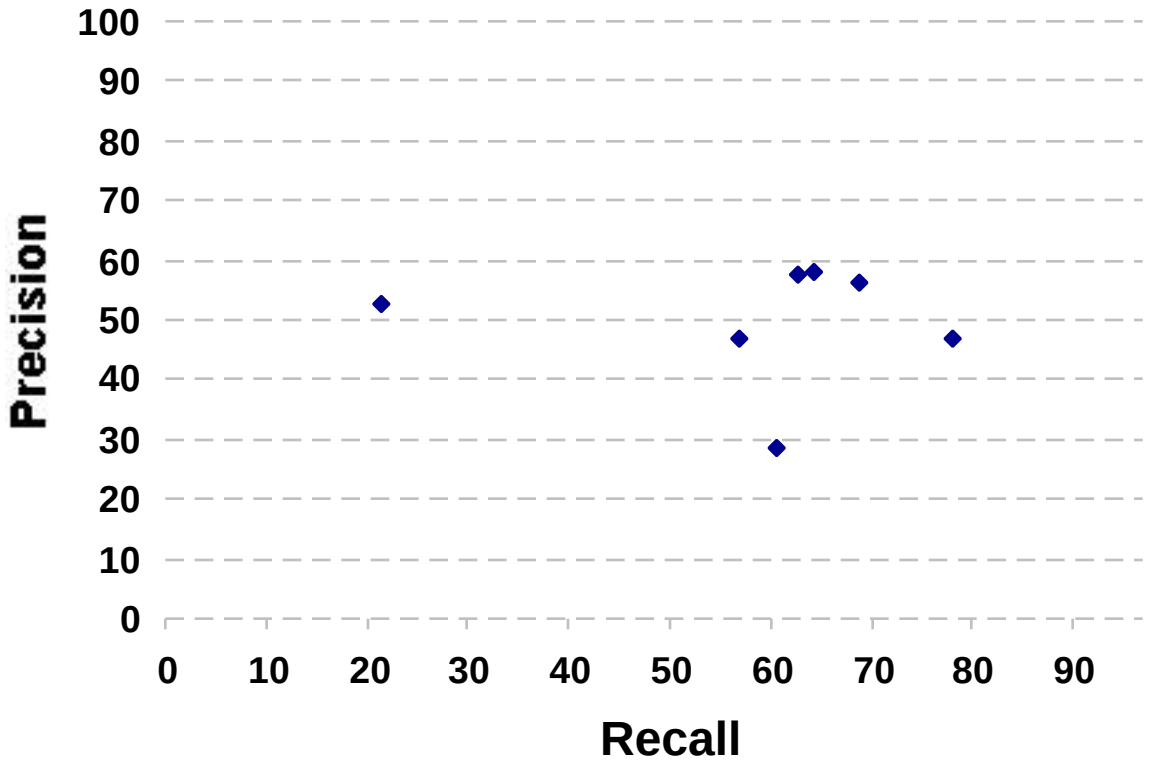


MUC

Coreference Task (CO)

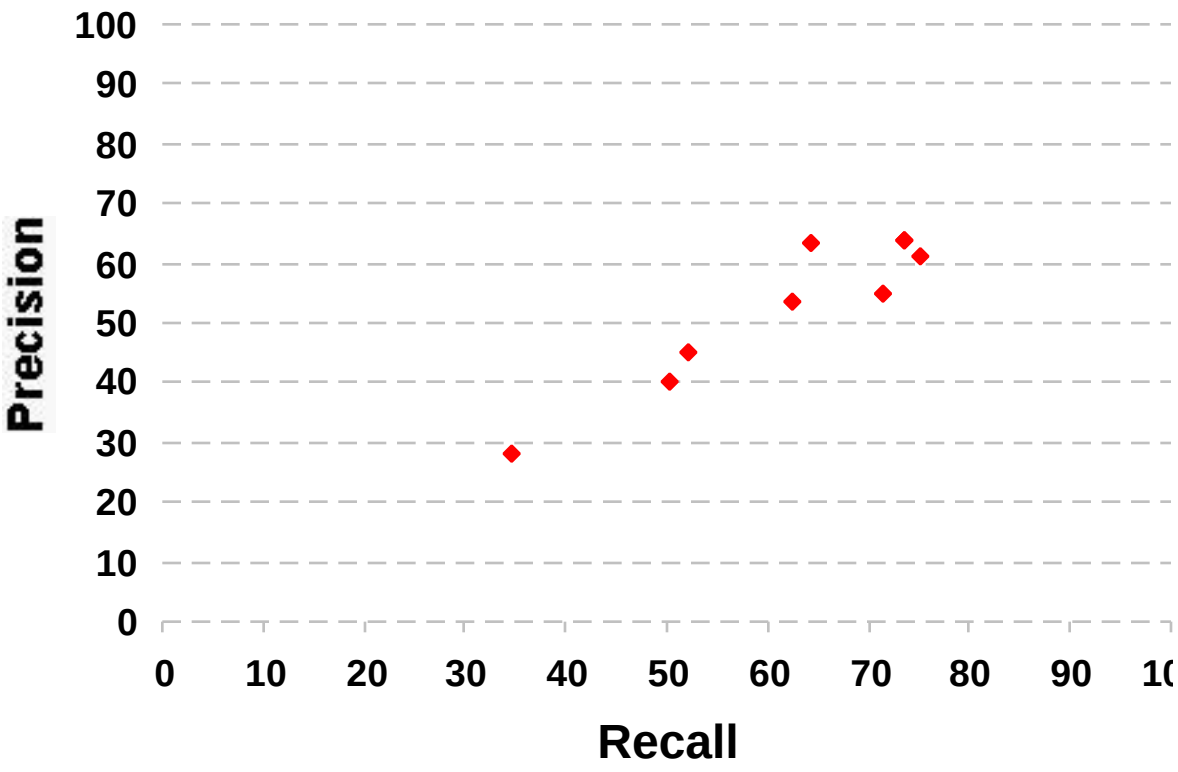
- Capture information on coreferring expressions: all mentions of a given entity, including those tagged in NE, TE tasks.
- Focused on the IDENTITY (IDENT) relation: symmetrical and transitive relation, equivalence classes used for scoring.
- Markables: Nouns, Noun Phrases, Pronouns

CO Results Overall



MUC

CO Overall Results



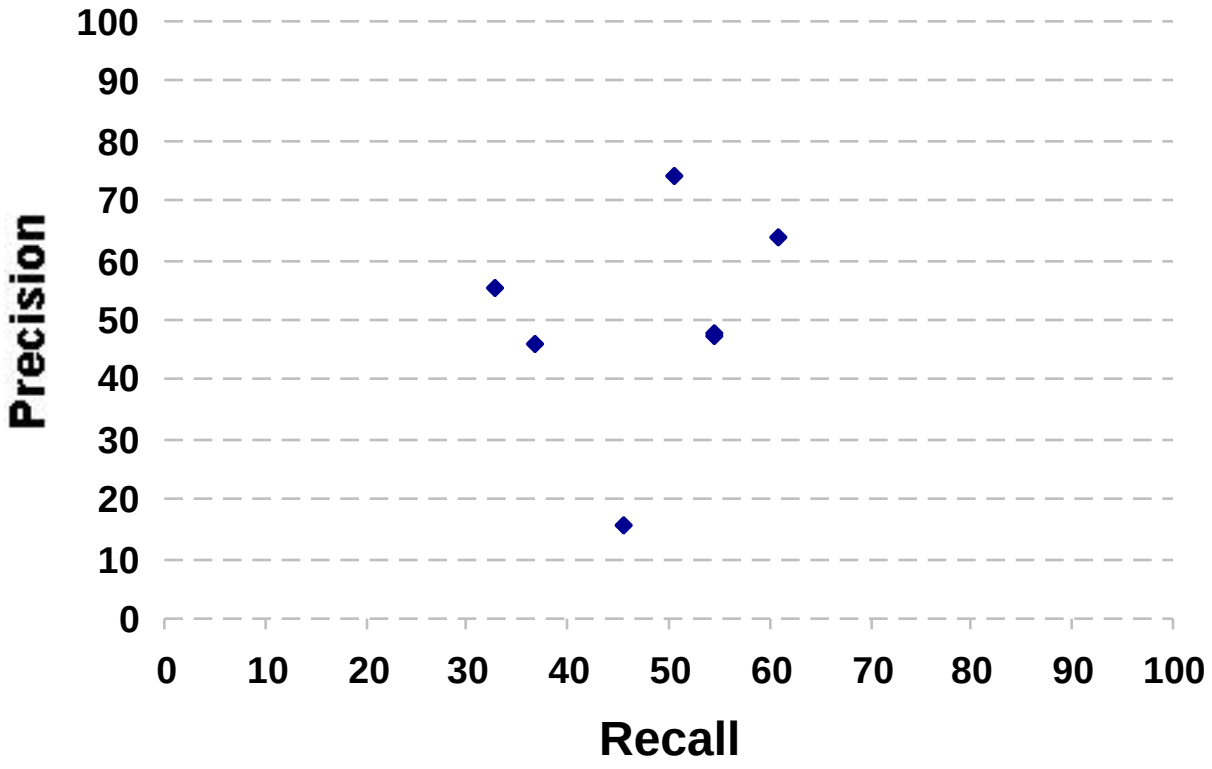
MUC 6

CO Results for Walkthrough

- **Walkthrough article non-relevant for ST**
- **F-measures range from 23.2-62.3%**
- **Missing:**
  - **Dates: Thursday, Sept. 10**
  - **Money: $30 Million**
  - **Unusual Conjunctions: *GM, GE PROJECTS***
  - **Miscellaneous:**

    Thursday's meeting, agency's meeting,
    FCC's allocation…, transmissions from satellites to earth stations
    US satellite industry, federal regulators
    satellite downlinks,
    NEEDED AIRWAVES.

CO Results on Walkthrough



MUC