Trends in Information Extraction

Qi Zhang 张奇

Fudan University

Information Extraction

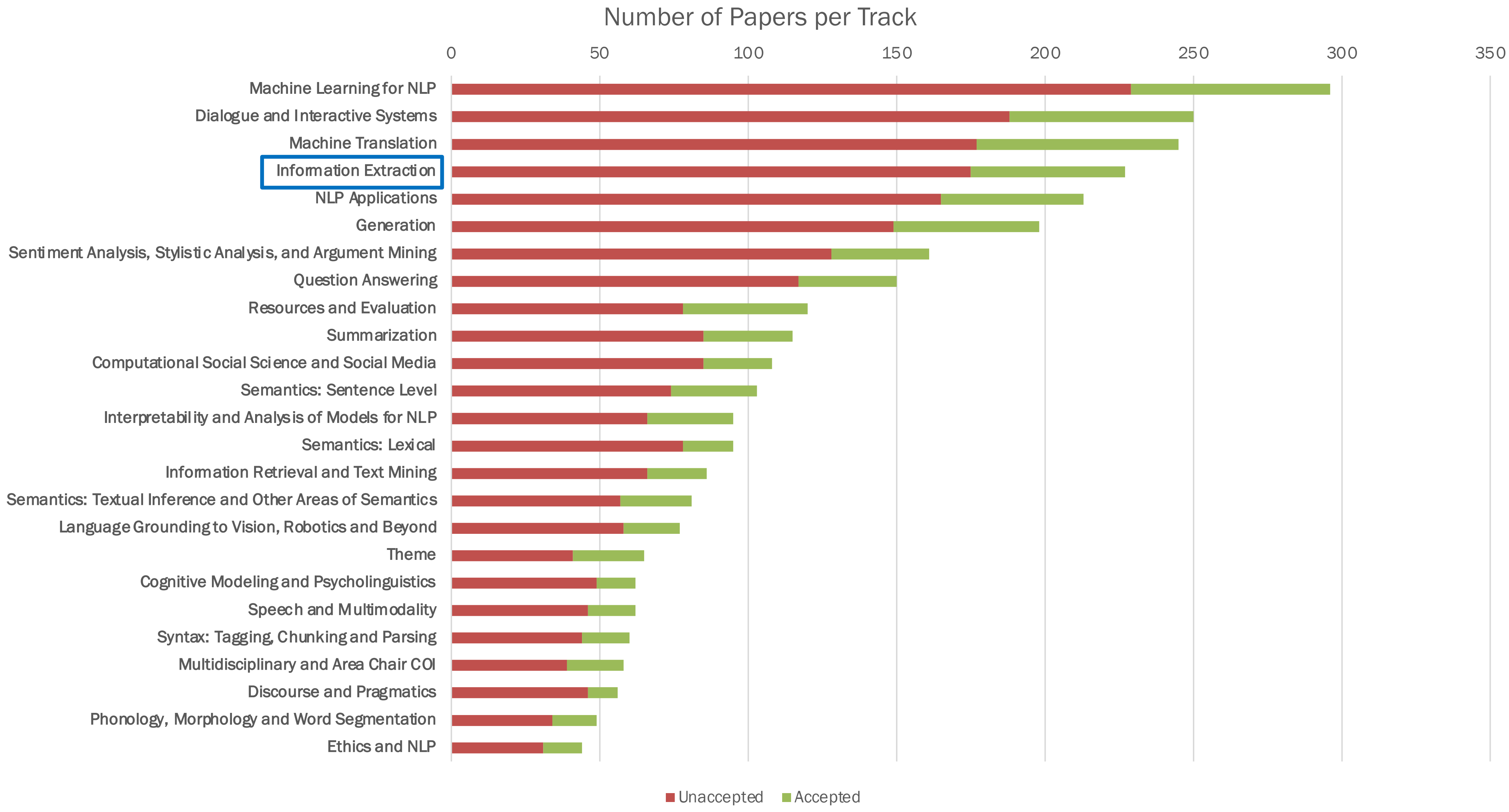
Information extraction (**IE**) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents and other electronically represented sources.

Information extraction dates back to the late 1970s in the early days of NLP.

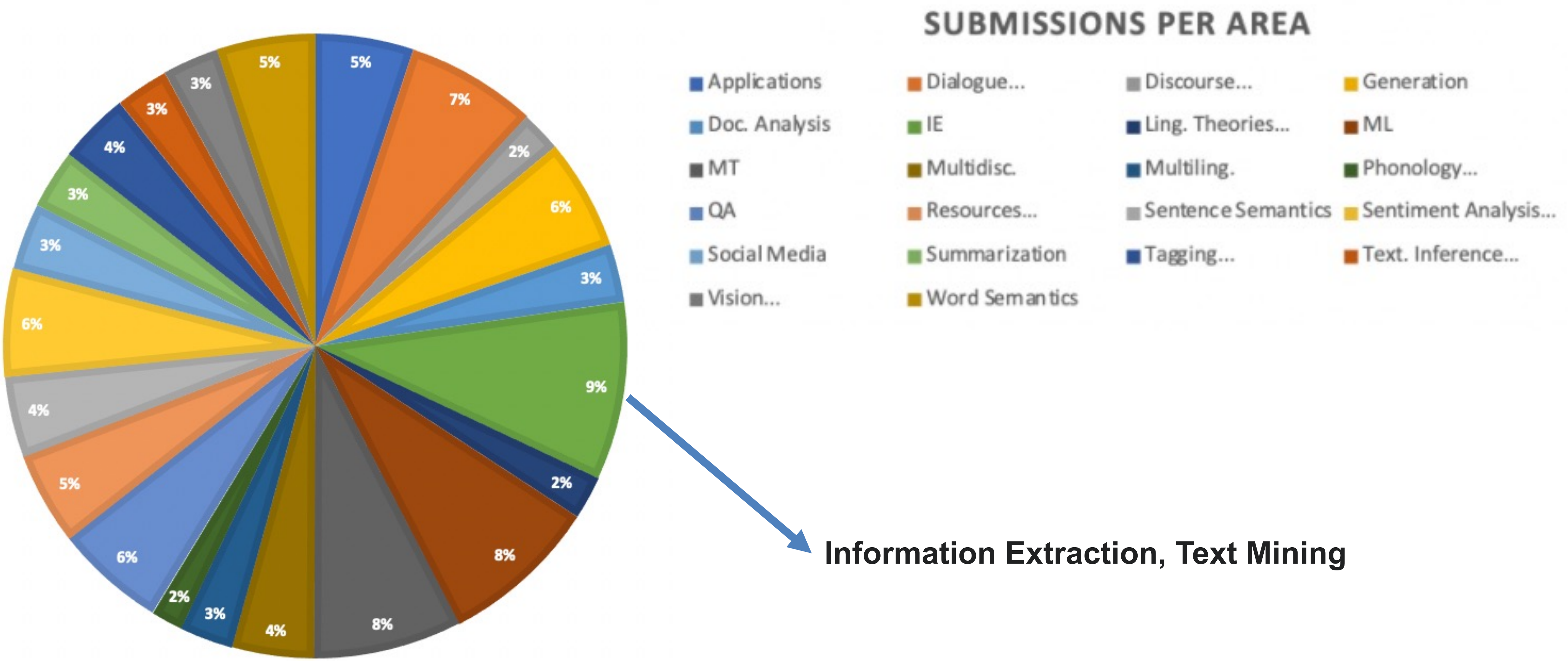
Tasks：

- Named Entity Recognition
- Relationship Extraction
- Coreference Resolution
- Event Extraction
- Table Extraction
- Table Information Extraction
- …

ACL 2020 Statistic

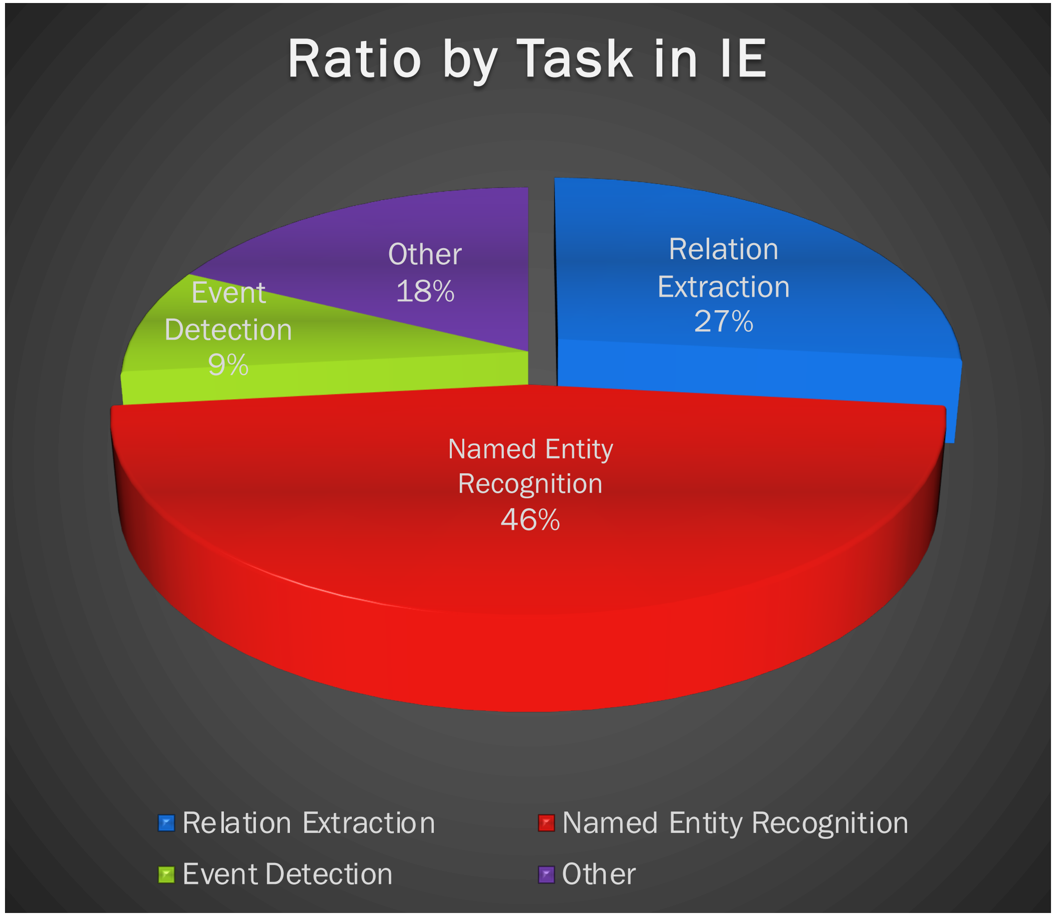


ACL 2019 Statistic



Information Extraction, Text Mining

Statistic in Information Extraction



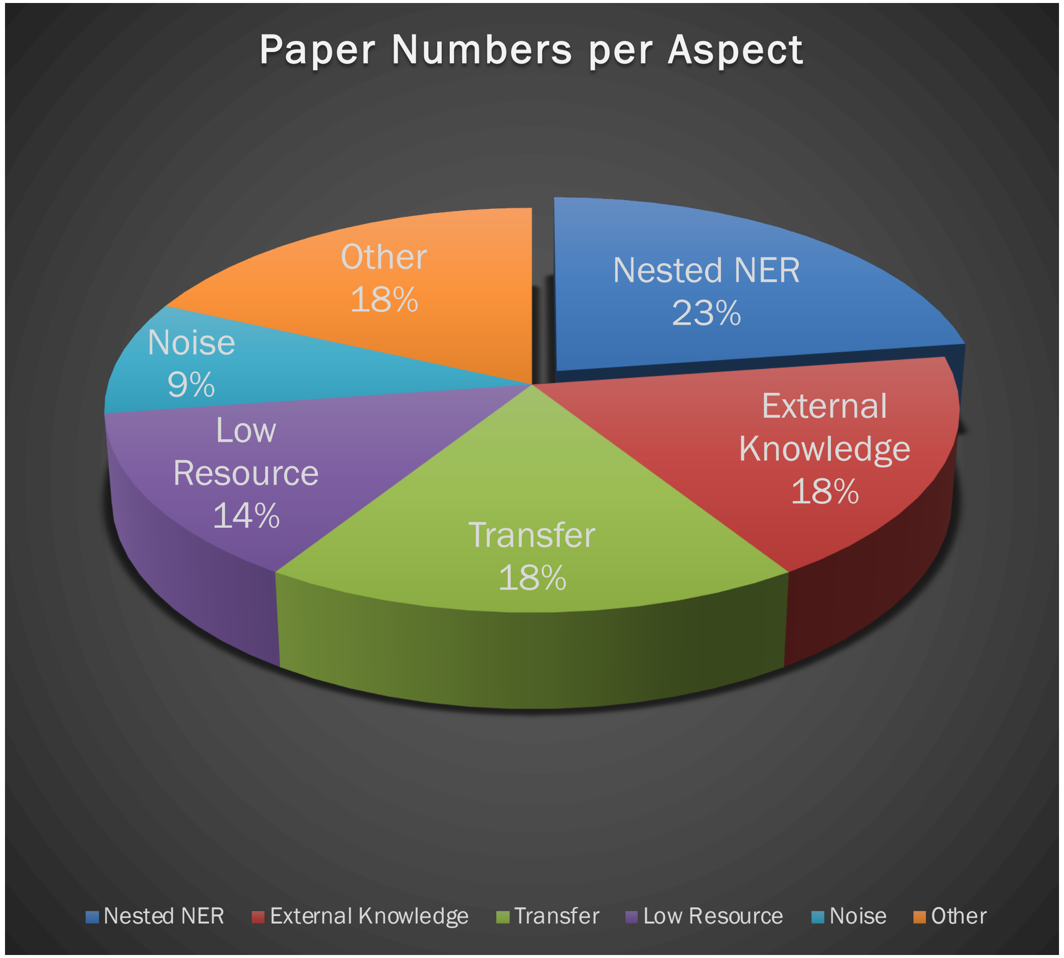
- About 46% of IE papers study Named Entity Recognition
- About 27% of IE papers study Relation Extraction
- About 9% of IE papers study Event Detection
- The others study diverse topics such as aspect-opinion mining, argument mining, and so on.

01 Name Entity Recognition

Statistic in Named Entity Recognition

Wordcloud generated from the titles of NER papers

Statistic in Named Entity Recognition



- About 18% of NER papers study cross-domain/cross-lingual NER
- About 18% of NER papers study the introduction of external knowledge
- About 14% of NER papers study distantly supervised NER
- About 9% of NER papers study the dealing with noise in NER

Nested Named Entity Recognition

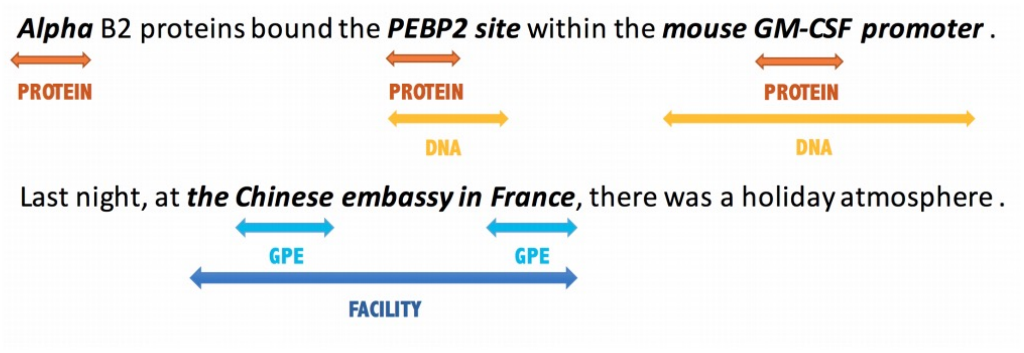


Figure 1: Examples for *nested* entities from GENIA and ACE04 corpora.

- NER → Machine Reading Comprehension
  - ➢ An entity type → a question
  - ➢ An annotated entity → $x_{start,end}$
  - ➢ X → Context

Dateset → (QUESTION, ANSWER, CONTEXT)

Entity	**Natural Language Question**
Location	Find locations in the text, including non-geographical locations, mountain ranges and bodies of water.
Facility	Find facilities in the text, including buildings, airports, highways and bridges.
Organization	Find organizations in the text, including companies, agencies and institutions.

Table 1: Examples for transforming different entity categories to question queries.

Li, Xiaoya, et al, 2021, A Unified MRC Framework for Named Entity Recognition

Shannon.AI

Nested Named Entity Recognition

English GENIA			
Model	**Precision**	**Recall**	**F1**
Hyper-Graph (Katiyar and Cardie, 2018)	77.7	71.8	74.6
ARN (Lin et al., 2019a)	75.8	73.9	74.8
Path-BERT (Shibuya and Hovy, 2019)	78.07	76.45	77.25
DYGIE (Luan et al., 2019)	-	-	76.2
Seq2seq-BERT (Straková et al., 2019)	-	-	78.31
BERT-MRC	**85.18**	**81.12**	**83.75 (+5.44)**
English KBP 2017			
Model	**Precision**	**Recall**	**F1**
KBP17-Best (Ji et al., 2017)	76.2	73.0	72.8
ARN (Lin et al., 2019a)	77.7	71.8	74.6
BERT-MRC	**82.33**	**77.61**	**80.97 (+6.37)**

Table 2: Results for *nested* NER tasks.

Chinese MSRA			
Model	**Precision**	**Recall**	**F1**
Lattice-LSTM (Zhang and Yang, 2018)	93.57	92.79	93.18
BERT-Tagger (Devlin et al., 2018)	94.97	94.62	94.80
Glyce-BERT (Wu et al., 2019)	95.57	95.51	95.54
BERT-MRC	**96.18**	**95.12**	**95.75 (+0.21)**
Chinese OntoNotes 4.0			
Model	**Precision**	**Recall**	**F1**
Lattice-LSTM (Zhang and Yang, 2018)	76.35	71.56	73.88
BERT-Tagger (Devlin et al., 2018)	78.01	80.35	79.16
Glyce-BERT (Wu et al., 2019)	81.87	81.40	81.63
BERT-MRC	**82.98**	**81.25**	**82.11 (+0.48)**

Table 3: Results for *flat* NER tasks.

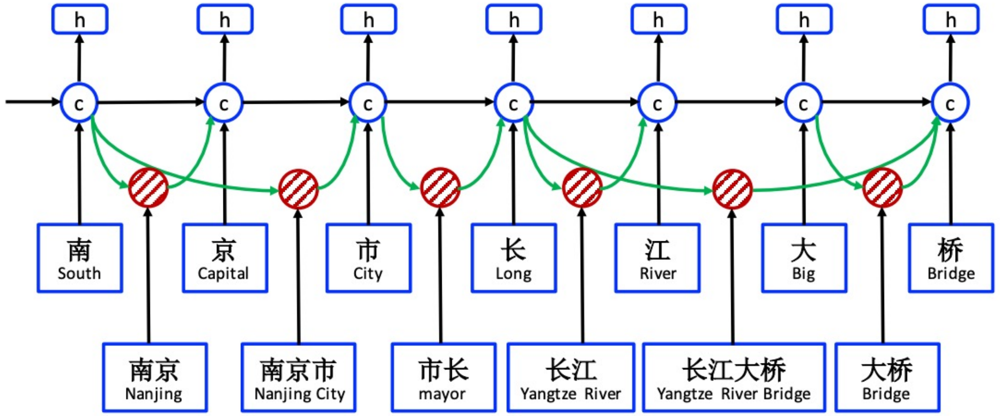
English OntoNotes 5.0	
Model	**F1**
BERT-Tagger	89.16
Position index of labels	88.29 (-0.87)
Keywords	89.74 (+0.58)
Wikipedia	89.66 (+0.59)
Rule-based template filling	89.30 (+0.14)
Synonyms	89.92 (+0.76)
Keywords+Synonyms	90.23 (+1.07)
Annotation guideline notes	91.11 (+1.95)

Table 5: Results of different types of queries.

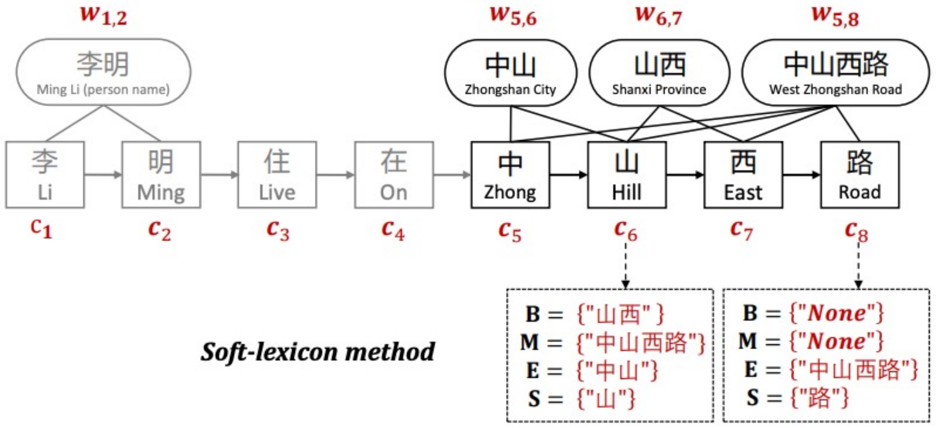
Li, Xiaoya, et al, 2020, A Unified MRC Framework for Named Entity Recognition

Named Entity Recognition with External Knowledge

- Simpler lexicons usage in Chinese NER



Lattice LSTM structure.



The SoftLexicon method.

Zhang, Yue, and Jie Yang. "Chinese NER Using Lattice LSTM."

Ma, Ruotian, et al. "Simplify the usage of lexicon in Chinese NER."

Named Entity Recognition with External Knowledge

- Simpler lexicons usage in Chinese NER

Models	OntoNotes	MSRA	Weibo	Resume
Lattice-LSTM	1×	1×	1×	1×
LR-CNN (Gui et al., 2019)	2.23×	1.57×	2.41×	1.44×
BERT-tagger	2.56×	2.55×	4.45×	3.12×
BERT + LSTM + CRF	2.77×	2.32×	2.84×	2.38×
Soft-lexicon (LSTM)	6.15×	5.78×	6.10×	6.13×
Soft-lexicon (LSTM) + bichar	6.08×	5.95×	5.91×	6.45×
Soft-lexicon (LSTM) + BERT	2.74×	2.33×	2.85×	2.32×

Table 2: Inference speed (average sentences per second, the larger the better) of our method with LSTM layer compared with Lattice-LSTM, LR-CNN and BERT.

Input	Models	P	R	F1
Gold seg	Yang et al., 2016	65.59	71.84	68.57
	Yang et al., 2016*†	72.98	**80.15**	**76.40**
	Che et al., 2013*	77.71	72.51	75.02
	Wang et al., 2013*	76.43	72.32	74.32
	Word-based (LSTM)	76.66	63.60	69.52
	+ char + bichar	**78.62**	73.13	75.77
Auto seg	Word-based (LSTM)	72.84	59.72	65.63
	+ char + bichar	73.36	70.12	71.70
No seg	Char-based (LSTM)	68.79	60.35	64.30
	+ bichar + softword	74.36	69.43	71.89
	+ ExSoftword	69.90	66.46	68.13
	+ bichar + ExSoftword	73.80	71.05	72.40
	Lattice-LSTM	76.35	71.56	73.88
	LR-CNN (Gui et al., 2019)	76.40	72.60	74.45
	Soft-lexicon (LSTM)	77.28	74.07	75.64
	Soft-lexicon (LSTM) + bichar	77.13	**75.22**	**76.16**
	BERT-Tagger	76.01	79.96	77.93
	BERT + LSTM + CRF	81.99	81.65	81.82
	Soft-lexicon (LSTM) + BERT	**83.41**	**82.21**	**82.81**

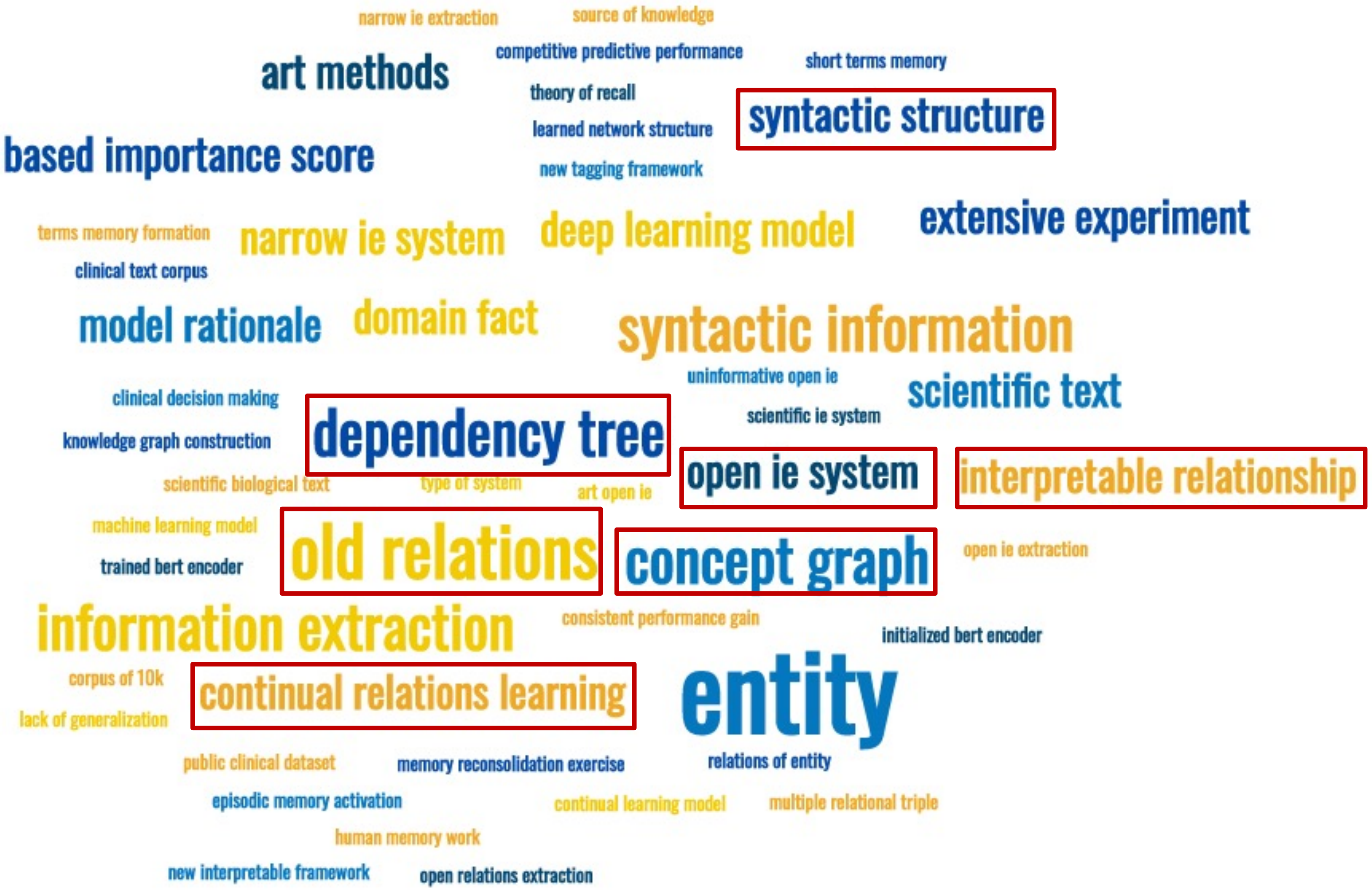
Table 3: Performance on OntoNotes. A model followed by (LSTM) (e.g., Proposed (LSTM)) indicates that its sequence modeling layer is LSTM-based.

Ma, Ruotian, et al. "Simplify the usage of lexicon in Chinese NER."

02

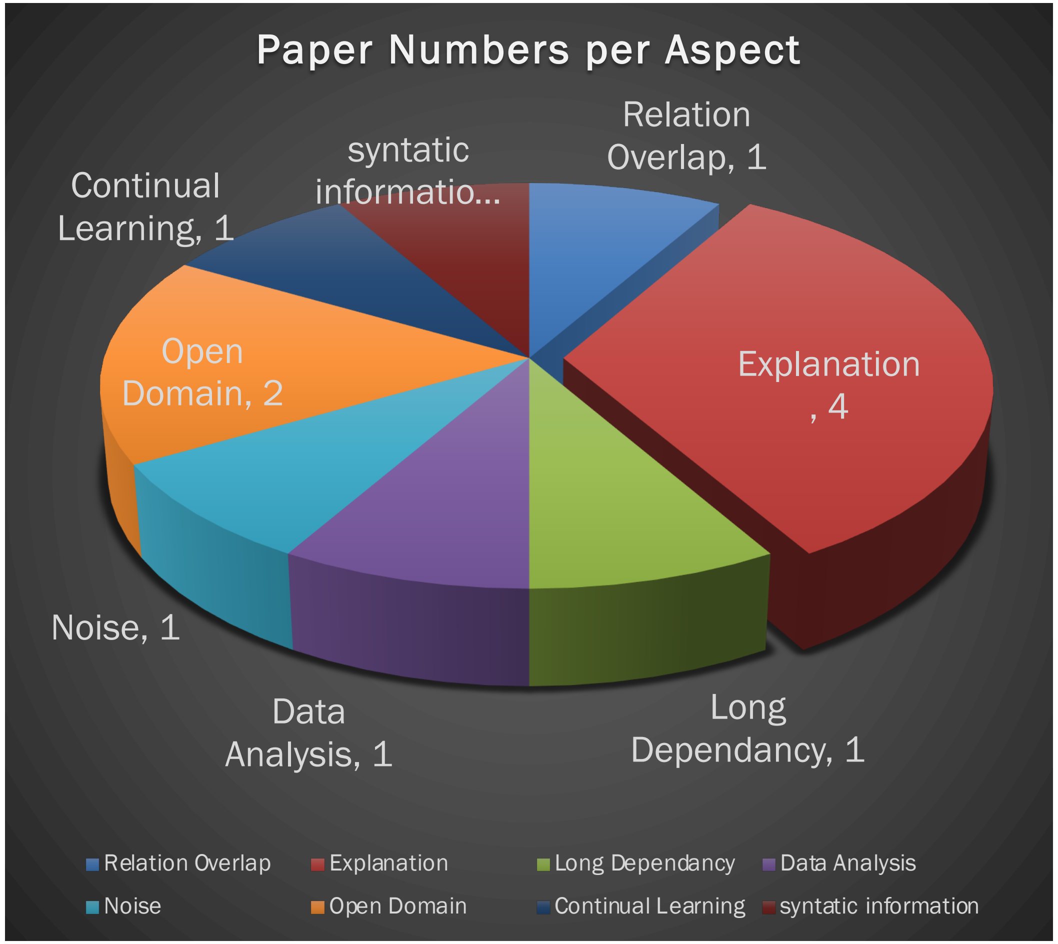
Relation Extraction

Statistic in Relation Extraction



Wordcloud generated from the abstracts of relation extraction papers

Statistic in Relation Extraction



- 4 papers study the interpretability of neural relation extraction
- 2 papers study open-domain relation extraction
- 1 paper studies cross-sentence relation extraction
- 1 paper studies the dealing of noise in NER

Interpretable & Open Domain Relation Extraction

- Learning interpretable relationships from open domain facts.

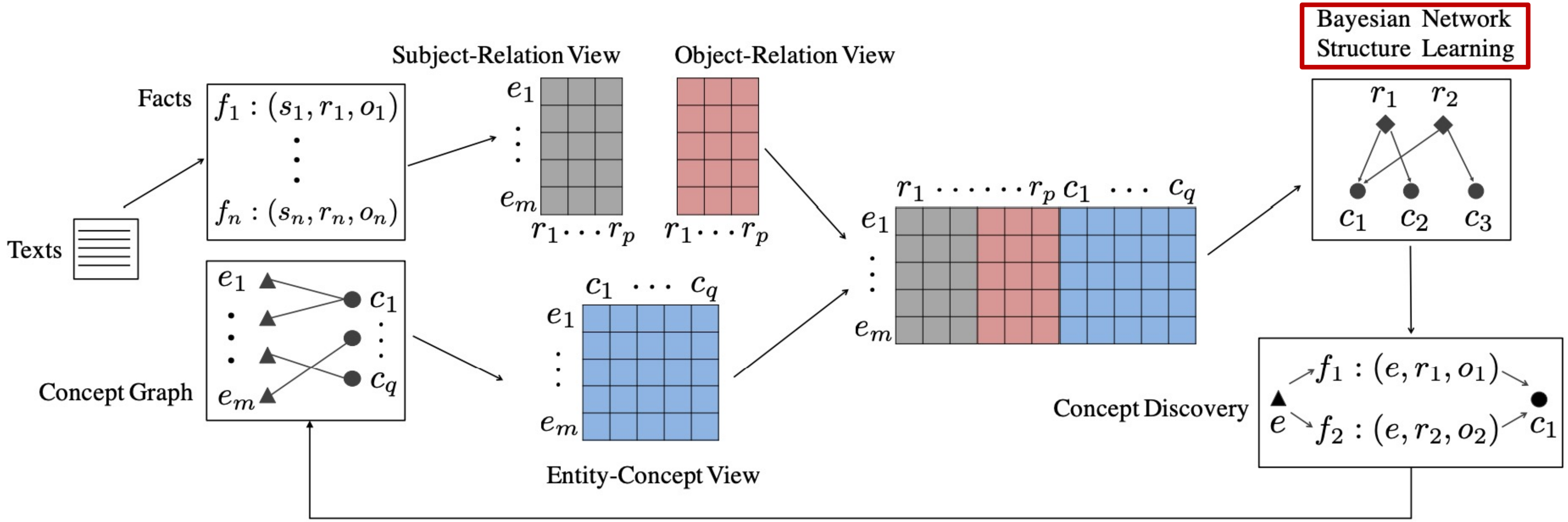


Figure 1: The workflow of learning interpretable relationships from open domain facts for concept discovery. $f_i = (s_i, r_i, o_i)$ represents a fact, where s_i and o_i are both entities, and r_i is a relation. We use e_i to denote an entity and c_i to represent a concept.

Zhang, Jingyuan, et al. "Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts."

Interpretable & Open Domain Relation Extraction

- Learning interpretable relationships from open domain facts.

Relation Selection		TF Selection				TFIDF Selection			
Dataset	Method	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
English	SVM(s)	58.19% (10)	55.17% (10)	87.43% (6)	67.65% (11)	72.38% (10)	67.28% (10)	87.12% (10)	75.93% (11)
	BNSL(s)	71.57% (5)	67.93% (5)	81.70% (10)	74.19% (6)	86.00% (2)	82.24% (2)	91.82% (2)	86.77% (2)
	SVM + BNSL(s)	71.62% (4)	68.36% (4)	80.48% (11)	73.93% (7)	82.04% (7)	78.31% (6)	88.63% (7)	83.15% (7)
	BNSL + SVM(s)	78.46% (1)	80.55% (1)	75.04% (12)	77.70% (3)	88.36% (1)	86.48% (1)	90.94% (4)	88.65% (1)
	SVM(o)	55.07% (12)	52.91% (12)	92.29% (1)	67.26% (12)	66.65% (12)	62.64% (12)	82.48% (12)	71.21% (12)
	BNSL(o)	71.14% (7)	65.68% (7)	88.54% (5)	75.42% (4)	82.64% (5)	78.99% (5)	88.95% (6)	83.67% (6)
	SVM + BNSL(o)	66.84% (9)	61.65% (9)	89.07% (3)	72.87% (8)	78.27% (9)	74.79% (8)	85.28% (11)	79.70% (9)
	BNSL + SVM(o)	77.02% (2)	73.10% (2)	85.50% (7)	78.81% (1)	84.16% (4)	81.49% (3)	88.40% (9)	84.80% (4)
	SVM	57.38% (11)	54.36% (11)	92.05% (2)	68.35% (10)	72.15% (11)	66.46% (11)	89.45% (5)	76.26% (10)
	BNSL	71.26% (6)	66.77% (6)	84.63% (9)	74.65% (5)	84.78% (3)	80.89% (4)	91.09% (3)	85.69% (3)
	SVM + BNSL	68.31% (8)	63.71% (8)	85.09% (8)	72.86% (9)	78.70% (8)	73.99% (9)	88.50% (8)	80.60% (8)
	BNSL + SVM	75.84% (3)	70.60% (3)	88.58% (4)	78.57% (2)	82.22% (6)	76.50% (7)	93.03% (1)	83.96% (5)
Chinese	SVM(s)	89.80% (8)	86.91% (8)	93.73% (5)	90.19% (8)	74.58% (8)	67.98% (6)	92.95% (8)	78.53% (8)
	BNSL(s)	92.23% (5)	90.24% (5)	94.71% (1)	92.42% (5)	75.01% (6)	67.90% (8)	94.88% (1)	79.16% (6)
	SVM + BNSL(s)	93.31% (4)	93.13% (4)	93.52% (8)	93.32% (4)	76.37% (3)	69.62% (3)	93.55% (6)	79.83% (3)
	BNSL + SVM(s)	95.56% (1)	97.36% (1)	93.65% (7)	95.47% (1)	77.54% (2)	70.64% (2)	94.27% (4)	80.76% (2)
	SVM(o)	51.16% (12)	50.71% (12)	82.58% (9)	62.84% (10)	50.55% (12)	50.33% (12)	84.65% (10)	63.12% (10)
	BNSL(o)	51.39% (10)	50.96% (10)	73.85% (11)	60.31% (12)	50.79% (10)	50.55% (10)	72.37% (12)	59.53% (12)
	SVM + BNSL(o)	51.33% (11)	50.82% (11)	82.41% (10)	62.87% (9)	50.66% (11)	50.39% (11)	84.73% (9)	63.20% (9)
	BNSL + SVM(o)	51.72% (9)	51.18% (9)	74.54% (12)	60.69% (11)	50.97% (9)	50.68% (9)	72.98% (11)	59.82% (11)
	SVM	90.35% (7)	87.69% (7)	93.88% (4)	90.68% (7)	74.68% (7)	67.95% (7)	93.45% (7)	78.68% (7)
	BNSL	92.15% (6)	90.16% (6)	94.62% (2)	92.34% (6)	75.12% (5)	68.08% (5)	94.61% (2)	79.18% (5)
	SVM + BNSL	93.61% (3)	93.55% (3)	93.68% (6)	93.61% (3)	76.33% (4)	69.57% (4)	93.60% (5)	79.82% (4)
	BNSL + SVM	95.46% (2)	96.59% (2)	94.25% (3)	95.40% (2)	77.68% (1)	70.77% (1)	94.32% (3)	80.87% (1)

Performance of relation selections on the entire data. The results are reported as "value + (rank)".

Dataset	English				Chinese			
Method	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
HypeNet	69.64%	75.09%	69.74%	72.31%	76.57%	87.17%	71.22%	78.39%
RNN(sen)	77.18%	80.74%	78.62%	79.67%	71.90%	72.85%	84.35%	78.18%
RNN(e)	67.77%	77.09%	61.62%	68.49%	57.67%	61.19%	79.53%	69.16%
RNN(s)	73.38%	80.35%	70.39%	75.04%	64.93%	64.02%	94.13%	76.21%
RNN(o)	70.95%	79.81%	65.46%	71.93%	64.97%	64.08%	94.01%	76.21%
RNN(f)	70.01%	79.08%	64.25%	70.90%	49.55%	61.23%	42.81%	49.95%
SVM(s)	76.68%	74.82%	88.93%	81.26%	85.06%	90.01%	84.33%	87.07%
SVM(o)	74.81%	72.72%	89.14%	80.10%	51.86%	57.54%	73.87%	64.69%
SVM	77.43%	74.38%	92.00%	82.25%	86.07%	90. 86%	85.22%	87.95%
BNSL(s)	86.03%	82.89%	**95.07%**	88.56%	87.54%	**92.40%**	86.21%	89.20%
BNSL(o)	86.22%	84.52%	92.76%	88.45%	49.03%	56.79%	61.10%	58.86%
BNSL	84.79%	81.87%	94.08%	87.55%	87.37%	92.32%	86.00%	89.05%
B + H	**91.27%**	**91.15%**	93.75%	**92.43%**	**87.88%**	86.01%	**95.18%**	**90.36%**

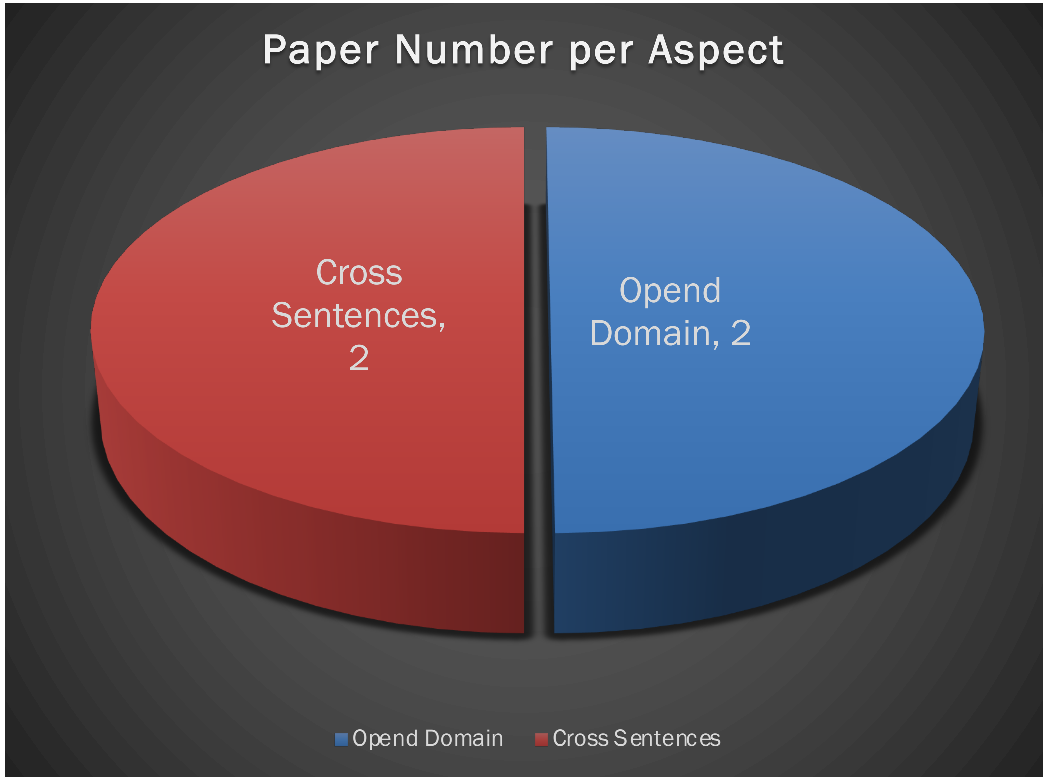
Performance on the co-occurred data. The best results are **in bold**.

Zhang, Jingyuan, et al. "Learning Interpretable Relationships between Entities, Relations and Concepts via Bayesian Structure Learning on Open Domain Facts."

03

Event Detection

Statistic in Event Detection



- **4** event detection papers was analyzed
- 2 papers study cross-sentence detection
- 2 paper study open-domain detection