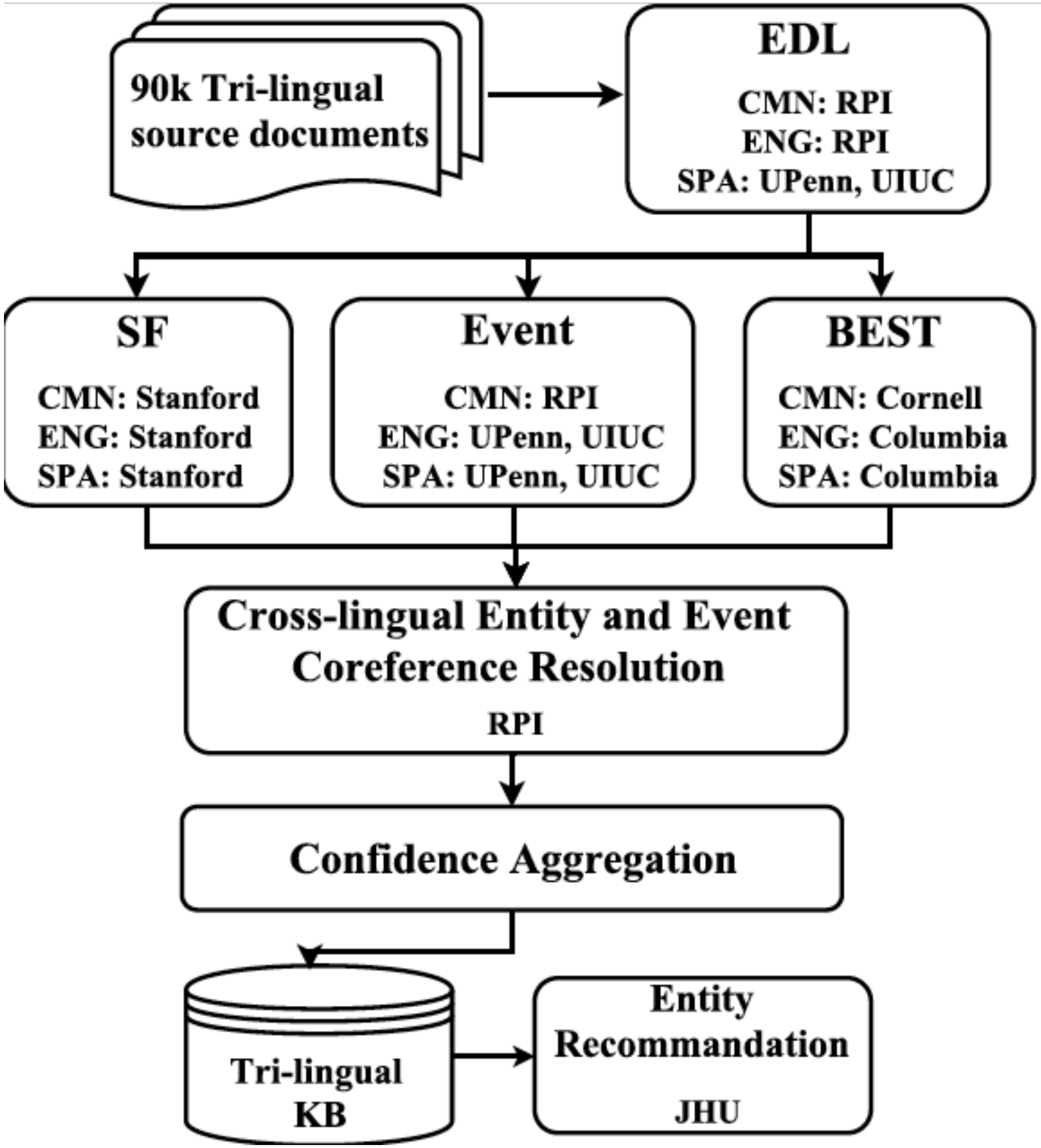
Cross-lingual Cold-Start Knowledge Base Construction

M. Al-Badrashiny, J. Bolton5, A. T. Chaganty, K. Clark, C. Harman, L. Huang, M. Lamm, J. Lei, D. Lu, X. Pan, A. Paranjape, E. Pavlick, H. Peng, P. Qi, P. Rastogi, A. See, K. Sun, M. Thomas, C. –T. Tsai, H. Wu, B. Zhang, C. Callison-Burch, C. Cardie, H. Ji, C. Manning, S. Muresan, O. C. Rambow, D. Roth, M. Sammons, B. Van Durme

TinkerBell

System Overview



The Devil's in the Details!

Overall Results

- Top performance at all cross-lingual tasks
  - We are the only team who did end-to-end KB construction for all languages and all tasks
- Compared with human performance (all hops)

slot types	#justifications	TinkerBell	Human	% Human
all	3	7.56%	47.1%	16.1%
all	1	13.32%	59.77%	22.3%
SF	3	11.43%	40.97%	27.9%
SF	1	17.30%	41.53%	41.7%

Novel Approaches

- EDL
  - A joint model of name tagging, linking and clustering based on multi-lingual multi-level common space construction
  - Joint transliteration and sub-word alignment for cross-lingual entity linking
- SF
  - Joint inference between EDL and SF
- Event extraction
  - dependency relation based attention mechanism for event argument extraction
- Sentiment Analysis (BeSt)
  - a target-focused method augmented with a polarity chooser and trained for the only entity-target task
- Cross-lingual cross-document entity and event coreference resolution

Entity Discovery and Linking

- Top performance for all languages in Cold-start++ KB construction

Team	NER			NERC			NERLC			KBIDs			CEAFmC+		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
3	**83.2**	**67.3**	**74.4**	**76.8**	**62.2**	**68.8**	**62.6**	**50.7**	**56.0**	**73.1**	**64.9**	**68.8**	**60.7**	**49.1**	**54.3**
13	52.8	54.8	53.8	29.8	30.9	30.3	22.6	23.4	23.0	64.1	46.9	54.2	19.7	20.5	20.1
8	81.7	53.0	64.3	71.7	46.5	56.4	5.5	3.5	4.3	0.0	0.0	0.0	4.8	3.1	3.7
Chinese															
3	84.8	62.9	**72.2**	79.6	59.1	**67.8**	**65.1**	**48.3**	**55.4**	79.9	**64.9**	**71.7**	**64.0**	**47.5**	**54.5**
18	75.0	60.5	67.0	70.0	56.5	62.6	47.8	38.5	42.7	**84.4**	38.7	53.1	46.3	37.4	41.4
13	68.2	47.4	55.9	38.8	26.9	31.8	31.5	21.9	25.8	62.3	44.4	51.8	30.6	21.3	25.1
17	79.8	56.2	66.0	73.9	52.0	61.1	14.7	10.3	12.1	0.0	0.0	0.0	13.9	9.8	11.5
23	56.2	**71.5**	63.0	51.7	**65.9**	57.9	9.9	12.7	11.1	0.0	0.0	0.0	8.9	11.4	10.0
8	**85.4**	50.8	63.7	**81.1**	48.3	60.5	5.0	3.0	3.7	0.0	0.0	0.0	4.6	2.8	3.5
English															
3	77.5	66.7	71.7	71.5	61.5	66.1	**57.9**	49.8	**53.5**	63.6	**68.2**	**65.8**	**54.1**	46.5	**50.1**
18	78.6	79.1	**78.8**	72.6	**73.0**	**72.8**	52.9	**53.2**	53.0	**70.4**	49.8	58.4	48.8	**49.1**	49.0
17	73.0	**79.5**	76.1	66.1	71.9	68.9	23.2	25.3	24.2	0.0	0.0	0.0	21.1	22.9	22.0
19	**90.8**	62.5	74.1	**83.3**	57.3	67.9	26.9	18.5	21.9	0.0	0.0	0.0	23.5	16.2	19.2
13	55.9	70.5	62.4	31.7	39.9	35.3	19.5	24.6	21.8	66.9	50.5	57.6	16.0	20.2	17.9
8	78.5	48.9	60.3	71.3	44.5	54.8	7.8	4.9	6.0	0.0	0.0	0.0	7.0	4.4	5.4
22	51.5	32.9	40.1	29.7	19.0	23.2	5.2	3.3	4.0	0.0	0.0	0.0	4.9	3.1	3.8
Spanish															
3	**86.6**	**74.3**	**80.0**	**78.5**	**67.4**	**72.5**	**64.1**	**55.0**	**59.2**	**76.4**	**62.1**	**68.5**	**62.8**	**53.9**	**58.0**
13	40.9	50.4	45.1	22.7	28.0	25.1	19.9	24.6	22.0	64.0	46.6	53.9	16.2	20.0	17.9
8	84.9	58.7	69.4	63.5	43.9	51.9	5.2	3.6	4.2	0.0	0.0	0.0	4.5	3.1	3.7

- English and Chinese EDL see tomorrow RPI's talk
- This talk: details about Spanish EDL

Event Coreference Resolution

- Construct an undirected weighted graph:
  - node: event nugget
  - edge: coreference link between two event nuggets
- Apply hierarchical clustering to classify event nuggets into hoppers

Features	Remarks(EM1: the first event mention, EM2: the second event mention)
type_subtype_match	1 if the types and subtypes of the event nuggets match
trigger_pair_exact_match	1 if the spellings of triggers in EM1 and EM2 exactly match
stem_of_the_trigger_match†	1 if the stems of triggers in EM1 and EM2 match
similarity_of_the_triggers(wordnet)*	quantized semantic similarity score (0-5) using WordNet resource
similarity_of_the_triggers(word2vec)	quantized semantic similarity score (0-5) using word2vec embedding
POS_match*	1 if two sentences have the same NNPCD
token_dist	how many tokens between triggers of EM1 and EM2 (quantized)
realis_conflict	1 if the realis in EM1 and EM2 exactly match
Entity_match	Number of entities appear both in sentences of EM1 and EM2
Entity_prior	Number of entities appear only in the sentence of EM1
Entity_act	Number of entities appear only in the sentence of EM2

- Event arguments our system found & missed by human in KB construction
  - **compound noun:** 日军*一有*伤亡,就会疯狂报复老百姓的 *(once Japanese army has* ***injures and deaths****, they will revenge civilians like crazy.)*
  - *Why should it be Apple's problem? Will it stop you form* ***buying*** *an iPhone?*

TINKERBELL – UIUC
EVENT NUGGETS AND EDL

DEFT @ UIUC

Mark Sammons

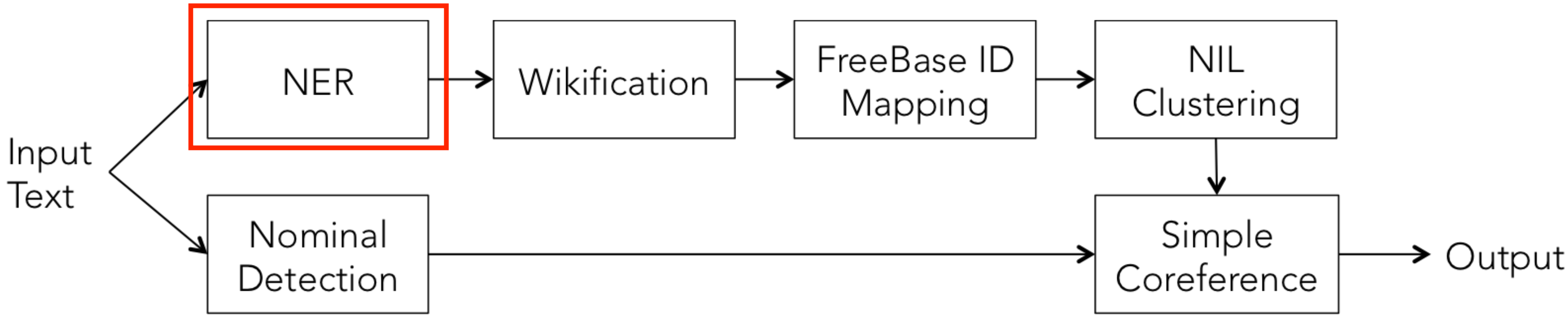
mssammon@illinois.edu

November 2017

Spanish Entity Detection and Linking

Chen-Tse Tsai

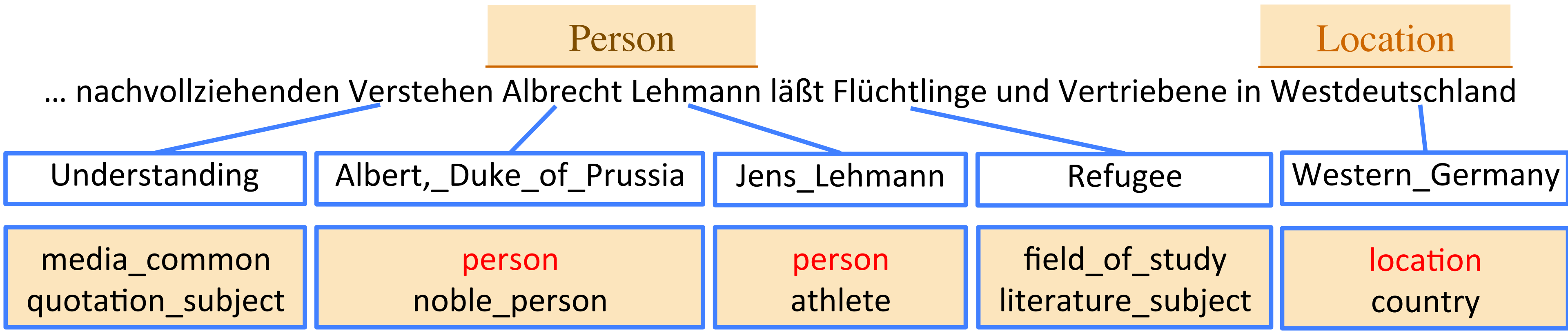
SPANISH EDL: NER



- NER (Chinese and Spanish)
  - Cross-Lingual NER via Wikification [Tsai et al., CoNLL 2016]
  - Wikify n-grams and add wikifier features to the Illinois NER model
  - Chinese/Spanish brown clusters
  - Chinese/Spanish gazetteers

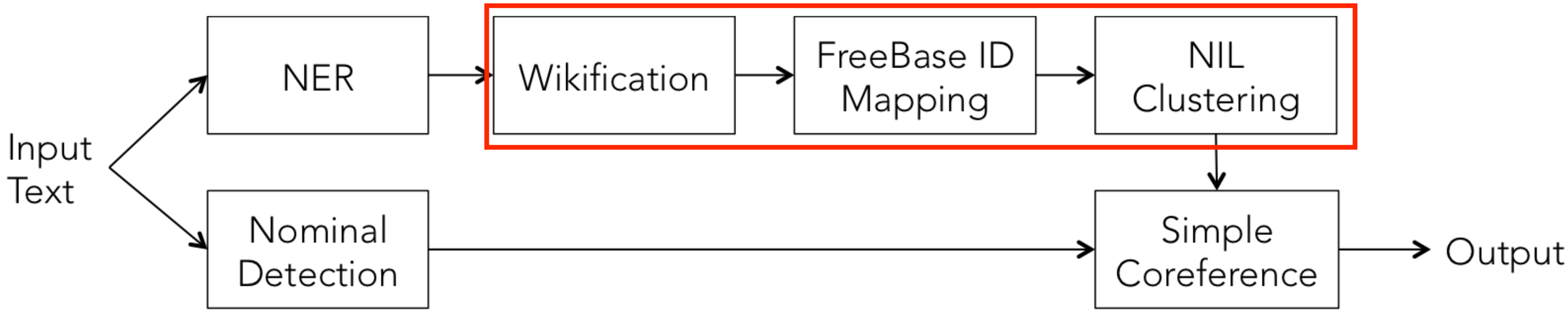
NER with no Target Language Training Data: Key Idea

- Cross-lingual Wikification generates good language-independent features for NER by grounding n-grams (TsaiMaRo2016)



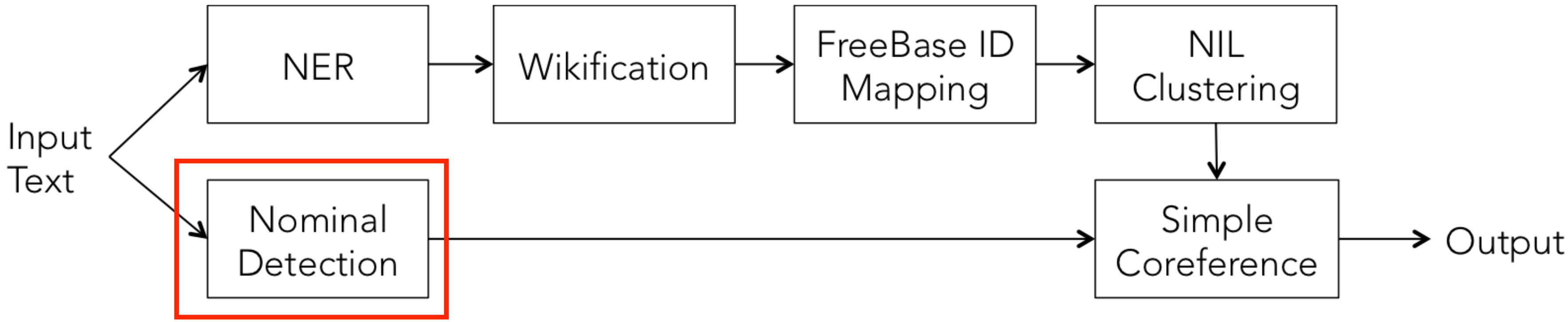
- Words in any language are grounded to the English Wikipedia
  - Features extracted based on the titles can be used across languages
- Instead of the traditional pipeline: NER → Wikification
  - Wikified n-grams provide features for the NER model
  - Turns out to be useful also when monolingual training data is available
  - Use TAC 2015 EDL train + eval, 2016 eval, DEFT ERE Spanish data to train

SPANISH EDL: WIKIFICATION



- Wikification
  - Uses cross-lingual word and title embeddings to compute similarities between a foreign mention and English title candidates [Tsai and Roth, NAACL 2016]
  - Obtain FreeBase ID using the links between Wikipedia titles and FreeBase entries if a mention is grounded to some Wikipedia entry.
  - NIL Clustering: unlinked mentions are clustered together if Jaccard similarity of surface forms > 0.5

SPANISH EDL: WIKIFICATION



- Nominal/Pronoun Detection
  - Train Illinois NER model on the nominal noun annotations
    - Only generic features – words themselves, Brown clusters
    - Train on nominal mentions in the TAC EDL 2016 Spanish evaluation data. (ERE nominal data does not help)
    - For pronouns, train on pronouns in DEFT ERE (no pronominal data in previous TAC evals)
- Co-ref to linked NE: Type + proximity + author heuristics

Results

- Hard to interpret cold start scores to extract EDL, so these are scores for UIUC's standalone EDL submission
  - Some improvements to nominal mention detection and linking, so almost certainly higher than Cold Start performance

2017 Evaluation Set			
Measure	Precision	Recall	F1
Spanish			
strong typed mention match	84.6	69.4	76.3
strong typed all match	77.3	48.9	59.9
typed mention ceaf	78.3	49.5	60.7

Cross-Lingual Wikification Evaluation [Tsai & Roth NAACL'16]

The baseline of simply choosing the title that maximizes Pr(title|mention) is good for many mentions:

Language	Method	Hard	Easy	Total
Spanish	EsWikifier	40.11	99.28	79.56
	MonoEmb	38.46	96.12	76.90
	WordAlign	48.75	95.78	80.10
	WikiME	**54.46**	94.83	**81.37**
Chinese	MonoEmb	43.73	97.85	79.81
	WikiME	**57.61**	98.03	**84.55**
Turkish	MonoEmb	40.47	98.15	78.93
	WikiME	**60.18**	97.55	**85.10**
Tamil	MonoEmb	34.51	98.65	77.30
	WikiME	**54.13**	99.13	**84.15**
Tagalog	MonoEmb	35.47	99.44	78.12
	WikiME	**56.70**	98.46	**84.54**

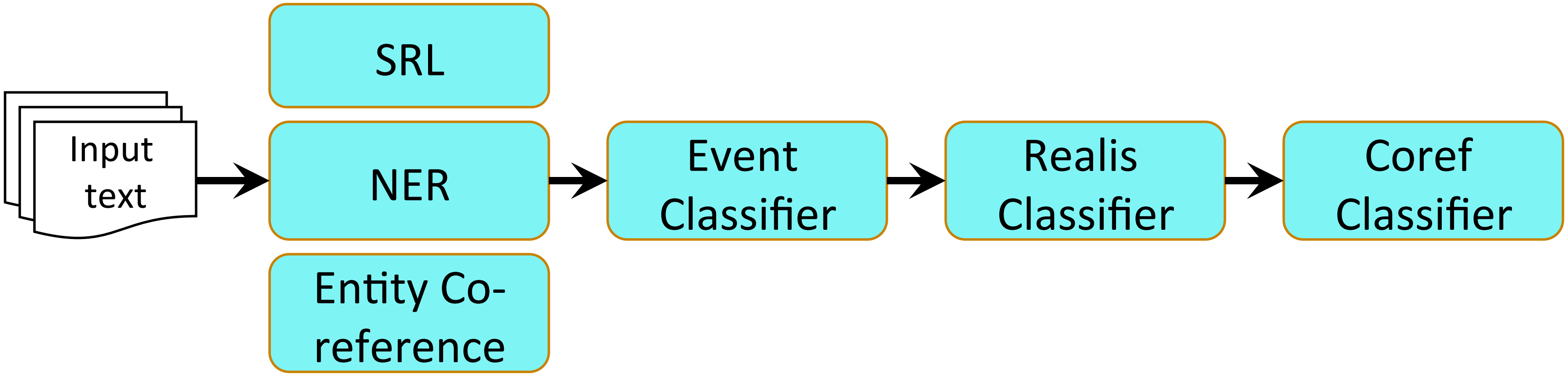
Citations

- Chen-Tse Tsai and Dan Roth, "Cross-lingual Wikification using Multilingual Embeddings", *NAACL* (2016)
- Chen-Tse Tsai, Stephen Mayhew, and Dan Roth, "Cross-lingual Named Entity Recognition via Wikification", *CoNLL* (2016)
- Haoruo Peng and Yangqiu Song and Dan Roth, "Event Detection and Co-reference with Minimal Supervision", *EMNLP* (2016)

Event Nugget Detection and Co-Reference

Haoruo Peng, Hao Wu

Event Nugget Detection and Coreference



- Pipeline architecture
- Use SRL predicates as event trigger candidates
- Classify triggers into 34 types, filter extraneous typed triggers
- Realis: Classify survivors into Actual/General/Other
- Binary classifier, applied to "Actual" pairs, into Coref/Non-coref

- Spanish: translate to English, process, map back

SRL ANNOTATION COVERAGE OF EVENTS

- From Peng et al. 2016, analysis of ACE 2005 and TAC 2015 event coverage by predicted SRL

ACE		Precision	Recall	F1
Predicates over Triggers	Verb-SRL	—	93.2	—
	Nom-SRL	—	87.5	—
	All	—	91.9	—
SRL Args over Event Args	Verb-SRL	90.4	85.7	88.0
	Nom-SRL	92.5	73.5	81.9
	All	90.9	82.3	86.4
TAC KBP		**Precision**	**Recall**	**F1**
Predicates over Triggers	Verb-SRL	—	90.6	—
	Nom-SRL	—	85.5	—
	All	—	$\boxed{88.1}$	—
SRL Args over Event Args	Verb-SRL	89.8	83.6	86.6
	Nom-SRL	88.2	69.9	78.0
	All	89.5	81.0	85.0

Tinkerbell English/Spanish Event Results

- Low scores for Tinkerbell system:
  - Only detected event nugget + coref, not event arguments
  - during later TAC event track, found several bugs
- Results from TAC event track: English Event Nugget Detection

	Precision	Recall	F1
Dev Set			
Span	61.40	55.46	58.28
Type	50.68	44.75	47.54
Realis	41.76	36.32	38.86
Overall	33.50	32.10	30.81
Test Set			
Span	53.44	41.72	46.86
Type	37.46	29.24	32.85
Realis	30.30	23.65	26.57
Overall	19.80	15.46	17.36

EVENT RESULTS FROM TAC EVENT TRACK (CONT'D)

- Event Nugget Co-reference: English

	BCUB	CEAFe	MUC	BLANC	AVG
Dev Set	36.86	35.67	13.43	9.77	23.93
Test Set	24.98	23.36	12.57	8.96	$\boxed{17.47}$

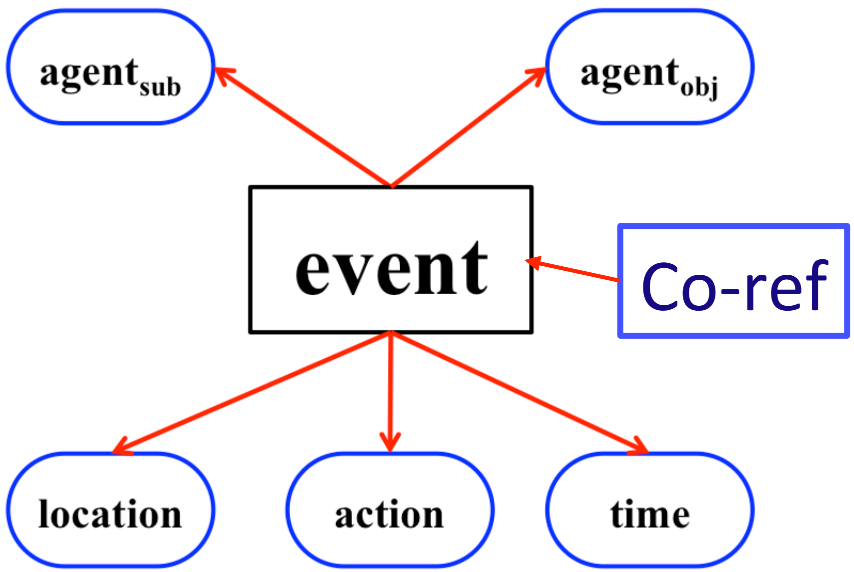
- Event Nugget Co-reference: Spanish

	BCUB	CEAFe	MUC	BLANC	AVG
Dev Set	22.06	20.81	13.52	7.37	15.94
Test Set	15.93	15.85	3.89	3.44	$\boxed{9.78}$

Current Work: Minimally Supervised Event Detection

- Peng & Roth EMNLP'16
- Deterministic Mapping from E-SRL to Event Components

  - Action: SRL predicate
  - Agent_{sub} : SRL subject
  - Agent_{obj} : SRL object
  - Time: Temporal Expression
  - Location: NER location
  - Entity Co-reference



EVENT VECTOR REPRESENTATION

ESA: A Wikipedia driven approach.
Represents a word as a (weighted) list of all Wikipedia titles it occurs in
[Gabrilovich & Markovitch 2009]

- Unsupervised Conversion
  - **Representations are generic**; do not depend on the task and data set but rather on a lot of, lazily read, text. It takes event structure into account.
- Text-Vector Conversion Methods
  - Explicit Semantic Analysis (ESA) is used for each component (sparse representation, up to 200 active coordinates)
  - (Found to be better than Brown Cluster(BC), Word2Vec, Dep. Embedding)
- Basic Vector Representation
  - Concatenate vector representations of all event components

[......] = [[......] [......] [......] [......] [......] [......]]

event — action, $\text{agent}_{\text{sub}}$, $\text{agent}_{\text{obj}}$, location, time, sentence or clause

- Augmented Vector Representation
  - Augment by concatenating more text fragments to enhance the interactions between the action and other arguments

[......] = [[......] [......] [......] [......] [......]]

$\text{event}_{\text{aug}}$ — event, $\text{agent}_{\text{sub}}$ + action, $\text{agent}_{\text{obj}}$ + action, location + action, time + action

Event Vector Representation Advantage

- Domain Transfer
  - Event Vector (MSEP) performs better outside training domains
  - Supervised methods are shown to over-fit and performance drops when transferring domains (here: Newswire and Forums)

	Train	Test	MSEP	Supervised
Event Detection				Span+Type F1
In Domain	NW	NW	58.5 *	**63.7**
Out of Domain	DF	NW	**55.1** *	54.8
In Domain	DF	DF	57.9	**62.6**
Out of Domain	NW	DF	**52.8**	52.3
Event Co-reference				AVG F1
In Domain	NW	NW	73.2	**73.6**
Out of Domain	DF	NW	**71.0**	70.1
In Domain	DF	DF	68.6	**68.9**
Out of Domain	NW	DF	**67.9**	67.0

*

Belief and Sentiment

- Belief and Sentiment are *cognitive states*
  - Analyze text to understand what people (the author, other people) think is true, and like and dislike
- TAC KBP 2016: BeSt track
  - Source-and-Target Belief and Sentiment
- Multiple conditions
  - 2 genres
    - Discussion forums
    - Newswire
  - 3 languages
    - English, Chinese, Spanish
  - 2 ERE conditions
    - Gold
    - Detected (RPI, UIUC -- thanks!)

ColdStart++: Belief and Sentiment

- Actually, only Sentiment
- Actually, only Sentiment towards Entities
- Columbia
  - English
  - Spanish
- Cornell
  - Chinese
- Both sites used the systems they developed for TAC KBP BeSt 2016, with small improvements
  - Addition of confidence measure

Results from 2016 BeSt Eval

Columbia English Results 2016 BeSt (best results in eval)

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	Disc. Forums	8.1%	70.6%	14.5%	3.7%	29.7%	6.5%
	Newswire	4.0%	35.5%	7.2%	2.3%	16.3%	4.0%
Columbia System 1	Disc. Forums	14.1%	38.5%	20.7%	6.2%	20.6%	9.5%
	Newswire	7.3%	16.5%	10.1%	2.7%	9.0%	4.2%

- Discussion Forums easier
  - There is more sentiment in DFs
- Predicted ERE hard

Results from 2016 BeSt Eval

Cornell Chinese Results 2016 BeSt (best resuts in eval)

System	Genre	Gold ERE			Predicted ERE		
		Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Baseline	Disc. Forums	5.0%	66.1%	9.2%	1.6%	6.1%	2.6%
	Newswire	0.7%	23.1%	1.4%	0.3%	2.0%	0.6%
Cornell System 1 (gold) System 2 (pred)	Disc. Forums	52.9%	27.5%	36.2%	12.1%	1.2%	2.1%
	Newswire	21.9%	4.3%	7.2%	5.9%	0.9%	1.6%

- Did relatively better on Gold than Columbia on E
- Discussion Forums easier
  - There is more sentiment in DFs
- Predicted ERE hard

Chinese Belief and Sentiment (Cornell)

- Hybrid approach based on our belief and sentiment system at TAC 2016 with the following changes:
  - More training data
    - BeSt 2016 eval
    - Chinese slangs and idioms to improve sentiment analysis
  - Confidence
    - We build 7 versions of the system, each optimized to a different F_{β} measure; then set the confidence of a sentiment $c_{sentiment}$ heuristically, based on the number of systems that report it
      - E.g., 0.1 if 1 system reports, 0.3 if 2, 0.5 if 3, 0.7 if 4, etc.
    - The final confidence c_{final} is obtained in two different ways
      - $c_{final} = c_{sentiment}$
      - $c_{final} = c_{sentiment} \cdot c_{target} \cdot c_{source}$

Columbia English/Spanish Sentiment

- Approach in 2016 assumes two defaults
  - Source is always author
  - Sentiment is always negative
- Approach based on:
  - Sentence segments
  - Whole posts
  - Author history
- We added a positive sentiment detector for CS++ 2017
- We added more training data
- Confidence: used ML confidence scores, and then added priors on target types
  - These priors made no difference whatsoever (why?)

Results

- Results are disappointing for Columbia systems (English, Spanish)
- K3, all hops

Language	LDC-Mean-All-Macro			SF-All-Macro		
	Prec.	Rec.	F-meas.	Prec.	Rec.	F-meas.
Chinese Sys1 Cornell	18.7%	41.1%	21.8%	20.0%	46.0%	23.9%
English Columbia	6.5%	16.3%	7.4%	6.8%	14.1%	6.8%
Spanish Columbia	2.4%	9.8%	3.2%	2.8%	11.1%	3.5%

Why are Results so Low for English and Spanish?

- Had already seen that predicted ERE decreases performance
  - CS++ results in line with BeSt 2016 results on predicted ERE
- Chinese system made more systematic use of outside resources than Columbia systems did
- As a result, some overfitting to training data for English and Spanish
- Obvious remedy: train on more varied data, use more external resources (sentiment dictionaries etc.)

Tinkerbell – Stanford

Tri-lingual Slot Filling

Arun Chaganty, Ashwin Paranjape, Jason Bolton, Jinhao Lei, Matthew Lamm, Abigail See, Kevin Clark, Yuhao Zhang, Peng Qi, **Christopher D. Manning**

CS Knowledge Base Population

Penner is survived by his brother, John, a copy editor at the Times, and his former wife, Times sportswriter Lisa Dillman.

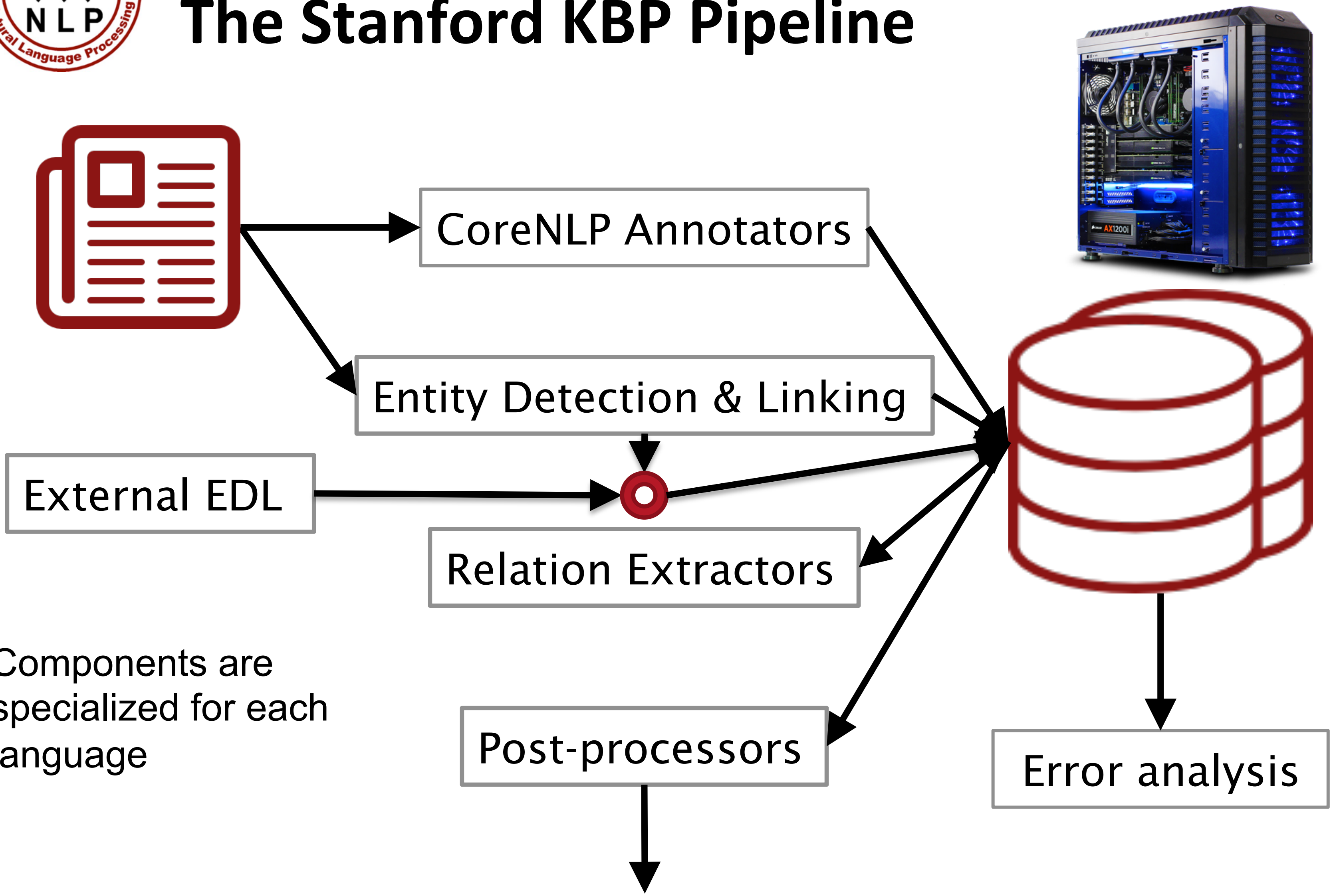
Subject	Relation/Slot	Object
Mike Penner	per:spouse	Lisa Dillman
Lisa Dillman	per:title	Sportswriter
Lisa Dillman	per:employee_of	Los Angeles Times
…	…	…

CS KB/SF 2017

- **Common system architecture**
- Entities
- English system
- Chinese system
- Spanish system
- Results

The Stanford KBP Pipeline

20 cores, 768GB RAM,
1.2TB SSD.



Components are specialized for each language

CS KB/SF 2017

- Common system architecture
- **Entities**
- English system
- Chinese system
- Spanish system
- Results

Entities for slot filling

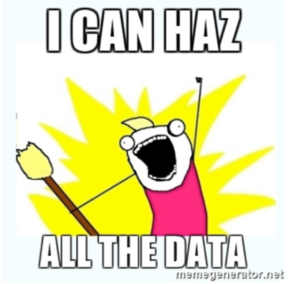
- Need to identify possible slot filling candidates, so annotate dates, titles, etc. with a rule based system.
  - Use lots of TokensRegex patterns, SUTime and HeidelTime (for Spanish).
- Our internal system also uses a named entity recognition system to identify name mentions and uses coreference for pronominal mentions. We ignore nominal mentions.
  - Use the neural coreference system in Stanford CoreNLP for English and Chinese and a rule based system for Spanish.
  - **This year: Improved named entity recognition**
- **This year: fusion with external EDL systems**

Improved named entity recognition

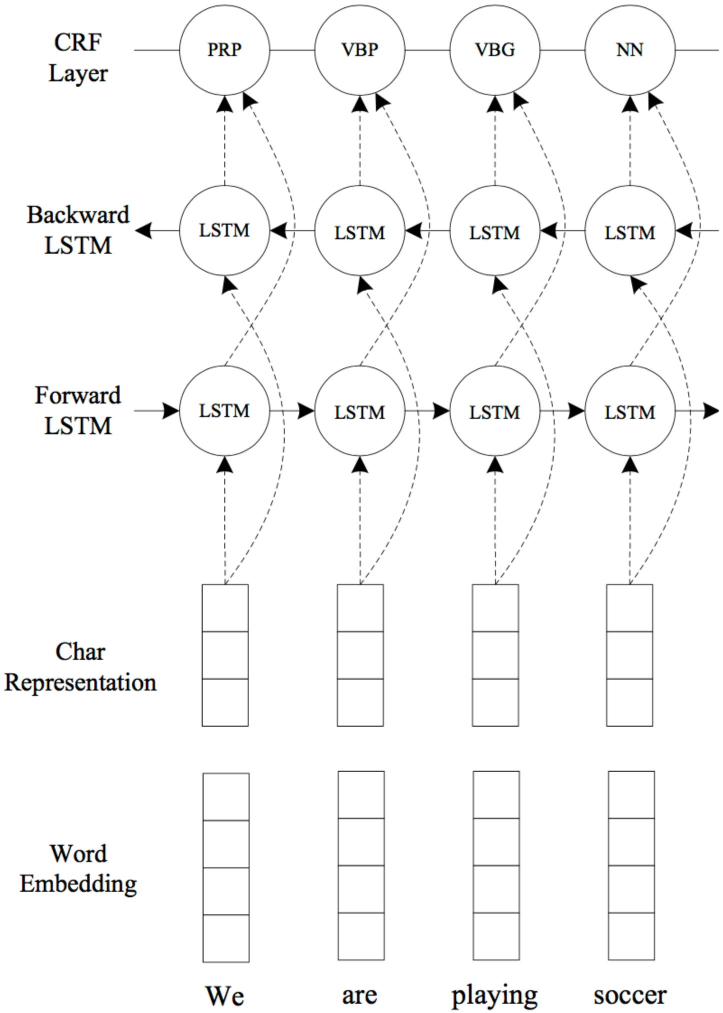
- Several new datasets for training

	Old	New in 2017
English	ACE 2002 / 2003 MUC 6 and 7 CoNLL 2003 OntoNotes	EDL Comprehensive Training Data 2014, 2015 ERE Discussion Forum Annotation 2014 ERE Chinese/English Parallel Annotation 2014 Rich ERE Training Annotation 2015 and 2016
Chinese	Ontonotes 5 ACE 2005 Multilingual	ACE 2004 Multilingual EDL Comprehensive Training Data 2015 ERE Chinese/English Parallel Annotation 2014, 2015 ERE Discussion Forum Annotation 2014 Rich ERE Chinese/English Parallel Annotation 2015 Rich ERE Training Annotation 2015
Spanish	Ancora Spanish Treebank DEFT Spanish Treebank v2	CoNLL 2003 ACE 2007 Multilingual EDL Comprehensive Training Data 2015 Rich ERE Annotation 2015 Light ERE Training Data 2015

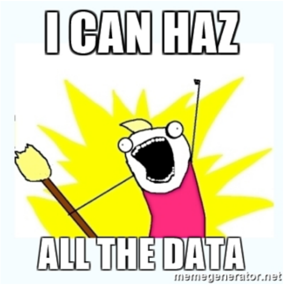
New Neural NER model for English



- We added a Bi-directional LSTM-CNNs-CRF Model for NER
- Based on [Xuezhe Ma, and Eduard Hovy. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.]



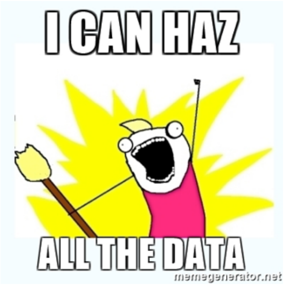
Improved named entity recognition: results



- Data from the EDL and ERE resources help significantly
  - Particularly provided in-domain data for discussion forums
  - More pronounced for Spanish and Chinese
- The neural bi-LSTM CRF model results in increased score for English

EDL 2015-16	Original training data	+ New training data	+ Neural model
Spanish	55.0	70.0	
Chinese	62.4	74.9	
English	75.5	80.0	80.9

Improved named entity recognition: impact on slot filling



- The dataset augmentation resulted in relatively minor improvements on its own, but the neural model helped significantly.

2017 KBP	Original training data	+ New training data	+ Neural model
Spanish	18.6	18.6	
Chinese	14.9	-	
English	22.2	22.2	25.4

EDL fusion for ColdStart++

EDL fusion for ColdStart++: results on 2016 eval (dev)

- Merge entities from other Tinkerbell teams with Stanford's entities and fine-grained typed slot candidates.
- Improvements across languages: better EDL helps in relation extraction!

KBP 2016	EDL System	P	R	F1
English	Stanford only	55.7	9.6	16.4
	+ RPI	49.8	11.3	18.4
Chinese	Stanford only	27.9	22.6	25.0
	+ RPI	16.5	27.3	20.6
Spanish	Stanford only	28.3	2.5	4.6
	+ UIUC	19.8	3.4	5.9

Scores are biased because of incompleteness!

EDL fusion for ColdStart++: results on 2017 evaluation

- EDL fusion made a huge impact on Chinese, and improved over our original English system, but the neural NER system outperformed both.

KBP 2017	EDL System	P	R	F1	AP
English	Stan. CRF only	21.3	29.1	22.2	26.2
	Stan. Neural only	23.8	33.3	25.4	27.5
	+ RPI	22.3	32.4	23.9	26.7
Chinese	Stanford only	16.3	14.9	14.9	16.8
	+ RPI	19.6	18.1	18.0	18.4
Spanish	Stanford only	-	-	-	-
	+ UIUC	19.2	19.8	18.6	16.3

CS KB/SF 2017

- Common system architecture
- Entities
- **English system**
- Chinese system
- Spanish system
- Results

English Extraction systems

- Pattern-based systems
  - TokensRegex
  - Semgrex
  - Coreference-based alternate names
  - Rule-based system for identifying webpage URLs.
  - Nested mention extractor for subsidiaries and headquarters
- Self-trained supervised classifier
- **New neural network system**

Position-aware LSTM with attention

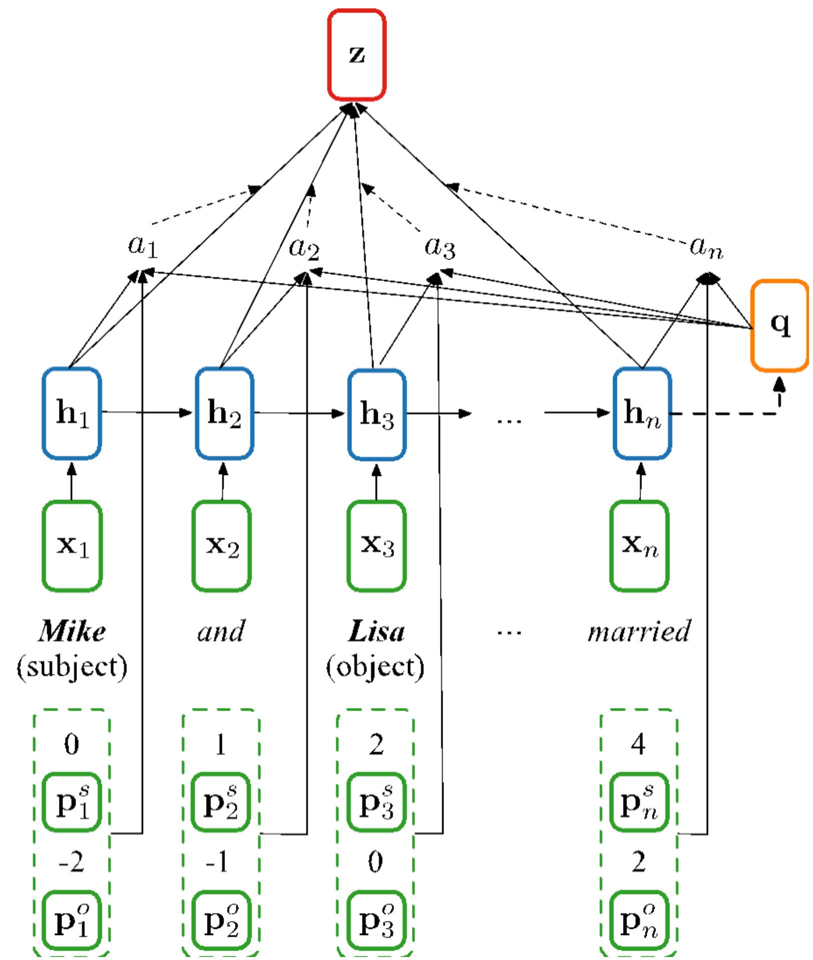
- Use our new position-aware NN relation extraction architecture (Zhang et al. EMNLP 2017)
- Needs supervised training data

  Summary vector: $\mathbf{q} = \mathbf{h}_n$
- Attention layer:

$$u_i = \mathbf{v}^\top \tanh(\mathbf{W}_h \mathbf{h}_i + \mathbf{W}_q \mathbf{q} + \mathbf{W}_s \mathbf{p}_i^s + \mathbf{W}_o \mathbf{p}_i^o)$$

$$a_i = \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)}$$

- Relations: $\mathbf{z} = \sum_{i=1}^{n} a_i \mathbf{h}_i$
- Softmax: $\mathbf{y} = \mathrm{softmax}(\mathbf{Wz})$



Results

- The neural system significantly outperforms the other systems
- Using multiple justifications increases recall at the expense of precision, results in a net decrease in average precision

KBP 2017	Relation Extraction	P	R	F1	AP (K=1)
English	Patterns only	19.9	18.1	17.6	16.4
	+ Supervised	20.3	21.9	19.5	19.0
	+ Neural system	22.7	27.5	22.6	21.6
	- Multiple justifications	24.0	26.4	23.1	21.9

The curious case of low macro-precision

- High precision systems were showing lower macro precision!

System	micro-precision	macro-precision
High Precision	**51.00**	18.91
High Recall	19.35	**21.14**

- **Reason** - All queries with no slot fills get zero precision. Reduces mean-precision over queries
- High precision systems often predict nothing for many queries. Their macro-precision gets penalized because of low recall
- **Proposed fix** - Compute mean precision only over queries with at least 1 proposed slot fill – then we get 59.5 macro-precision for high precision and 38.49 for high recall system

CS KB/SF 2017

- Common system architecture
- Entities
- English system
- **Chinese system**
- Spanish system
- Results

Chinese Extraction systems

- Pattern-based systems
    - TokensRegex + Semgrex
    - **(New)** Nested-mention extractor for headquarters
- Logistic regression trained using distant-supervision
- Other improvements:
    - An improved Chinese segmentation model
    - Improved extractor for subsidiaries

Results

- Including the distant supervision system helps a little bit.

KBP 2017	Relation Extraction	P	R	F1	AP (K=1)
Chinese	Patterns only	20.1	18.6	18.5	17.3
	+ Distant supervision	20.5	18.7	18.8	17.4
	- Multiple justifications	20.5	18.7	18.8	17.4

CS KB/SF 2017

- Common system architecture
- Entities
- English system
- Chinese system
- **Spanish system**
- Results

New Spanish slot filling system

- Built from scratch!

New Spanish slot filling system

- Made from 2,400+ TokensRegex and 500 Semgrex patterns.
  - These are our CoreNLP systems for regex-like patterns over token sequences and dependency trees respectively
  - TokensRegex (for per:title): `$ENTITY_PER /fue/ /elegido|elegida/ /como/ $TITLE`
  - Semgrex (for per:title) `{ner:/TITLE/}=slot >/cop/ {ner:/PERSON/}=entity`
- Trace ingredients:
  - HeidelTime for date-time expressions
  - Large fine-grained NER lexicon, some translated from English

New Spanish slot filling system

- Secret sauce: good syntactic dependencies using Dozat et al. (2017) neural POS tagger and UD parser (91.65% LAS)

The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted. Restart your computer, and then open the file again. If the red x still appears, you may have to delete the image and then insert it again.

New Spanish slot filling system

Semgrex patterns are able to generalize many different contexts!

KBP 2016 (dev)	P	R	F1
Best 2016 system	17.6	36.4	23.7
Tokensregex	19.8	3.4	5.9
+ Semgrex	17.5	10.0	12.6

Scores are very biased because 2016 data is extremely incomplete!

KBP 2017	Relation Extraction	P	R	F1	AP (K=1)
Spanish	Patterns only	14.4	14.9	13.7	13.4
	- Multiple justifications	15.2	15.2	14.4	13.8

CS KB/SF 2017

- Common system architecture
- Entities
- English system
- Chinese system
- Spanish system
- **Results**

Slot filling results and takeaways

- Tinkerbell (and Stanford) SF systems were amongst the top-ranked!
- Improved EDL performance leads to better slot filling.
- Neural relation extraction system leads to significant improvement in English slot filling scores.

Tinkerbell	P	R	F1	AP
English	23.4	31.3	24.7	13.9
Chinese	17.4	15.5	15.6	8.6
Spanish	14.8	15.8	14.3	9.8
Cross-lingual	17.3	19.9	16.8	9.3

TinkerBell

COLUMBIA UNIVERSITY

CORNELL UNIVERSITY FOUNDED A.D. 1865

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Rensselaer

Stanford University
S
NLP
Natural Language Processing

JOHNS HOPKINS UNIVERSITY

Penn
UNIVERSITY of PENNSYLVANIA