Lightweight Multilingual Entity Extraction and Linking

Speaker: Shih-Han Lo

Advisor: Professor Jia-Ling Koh

Author: Aasish Pappu, Roi Blanco, Yashar Mehdad, Amanda Stent, Kapil Thadani

Date: 2017/09/19

Source: WSDM ’17

Outline

- Introduction
- Method
- Experiment
- Conclusion

Introduction

- Key tasks for text analytic systems:
  - Named Entity Recognition (NER)
  - Named Entity Linking (NEL)
- Some systems perform NER and NEL jointly.

Introduction

Motivation

- Most approaches involve (some of) the following steps:
  - Mention detection
  - Mention normalization
  - Candidate entity retrieval for each mention
  - Entity disambiguation for mentions with multiple candidate entities
  - Mention clustering for mentions that do not link to any entity

Outline

- Introduction
- Method
- Experiment
- Conclusion

Mention Detection

- Typically consists of running an NER system over input text.
- We use simple CRFs and only a few lexical, syntactic and semantic features.

System Description

Feature	Description
Tokens	w_i for i in $\{-2, ..., +2\}$, $w_i \& w_{i+1}$ for i in $\{-1, 0\}$
Embeddings	$emb[100]$ for i in $\{-2, ..., +2\}$
Morphological	$morpho_i$ for i in $\{-2, ..., +2\}$
POS	pos_i for i in $\{-2, ..., +2\}$, $pos_i \& pos_{i+1}$ for i in $\{-2, .., 1\}$

Features	**EN**			**ES**			**ZH**		
	P	R	F1	P	R	F1	P	R	F1
Token + Embeddings	91	82	86	86	79	82	76	54	64
+ POS	90	87	88	86	80	83	77	54	65
+ Morphological	90	88	89	85	84	85	74	60	67
+ POS + Morphological	89	88	89	85	84	84	75	61	67

Systems	**EN**	**ES**	**ZH**
This Work	88.6	84.6	67.2
Al-Rfou et al. [1]†	71.3	63.0	-
Stanford [17]*	86.3	81.1	64.1/69.5
Suzuki and Isozaki [48]	89.9	-	-
Che et al. [7]*	-	-	64.1/69.5
Lample et al. [29]+	90.9	85.8	-
Ma and Hovy [33]+	91.2	-	-
Luo et al. [32]*	91.2	-	-

Candidate Entity Retrieval

- **Entity Embeddings**

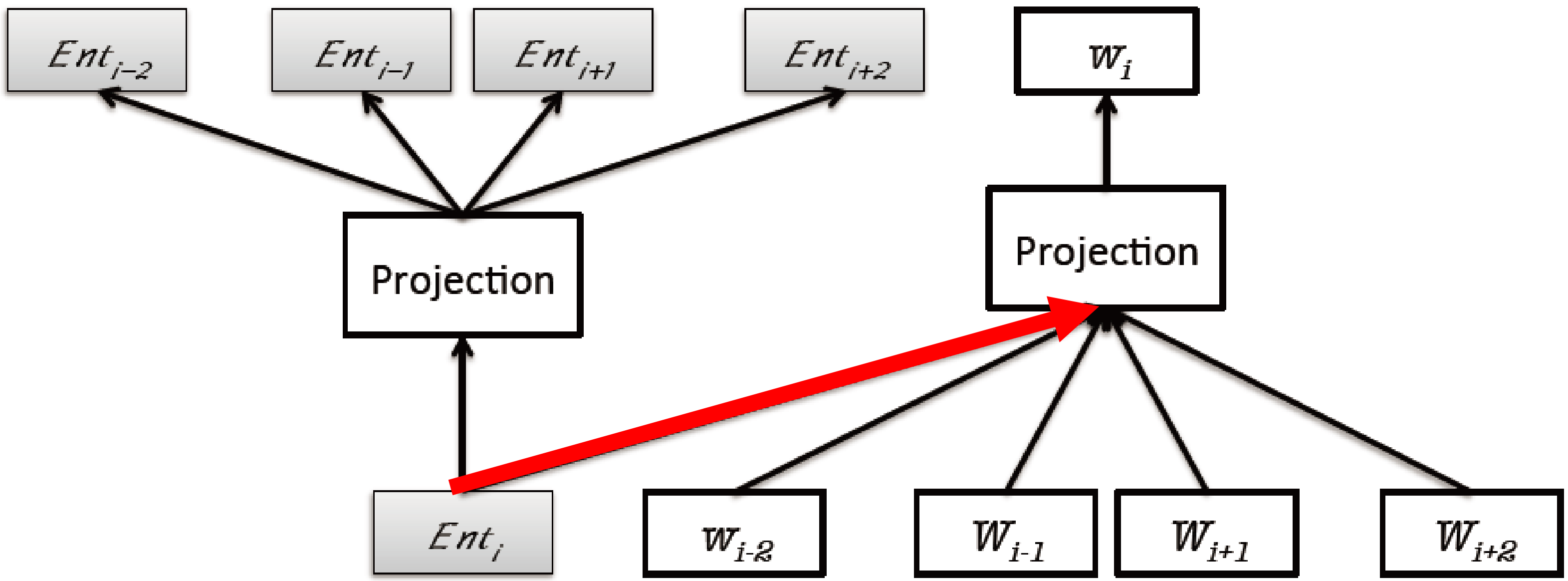
$$(ent_1, ent_2, \ldots, ent_n), \text{ where } ent_i \in Ent$$

$$(w_1, w_2 \ldots, w_m), \text{ where } w_j \in W$$

  - We aim to simultaneously learn *D*-dimensional representations of *Ent* and *W* in a common vector space.
  - Training our embedding model: continuous skip-grams with 300 dimensions and a window size of 10.

Candidate Entity Retrieval

- **Entity Embeddings**



Candidate Entity Retrieval

- **Fast Entity Linking**
  - Fast Entity Linker (FEL) is an unsupervised approach.
  - FEL imposes contextual dependencies by calculating the cosine distance between two entities.
    - Candidate ⇔ From the substrings of the input string

  - Minimal perfect hash function
  - Elias-Fano integer coding

Entity Disambiguation

- Task of figuring out to which candidate entity a mention refers.
- The task is complex because mentions may refer to different entities, depend on local context.

Entity Disambiguation

- **Forward-Backward Algorithm (FwBw)**

Algorithm 1 ForwardBackward

Input: $M \leftarrow$ mentions, $NB \leftarrow$ N-BestLinks,
$P \leftarrow$ Posterior probability from NB
Output: $\hat{L} \leftarrow$ 1-best Entities
procedure FwBw
  $fwd \leftarrow$ Forward(NB, M)
  $bkwd \leftarrow$ Forward(NB_{rev}, M_{rev})
  for $i \leftarrow 1, 3, ..., |M|$ **do**
    $\hat{L}_i \leftarrow \arg\max_k(fwd_{i,k} \cdot bkwd_{|M|-i,k})$
  end for
    return $\hat{L}_{1,2..,i,..|M|}$
end procedure
procedure JOINT_SIM(u,v)
  $sem \leftarrow$`semSim`(u, v), $lex \leftarrow$`textSim`(u, v)
    return $(\lambda \cdot sem + (1-\lambda) \cdot lex)$
end procedure
procedure Forward
  for l_i in NB_1 **do**
    $S_{i,1} \leftarrow$ JOINT_SIM(l_i, M_1)
    $\theta_{0,i} = P(l_i, M_1) \cdot S_{l_i,M_1}$
  end for
  for $i \leftarrow 2, 3, ..., |M|$ **do**
    for each link l_j **do**
      $S_{M_i,l_j} \leftarrow$ JOINT_SIM(M_i, l_j)
    $\theta_{j,i} \leftarrow \max_k(\theta_{k,i-1} \cdot S_{M_i,l_j} \cdot S_{l_k,l_j} \cdot P(M_i, l_k))$
    end for
  end for
    return θ
end procedure

Entity Disambiguation

- **Exemplar (Clustering)**

Algorithm 2 Exemplar Clustering

Input: M, NB, $pref_{1\times n} \leftarrow$ Posterior probability from N-BestLinks

Output: $\hat{L} \leftarrow$ 1-best Entities

$X_{n\times d} \leftarrow embeddings(M) \oplus embeddings(NB)$

$S_{n\times n} \leftarrow pairwiseSim(X)$

$R_{n\times n}, A_{n\times n} \leftarrow zeros, zeros$

$diag(S) \leftarrow diag(S) + pref$

λ is *damping factor* to discourage oscillations

while convergence `OR` $T \leq max\_iterations$ **do**

  $R_{i,k} \leftarrow S_{i,k} - \max_{k' \neq k}\{A_{i,k'} + S_{i,k'})\}$

  $A_{i,k} \leftarrow \min\left(0, A_{k,k} + \sum_{i' \notin \{i,k\}} max(0, R_{i',k})\right)$

  $A_{k,k} \leftarrow \sum_{i' \neq k} max(0, R_{i',k})$

end while

$I \leftarrow R_{i,i} + A_{i,i} > 0$

$CI = \arg\max_{k \in I} S_{k,k}$

  return $\hat{L} \leftarrow (\forall_{k \in |CI|} CI_k)$

Entity Disambiguation

- **Label Propagation (LabelProp)**
  - Modified adsorption (MAD)
  - For $G \leftarrow (V, E_w)$, we inject seed labels L on a few nodes.
  - For nodes V', we assign a label distribution: $\{l_1 : p_1, l_2 : p_2, \ldots, l_n : p_n\}$
  - Along with $\{L,\ G\}$, MAD takes three hyper-parameters $\{\mu_1, \mu_2, \mu_3\}$ as input.

  - We pick the highest ranked label for each node in V as the final candidate.

Outline

- Introduction
- Method
- Experiment
- Conclusion

Experiment

- **Datasets:**
  - Cross-lingual TAC KBP 2013
  - Mono-lingual AIDA-CONLL 2003

Data	Docs	Entities	Unique entities	Mentions
KBP-EN	1820	1183	349	150144
KBP-ES	1175	1305	583	6321
KBP-ZH	1224	1229	159	15092
AIDA-all	1392	37922	5598	50758

Experiment

- **Setup**
  - N-best: N = 10
  - **FwBw**: $\lambda = 0.5$
  - **Exemplar**: max_iterations = 300, $\lambda = 0.5$
  - **LabelProp**: $\mu_1 = 1, \mu_2 = 1e-2, \mu_3 = 1e-2$

Experiment

- **TAC KBP Evaluation Results**

Dataset	1-best		FwBw		Exemplar		LabelProp		BasisTech
	KNN	FEL	KNN	FEL	KNN	FEL	KNN	FEL	
KBP-EN	32.0	50.6	29.1	**61.0**	52.6	52.8	29.8	53.6	56.5
KBP-ES	31.3	50.8	27.7	46.7	24.0	50.5	28.5	48.3	**61.2**
KBP-ZH	17.0	**67.3**	7.5	54.7	9.8	57.5	12.3	49.8	62.1

Experiment

- **Analysis**

Dataset	#docs	#words per doc	Precision				Recall				F_1			
			1-best	FwBw	Exemplar	LabelProp	1-best	FwBw	Exemplar	LabelProp	1-best	FwBw	Exemplar	LabelProp
KBP-EN	1820	3118.1	42.5	**62.1**	45.3	59.9	62.6	59.9	**63.1**	48.5	50.6	**61.0**	52.8	53.6
├ News	924	300.6	42.8	**64.4**	44.9	62.6	66.3	61.7	**66.8**	61.7	52.0	**63.0**	53.7	62.1
├ Forums	607	7555.9	50.1	**64.3**	54.9	61.3	**60.4**	58.5	**60.4**	31.9	54.8	**61.3**	57.5	41.9
└ Newsgroups	288	2813.8	24.6	**46.5**	26.7	45.4	53.4	**56.8**	55.9	50.0	33.7	**51.2**	36.2	47.6
KBP-ES	1175	168.7	60.5	67.4	**71.0**	62.5	**43.8**	35.7	39.2	39.4	**50.8**	46.7	50.5	48.3
├ Spanish news	775	160.5	58.1	**65.3**	64.5	60.3	**37.9**	30.6	26.9	32.4	**45.9**	41.7	37.9	42.2
└ English news	397	180.7	64.3	70.8	**74.8**	65.5	**56.3**	46.3	42.8	53.9	**60.0**	56.0	54.4	59.1
KBP-ZH	1224	752.5	74.3	75.5	**77.1**	61.8	**61.5**	42.9	45.8	41.7	**67.3**	54.7	57.5	49.8
├ Newsgroups	415	1215.9	78.6	74.4	**81.0**	59.1	**66.6**	43.7	48.9	38.9	**72.1**	55.0	61.0	46.9
├ Chinese news	406	323.0	70.5	76.9	**82.6**	55.2	**51.7**	40.2	48.0	33.7	59.7	52.8	**60.8**	41.9
├ English news	230	217.9	71.4	71.4	52.7	**71.7**	**69.6**	43.5	30.0	60.4	**70.5**	54.1	38.2	65.6
└ Blogs	173	1360.0	75.9	81.0	**85.5**	65.8	**61.5**	46.6	54.0	42.0	**67.9**	59.1	66.2	51.2

Experiment

- **Analysis**

Measure	Lang	1-best	FwBw	Exemplar	LabelProp
#words	EN	-5.2	-4.9	-5.9	-41.3
	ES	6.0	-4.8	2.5	-3.9
	ZH	3.8	-2.9	6.5	-19.8
#mentions	EN	-1.2	-2.7	-1.6	-44.8
	ES	16.0	12.3	12.7	1.1
	ZH	-1.3	-4.3	8.9	-22.2
#mentions per word	EN	16.5	15.9	15.1	37.8
	ES	13.1	22.6	18.5	15.7
	ZH	6.0	12.1	16.2	4.8

Dataset	Precision		Recall		F_1	
	KNN	FEL	KNN	FEL	KNN	FEL
KBP-EN	**55.4**	45.3	50.1	**63.1**	52.6	**52.8**
├ News	**53.6**	44.9	49.2	**66.8**	51.3	**53.7**
├ Forums	**65.5**	54.9	53.5	**60.4**	**58.7**	57.5
└ Newsgroups	**34.3**	26.7	41.5	**55.9**	**37.5**	36.2

Experiment

- **AIDA Evaluation**

System	A_{macro}	A_{micro}
1-best	83.48	81.07
FwBw	**83.63**	80.98
Exemplar	83.50	81.08
Alhelbawy and Gaizauskas [2]	82.80	**86.10**
Cucerzan [10]	43.74	51.03
Kulkarni et al. [27]	76.74	72.87
Hoffart et al.[25]	81.91	81.82
Shirakawa et al. [45]	83.02	82.29
He et al. [24]	83.37	84.82

Experiment

- **Runtime Performance**

	# Entities	Data pack	# Vectors	Wiki
EN	4.9M	1.6GB	1.5GB	45GB
ES	1.10M	114M	877MB	9.8GB
ZH	870K	272MB	864MB	5.3GB

Datasets	Docs	Average mentions	Sec/doc
AIDA-all	1392	36.43	0.178
KBP-EN	1820	82.4	0.473
KBP-ES	1175	5.37	0.004
KBP-ZH	1224	14.54	0.013

Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- Our NER implementation is outperformed only by NER systems that use much more complex feature engineering and/or modeling methods.
- In future work, we plan to improve the performance of our system for other languages, by expanding the pool of entities for which we have information.
  - Candidate retrieval in Spanish is relatively poor compared to English and Chinese.