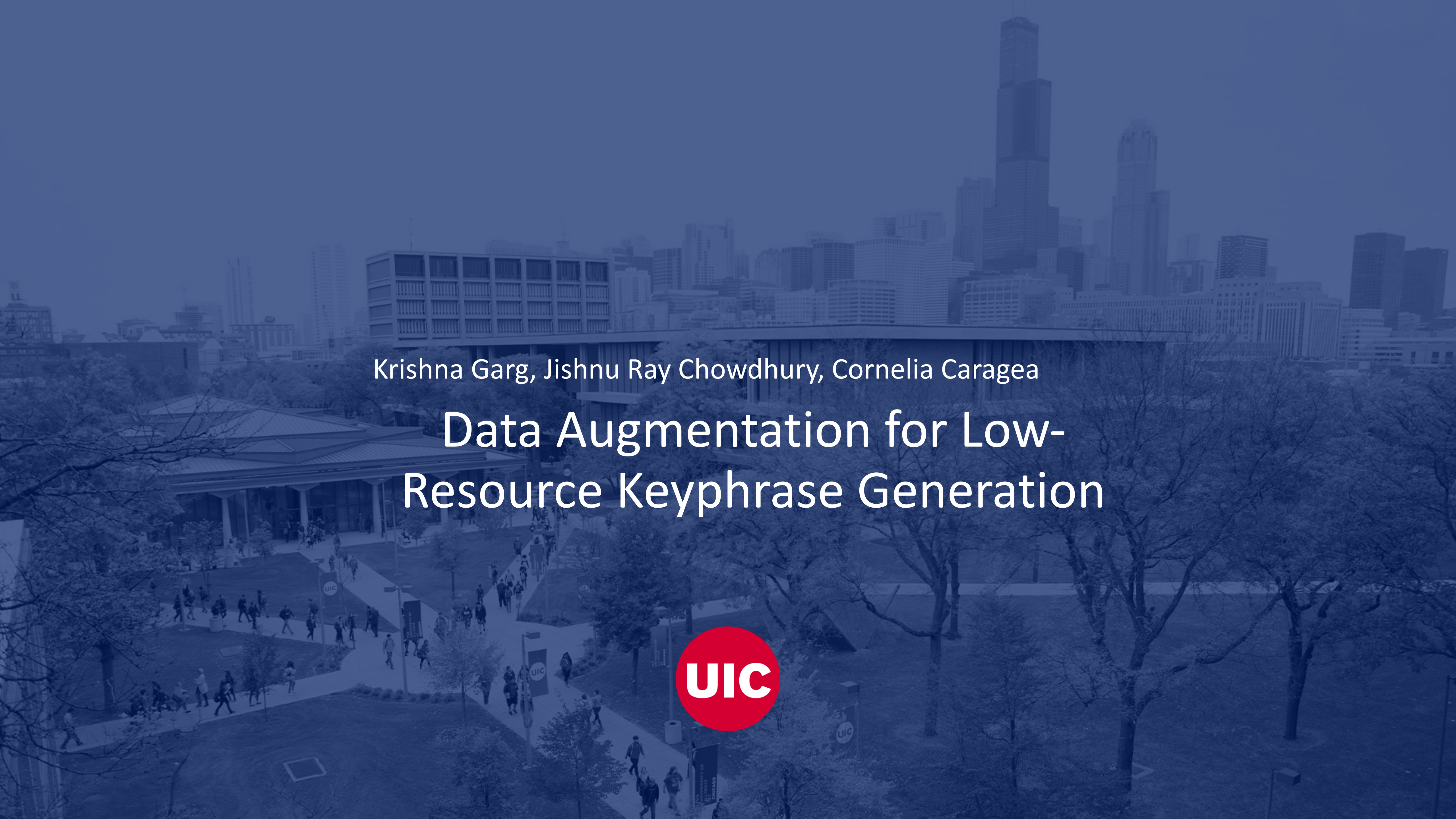

Krishna Garg, Jishnu Ray Chowdhury, Cornelia Caragea

Data Augmentation for Low-Resource Keyphrase Generation

Keyphrase Generation Task

Ranking-based Method for News Stance Detection

Qiang Zhang
University College London
London, United Kingdom
qiang.zhang.16@ucl.ac.uk

Emine Yilmaz
University College London
London, United Kingdom
emine.yilmaz@ucl.ac.uk

Shangsong Liang
KAUST
Thuwal, Saudi Arabia
shangsong.liang@kaust.edu.sa

ABSTRACT

A valuable step towards news veracity assessment is to understand stance from different information sources, and the process is known as the stance detection. Specifically, the stance detection is to detect four kinds of stances ("agree", "disagree", "discuss" and "unrelated") of the news towards a claim. Existing methods tried to tackle the stance detection problem by classification-based algorithms. However, classification-based algorithms make a strong assumption that there is clear distinction between any two stances, which may not be held in the context of stance detection. Accordingly, we frame the detection problem as a ranking problem and propose a ranking-based method to improve detection performance. Compared with the classification-based methods, the ranking-based method compare the true stance and false stances and maximize the difference between them. Experimental results demonstrate the effectiveness of our proposed method.

CCS CONCEPTS

• **Computing methodologies → Information extraction**;

KEYWORDS

Fake news; stance detection; learning to rank

Stance detection has been suggested as a crucial first step to detect fake news. [1] Researchers from both academia and industry initiated the Fake News Challenge (FNC) [1] and the first stage of FNC (FNC-1) aimed to accelerate the establishment of automatic systems for evaluating the positions that a news source holds about a particular claim. More specifically, given a news headline as a claim and its article body, FNC-1 tried to develop models to estimate the stance of the article body towards its headline. The stance could be one of the labels: "agree", "disagree", "discuss" and "unrelated". All the news with the "agree", "disagree" and "discuss" are assumed as "related". According to the FNC-1, formal definitions of the four stances are as: "**Agree**" – the body text agrees with the headline; "**Disagree**" – the body text disagrees with the headline; "**Discuss**" – the body text discusses the same claim as the headline, but does not take a position; and "**Unrelated**" – the body text discusses a different claim but not that in the headline.

A GradientBoosting classifier is implemented as the FNC-1 official baseline with a relative score of 75.20%. This classifier makes use of semantic analysis and overlap between headlines and bodies. [1] As for the FNC-1 submissions, an ensemble of convolutional neural network and gradient-boosted decision trees achieves the highest detection performance [2]. Another ensemble of five multi-layer perceptrons (MLP) achieves a little worse accuracy. These two

Applications

- Reviewer matching systems for conferences/ journals
- Mining literature
- Recommendations to readers

Task

- Input:
  - Title || Abstract || Sent_1 ||Sent_2 || …. Sent_k
- Output
  - kp_1 || kp_2 || … || kp_n

Motivation

- Lack of annotated data for different domains
- Semi- or Un-supervised methods may not work
    - Automatic annotation could be inaccurate
    - Unlabeled data may not be available
- Garg et al. (2022) showed benefits of using body
    - Body could be useful for data augmentation methods

Contributions

- Propose data augmentation techniques for “purely” low-resource domains
- Demonstrate effectiveness for three datasets

Data Augmentation Methods

- Synonym Replacement (SR)
  - Replace 10% words randomly with the synonyms from WordNet
- BackTranslation (BT)
  - English -> French -> English
- Keyphrase Dropout (KPD)
  - Mask keyphrases – intuition similar to Masked Language Modeling
- Keyphrase Synonym Replacement (KPSR)
  - Replace keyphrases with the synonyms

Concrete Example

Methods	Excerpts from different data augmentation methods
TITLE ‖ ABSTRACT	casesian : a knowledge-based system using statistical and experiential perspectives for improving the knowledge sharing in the medical prescription process [SEP] objectives : knowledge sharing is crucial for better patient care in the healthcare industry
AUG_TA_SR	casesian : a knowledge based system using statistical and experiential perspectives for better the knowledge sharing in the medical examination prescription [SEP] objectives : knowledge sharing is crucial for advantageously patient role care in the healthcare industry
AUG_TA_BT	cassian : a knowledge-based system that uses statistical and experiential perspectives to improve the sharing of knowledge in the medical prescription process [SEP] objectives : knowledge sharing is essential to improve patient care in the health sector
AUG_TA_KPD	casesian : a [MASK] using statistical and experiential perspectives for improving the [MASK] in the [MASK] process [SEP] objectives : [MASK] is crucial for better patient care in the healthcare industry
AUG_TA_KPSR	casesian : a cognition based system using statistical and experiential perspectives for improving the noesis sharing in the checkup prescription process [SEP] objectives : noesis sharing is crucial for better patient care in the healthcare industry
AUG_BODY	numerous methods have been investigated for improving the knowledge sharing process in medical prescription [SEP] case-based reasoning is one of the most prevalent knowledge extraction methods
GOLD KEYPHRASES	case-based reasoning , medical prescription , knowledge-based system , knowledge sharing , bayesian theorem

Table 1: An example depicting different augmentation methods used in the paper. The text is highlighted as follows: DIVERSITY introduced in the augmented samples , ABSENT KEYPHRASES , PRESENT KEYPHRASES (highlighted only in TITLE ‖ ABSTRACT for brevity). Note that all AUG prefixed methods augment as a separate article to the original article T ‖ A. For specific details about each method, please refer to §3.2. Best viewed in color.

Baselines

- T || A
- T || A || Body
  - Concatenation of all three
- Aug_TA_xxx
  - First article: T || A
  - Second article: Modified T || A using data augmentation methods
- Aug_Body
  - First article: T || A
  - Second article: BODY
- Aug_Body_xxx
  - First article: T || A
  - Second article: Modified BODY using data augmentation methods

Datasets

Datasets	#Train	#Dev	#Test
LDKP3K	50,000	3,339	3,413
LDKP10K	50,000	10,000	10,000
KPTimes	259,923	10,000	20,000

Results

Results – Present Keyphrase (LDKP3K dataset)

LDKP3K	1,000		2,000		4,000		8,000	
	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M
T || A	4.68_1	9.10_6	6.19_1	11.89_2	9.67_2	18.47_8	11.97_1	22.86_1
T || A || Body	4.94_1	9.55_5	5.99_2	11.61_2	10.14_1	19.57_0	12.30_0	$\mathbf{23.53}_0$
AUG_TA_SR	4.75_1	9.34_3	6.66_2	12.74_0	9.19_3	17.65_{10}	11.37_0	21.95_0
AUG_TA_BT	4.41_1	8.62_2	6.32_2	12.27_3	10.42_0	19.96_1	$\mathbf{12.34}_0$	23.32_2
AUG_TA_KPD	4.67_1	9.19_1	6.00_0	11.63_1	7.92_2	15.48_5	10.53_0	20.55_1
AUG_TA_KPSR	4.55_0	8.95_1	5.70_1	10.90_1	7.14_1	13.87_5	9.33_0	18.29_1
AUG_Body	$\mathbf{5.33}_2$	$\mathbf{10.42}_5$	$\mathbf{7.10}_6$	$\mathbf{13.92}_{18}$	9.97_5	19.25_{18}	11.82_2	22.67_4
AUG_Body_SR	4.88_1	9.69_4	6.50_0	12.53_2	9.36_9	18.15_{30}	12.19_1	23.04_3
AUG_Body_BT	4.59_0	9.04_2	6.36_3	12.26_5	$\mathbf{10.50}_0$	$\mathbf{20.09}_1$	12.31_1	23.19_3
AUG_Body_KPD	4.72_2	9.31_6	6.12_1	11.92_3	8.82_7	17.04_{18}	11.61_0	22.14_1
AUG_Body_KPSR	4.60_0	9.15_1	5.78_1	11.21_6	7.44_2	14.60_8	11.40_1	21.64_3

1. Augmentation with body (AUG_Body) is useful

Results – Present Keyphrase (LDKP3K dataset)

LDKP3K	1,000		2,000		4,000		8,000	
	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M
T || A	4.68_1	9.10_6	6.19_1	11.89_2	9.67_2	18.47_8	11.97_1	22.86_1
T || A || Body	4.94_1	9.55_5	5.99_2	11.61_2	10.14_1	19.57_0	12.30_0	$\mathbf{23.53}_0$
AUG_TA_SR	4.75_1	9.34_3	6.66_2	12.74_0	9.19_3	17.65_{10}	11.37_0	21.95_0
AUG_TA_BT	4.41_1	8.62_2	6.32_2	12.27_3	10.42_0	19.96_1	$\mathbf{12.34}_0$	23.32_2
AUG_TA_KPD	4.67_1	9.19_1	6.00_0	11.63_1	7.92_2	15.48_5	10.53_0	20.55_1
AUG_TA_KPSR	4.55_0	8.95_1	5.70_1	10.90_1	7.14_1	13.87_5	9.33_0	18.29_1
AUG_Body	$\mathbf{5.33}_2$	$\mathbf{10.42}_5$	$\mathbf{7.10}_6$	$\mathbf{13.92}_{18}$	9.97_5	19.25_{18}	11.82_2	22.67_4
AUG_Body_SR	4.88_1	9.69_4	6.50_0	12.53_2	9.36_9	18.15_{30}	12.19_1	23.04_3
AUG_Body_BT	4.59_0	9.04_2	6.36_3	12.26_5	$\mathbf{10.50}_0$	$\mathbf{20.09}_1$	12.31_1	23.19_3
AUG_Body_KPD	4.72_2	9.31_6	6.12_1	11.92_3	8.82_7	17.04_{18}	11.61_0	22.14_1
AUG_Body_KPSR	4.60_0	9.15_1	5.78_1	11.21_6	7.44_2	14.60_8	11.40_1	21.64_3

2. AUG_Body_xxx > AUG_TA_xxx

Results – Present Keyphrase (LDKP3K dataset)

LDKP3K	1,000		2,000		4,000		8,000	
	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M
T || A	4.68_{1}	9.10_{6}	6.19_{1}	11.89_{2}	9.67_{2}	18.47_{8}	11.97_{1}	22.86_{1}
T || A || Body	4.94_{1}	9.55_{5}	5.99_{2}	11.61_{2}	10.14_{1}	19.57_{0}	12.30_{0}	$\mathbf{23.53}_{0}$
AUG_TA_SR	4.75_{1}	9.34_{3}	6.66_{2}	12.74_{0}	9.19_{3}	17.65_{10}	11.37_{0}	21.95_{0}
AUG_TA_BT	4.41_{1}	8.62_{2}	6.32_{2}	12.27_{3}	10.42_{0}	19.96_{1}	$\mathbf{12.34}_{0}$	23.32_{2}
AUG_TA_KPD	4.67_{1}	9.19_{1}	6.00_{0}	11.63_{1}	7.92_{2}	15.48_{5}	10.53_{0}	20.55_{1}
AUG_TA_KPSR	4.55_{0}	8.95_{1}	5.70_{1}	10.90_{1}	7.14_{1}	13.87_{5}	9.33_{0}	18.29_{1}
AUG_Body	$\mathbf{5.33}_{2}$	$\mathbf{10.42}_{5}$	$\mathbf{7.10}_{6}$	$\mathbf{13.92}_{18}$	9.97_{5}	19.25_{18}	11.82_{2}	22.67_{4}
AUG_Body_SR	4.88_{1}	9.69_{4}	6.50_{0}	12.53_{2}	9.36_{9}	18.15_{30}	12.19_{1}	23.04_{3}
AUG_Body_BT	4.59_{0}	9.04_{2}	6.36_{3}	12.26_{5}	$\mathbf{10.50}_{0}$	$\mathbf{20.09}_{1}$	12.31_{1}	23.19_{3}
AUG_Body_KPD	4.72_{2}	9.31_{6}	6.12_{1}	11.92_{3}	8.82_{7}	17.04_{18}	11.61_{0}	22.14_{1}
AUG_Body_KPSR	4.60_{0}	9.15_{1}	5.78_{1}	11.21_{6}	7.44_{2}	14.60_{8}	11.40_{1}	21.64_{3}

3. AUG_Body_xxx > T || A || Body (Garg et al.)

Results – Present Keyphrase (LDKP3K dataset)

LDKP3K	1,000		2,000		4,000		8,000	
	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M
T || A	4.68_1	9.10_6	6.19_1	11.89_2	9.67_2	18.47_8	11.97_1	22.86_1
T || A || Body	4.94_1	9.55_5	5.99_2	11.61_2	10.14_1	19.57_0	12.30_0	$\mathbf{23.53}_0$
AUG_TA_SR	4.75_1	9.34_3	6.66_2	12.74_0	9.19_3	17.65_{10}	11.37_0	21.95_0
AUG_TA_BT	4.41_1	8.62_2	6.32_2	12.27_3	10.42_0	19.96_1	$\mathbf{12.34}_0$	23.32_2
AUG_TA_KPD	4.67_1	9.19_1	6.00_0	11.63_1	7.92_2	15.48_5	10.53_0	20.55_1
AUG_TA_KPSR	4.55_0	8.95_1	5.70_1	10.90_1	7.14_1	13.87_5	9.33_0	18.29_1
AUG_Body	$\mathbf{5.33}_2$	$\mathbf{10.42}_5$	$\mathbf{7.10}_6$	$\mathbf{13.92}_{18}$	9.97_5	19.25_{18}	11.82_2	22.67_4
AUG_Body_SR	4.88_1	9.69_4	6.50_0	12.53_2	9.36_9	18.15_{30}	12.19_1	23.04_3
AUG_Body_BT	4.59_0	9.04_2	6.36_3	12.26_5	$\mathbf{10.50}_0$	$\mathbf{20.09}_1$	12.31_1	23.19_3
AUG_Body_KPD	4.72_2	9.31_6	6.12_1	11.92_3	8.82_7	17.04_{18}	11.61_0	22.14_1
AUG_Body_KPSR	4.60_0	9.15_1	5.78_1	11.21_6	7.44_2	14.60_8	11.40_1	21.64_3

4. Standard SR, BT > Keyphrase specific KPD, KPSR

Results – Present Keyphrase (LDKP3K dataset)

LDKP3K	1,000		2,000		4,000		8,000	
	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M
T || A	4.68_1	9.10_6	6.19_1	11.89_2	9.67_2	18.47_8	11.97_1	22.86_1
T || A || Body	4.94_1	9.55_5	5.99_2	11.61_2	10.14_1	19.57_0	12.30_0	$\mathbf{23.53}_0$
AUG_TA_SR	4.75_1	9.34_3	6.66_2	12.74_0	9.19_3	17.65_{10}	11.37_0	21.95_0
AUG_TA_BT	4.41_1	8.62_2	6.32_2	12.27_3	10.42_0	19.96_1	$\mathbf{12.34}_0$	23.32_2
AUG_TA_KPD	4.67_1	9.19_1	6.00_0	11.63_1	7.92_2	15.48_5	10.53_0	20.55_1
AUG_TA_KPSR	4.55_0	8.95_1	5.70_1	10.90_1	7.14_1	13.87_5	9.33_0	18.29_1
AUG_Body	$\mathbf{5.33}_2$	$\mathbf{10.42}_5$	$\mathbf{7.10}_6$	$\mathbf{13.92}_{18}$	9.97_5	19.25_{18}	11.82_2	22.67_4
AUG_Body_SR	4.88_1	9.69_4	6.50_0	12.53_2	9.36_9	18.15_{30}	12.19_1	23.04_3
AUG_Body_BT	4.59_0	9.04_2	6.36_3	12.26_5	$\mathbf{10.50}_0$	$\mathbf{20.09}_1$	12.31_1	23.19_3
AUG_Body_KPD	4.72_2	9.31_6	6.12_1	11.92_3	8.82_7	17.04_{18}	11.61_0	22.14_1
AUG_Body_KPSR	4.60_0	9.15_1	5.78_1	11.21_6	7.44_2	14.60_8	11.40_1	21.64_3

5. Keyphrase-specific DA techniques hurt the performance

Results – Absent Keyphrase (LDKP3K dataset)

LDKP3K	1,000		2,000		4,000		8,000	
	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M	F1@5	F1@M
T || A	0.078_0	0.169_0	0.129_0	0.281_0	0.044_0	0.093_0	0.044_0	0.099_0
T || A || Body	0.079_0	0.165_0	0.130_0	0.282_0	0.047_0	0.105_0	0.031_0	0.073_0
AUG_TA_SR	0.132_0	0.290_0	0.136_0	0.300_0	0.096_0	0.207_0	0.067_0	0.141_0
AUG_TA_BT	0.128_0	0.279_0	0.139_0	0.305_0	0.068_0	0.140_0	0.121_0	0.266_0
AUG_TA_KPD	0.140_0	0.311_0	0.145_0	0.318_0	0.141_0	0.307_0	0.099_0	0.218_0
AUG_TA_KPSR	0.142_0	0.307_0	0.177_0	0.393_0	0.151_0	0.321_0	0.154_0	0.325_0
AUG_Body	0.129_0	0.291_0	0.130_0	0.292_0	0.061_0	0.138_0	0.079_0	0.175_0
AUG_Body_SR	0.141_0	0.319_0	0.157_0	0.342_0	0.076_0	0.161_0	0.149_0	0.322_0
AUG_Body_BT	0.130_0	0.287_0	0.121_0	0.265_0	0.081_0	0.183_0	0.120_0	0.253_0
AUG_Body_KPD	0.144_0	0.328_0	0.189_0	0.407_0	0.136_0	0.298_0	0.182_0	0.398_0
AUG_Body_KPSR	$\mathbf{0.162}_0$	$\mathbf{0.359}_0$	$\mathbf{0.200}_0$	$\mathbf{0.441}_0$	$\mathbf{0.184}_0$	$\mathbf{0.405}_0$	$\mathbf{0.227}_0$	$\mathbf{0.495}_0$

1. Body is still useful
2. Unlike present KP, all DA techniques are useful here
3. Keyphrase specific KPD, KPSR > Standard SR, BT
4. KPSR is the best
5. Garg et al. method (T || A || Body) weakens the performance

Sample Predictions

Excerpts from test dataset samples	Methods	Predicted Keyphrases
committees of learning agents [SEP] we describe how machine learning and decision theory is combined in an application that supports control room operators of a combined heating and power plant ... **Gold:** machine learning ; committees ; decision analysis	T || A	learning
	Aug_Body	machine learning
	Aug_Body_SR	learning
compositional analysis for linear control systems [SEP] the complexity of physical and engineering systems , both in terms of the governing physical phenomena and the number of subprocesses involved ... **Gold:** compositional reasoning ; linear systems ; simulation relations ; assume-guarantee reasoning	T || A	control
	Aug_Body	linear control; linear systems
	Aug_Body_SR	linear control; linear systems
the bits and flops of the n-hop multilateration primitive for node localization problems [SEP] the recent advances in mems , embedded systems and wireless communication technologies are making the realization ... **Gold:** technologies ; ad-hoc localization ; sensor networks ; embedded systems ; wireless ; network	T || A	tangible
	Aug_Body	wireless networks
	Aug_Body_SR	sensors

Table 6: Sample predictions using models trained with different (representative) augmentation methods and the baseline (T || A). The text is highlighted as follows: PRESENT KEYPHRASES , ABSENT KEYPHRASES . Note that the test samples contain only T || A. Best viewed in color.

Analysis – LDKP3K dataset with 1000 samples

Methods	Pres.KP	Abs.KP	TotalKP
T || A	3374	2093	5467
T || A || Body	3985	1482	5467
AUG_TA_SR	5761	5173	10934
AUG_TA_BT	5499	5435	10934
AUG_TA_KPD	4586	6348	10934
AUG_TA_KPSR	4532	6402	10934
AUG_Body	**6309**	4625	10934
AUG_Body_SR	5402	5532	10934
AUG_Body_BT	5291	5643	10934
AUG_Body_KPD	4590	**6344**	10934
AUG_Body_KPSR	4591	6343	10934

Table 7: Number of present, absent, total keyphrases in the training set of LDKP3K with 1000 samples for the different augmentation methods.

1. TotalKP doubles for all DA techniques
2. AUG_Body has highest Present Keyphrases
3. Keyphrase-specific KPSR, KPD rich in Absent KPs
4. Standard SR, BT rich in Present KPs

Inference Strategies & Potential Future Work

Methods	Present		Absent	
	F1@5	F1@M	F1@5	F1@M
T || A	4.68	9.10	0.078	0.169
AUG_TA_BT	4.41	8.62	0.128	0.279
AUG_TA_KPSR	4.55	8.95	0.132	0.290
AUG_Body	**5.33**	**10.42**	0.129	0.291
AUG_Body_BT	4.59	9.04	0.130	0.287
AUG_Body_KPD	4.72	9.31	0.144	0.328
AUG_Body_KPSR	4.60	9.15	**0.162**	**0.359**
Inference Strategies				
Body ∪ Body-KPSR	6.41	**11.95**	0.196	0.428
TA-BT ∪ Body-BT	5.39	10.19	0.160	0.342
TA-KPSR ∪ Body-KPSR	6.17	11.47	**0.220**	**0.462**
Body-BT ∪ Body-KPD	**6.45**	11.81	0.204	0.435
Body-KPSR ∪ Body-KPD	5.94	11.18	0.204	0.444

Table 8: A comparison of various Inference Strategies using *Union* (see §6) with the individual (AUG_) methods on LDKP3K with 1000 samples in the training set.

Conclusion

- Data augmentation techniques including Standard & Keyphrase-specific
- Demonstrate effectiveness for 12 different settings (3 datasets * 4 low-resource settings)
- Leverage full text of the articles
- Keyphrase-specific DA techniques work well for Absent Keyphrases, while standard DA techniques work well for Present Keyphrases
  - Future work includes developing DA techniques which perform good for both