

# *An Evaluation of Kea--A Keyphrase Extraction System*

School of Communication, Information, and Library studies

Lu Liu



# *Introduction*

- Keyphrase: an important means of document indexing, summarization, browsing, and clustering.
  - Problem: costly and time consuming to manually assign keyphrases to documents.
  - Kea: a system that automatically extract keyphrases from document
  - Kea uses supervised learning method to extract keyphrases based on a set of documents and their manually assigned keyphrases
- 
-

# *The Purpose of This Study*

- Purpose: evaluate Kea algorithm
  - Use a micro-biology corpus—GENIA for evaluation
  - The advantages
    - Keyphrases for training appear identically in the text of the documents in the GENIA corpus
    - GENIA provides enough keyphrases for training
    - Domain specificity
  - The disadvantage
    - Lack of human assessment
- 
-

# Kea Algorithm

- Generating candidate phrases
  - Calculating Feature Value
    - TFIDF
    - The position of the candidate phrase's first appearance (pos)
  - Naïve Bayes learning
    - $\Pr(\text{key}|\text{TFIDF}, \text{pos}) = \frac{\Pr(\text{TFIDF}|\text{key}) * \Pr(\text{pos}|\text{key}) * \Pr(\text{key})}{\Pr(\text{TFIDF}, \text{pos})}$
  - Extracting keyphrases: output a ranked list of phrases based on  $\Pr(\text{key}|\text{TFIDF}, \text{pos})$
- 
-

# *Evaluation of Kea*

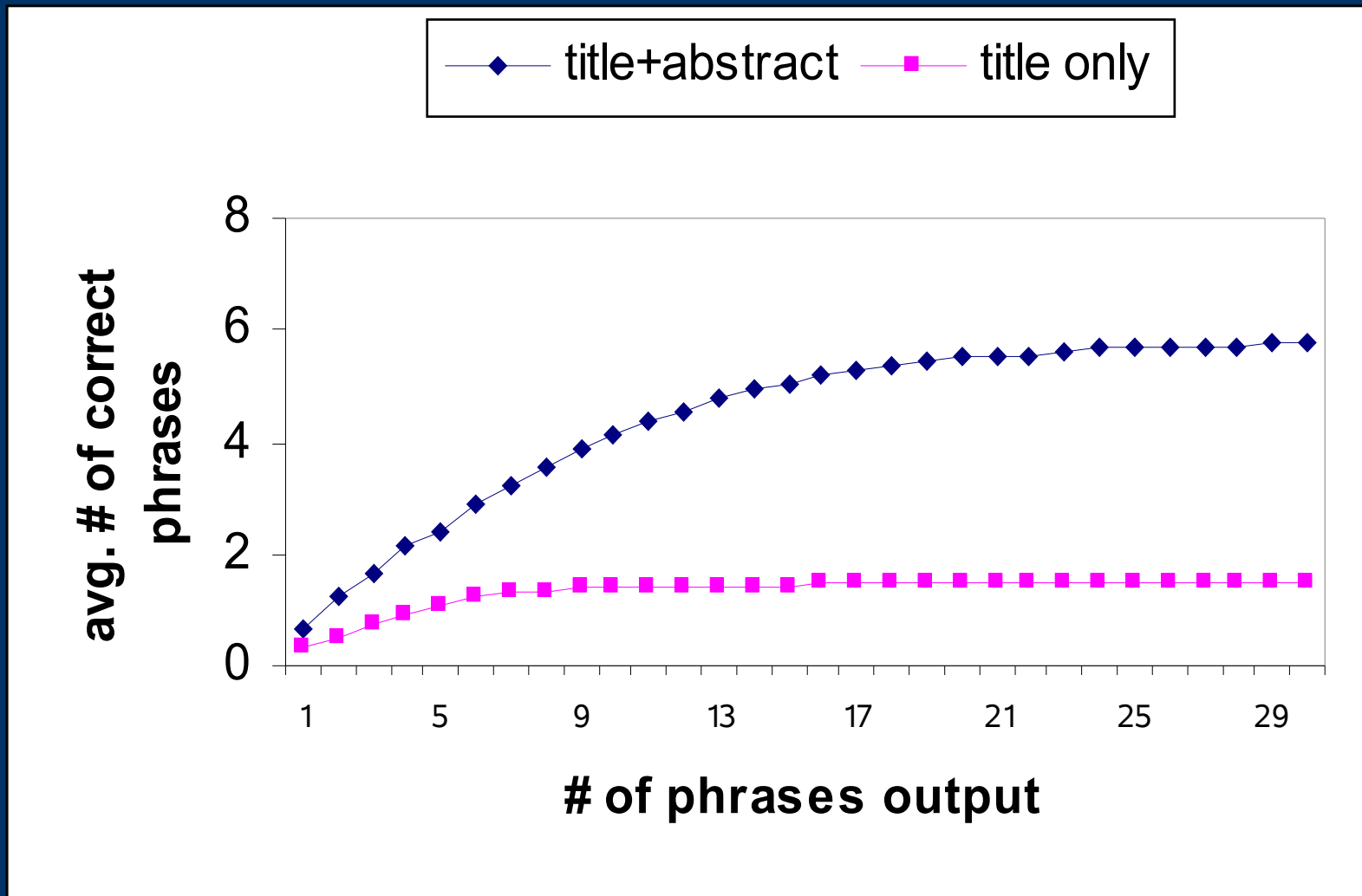
- Methodology
  - Total 1916 documents
  - 500 documents for testing
  - Training documents are randomly selected from remaining 1416 documents
- Measure
  - The number of correct phrases extracted by Kea against the human annotated phrases



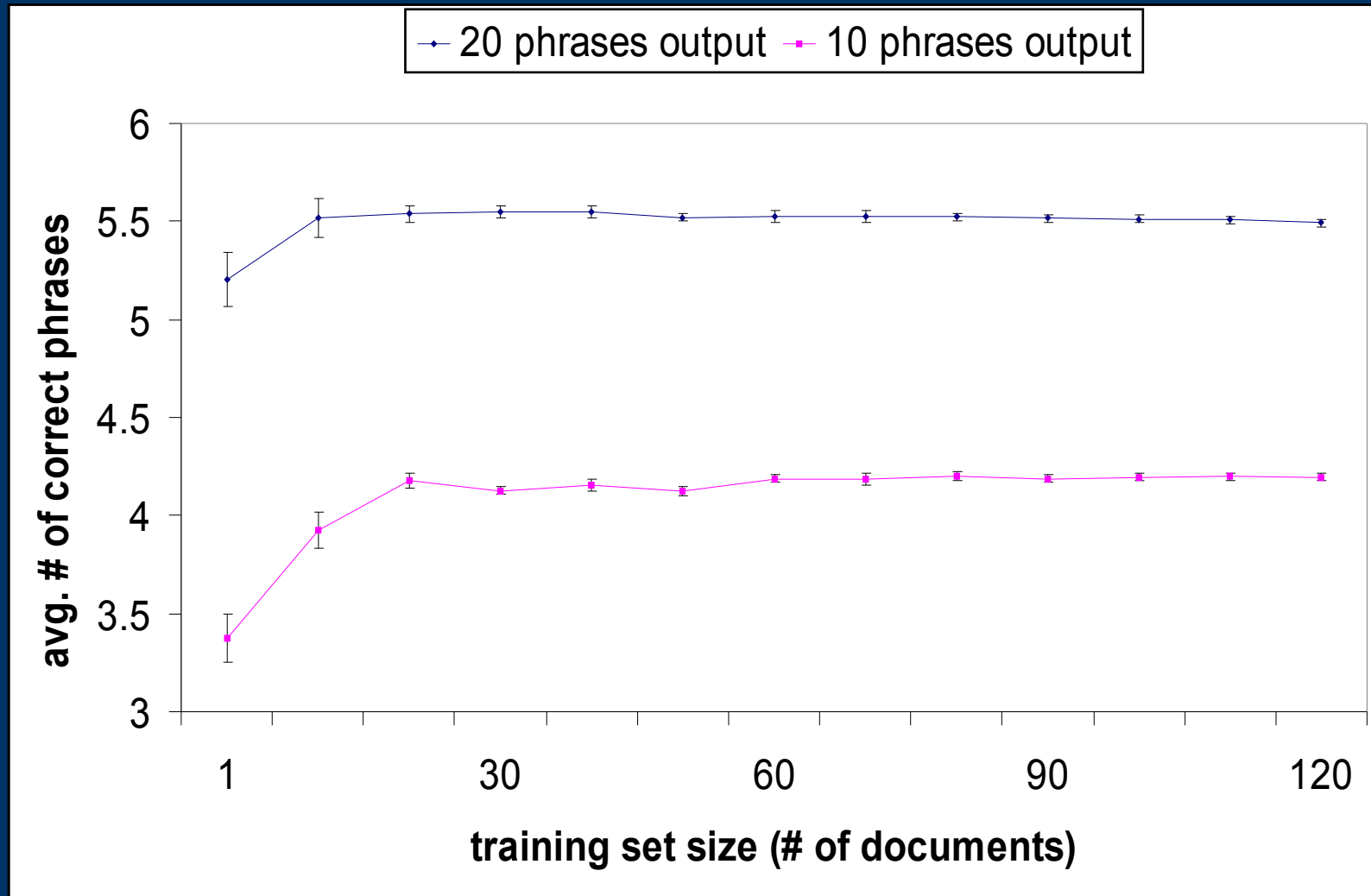
# Evaluation of Kea—Overall Performance

Keyphrases Extracted	Avg. Matches with Annotated Phrases from Full Text	Avg. Matches with Annotated Phrases from Title only
5	$2.54 \pm 1.16$	$1.09 \pm 0.89$
10	$4.13 \pm 1.74$	$1.41 \pm 1.07$
15	$5.06 \pm 2.22$	$1.45 \pm 1.11$
20	$5.521 \pm 2.6$	$1.46 \pm 1.12$
25	$5.68 \pm 2.8$	$1.46 \pm 1.12$
30	$5.74 \pm 2.89$	$1.46 \pm 1.12$

# Evaluation of Kea—Overall Performance (cont'd)



# Evaluation of Kea—Effect of Training Set Size

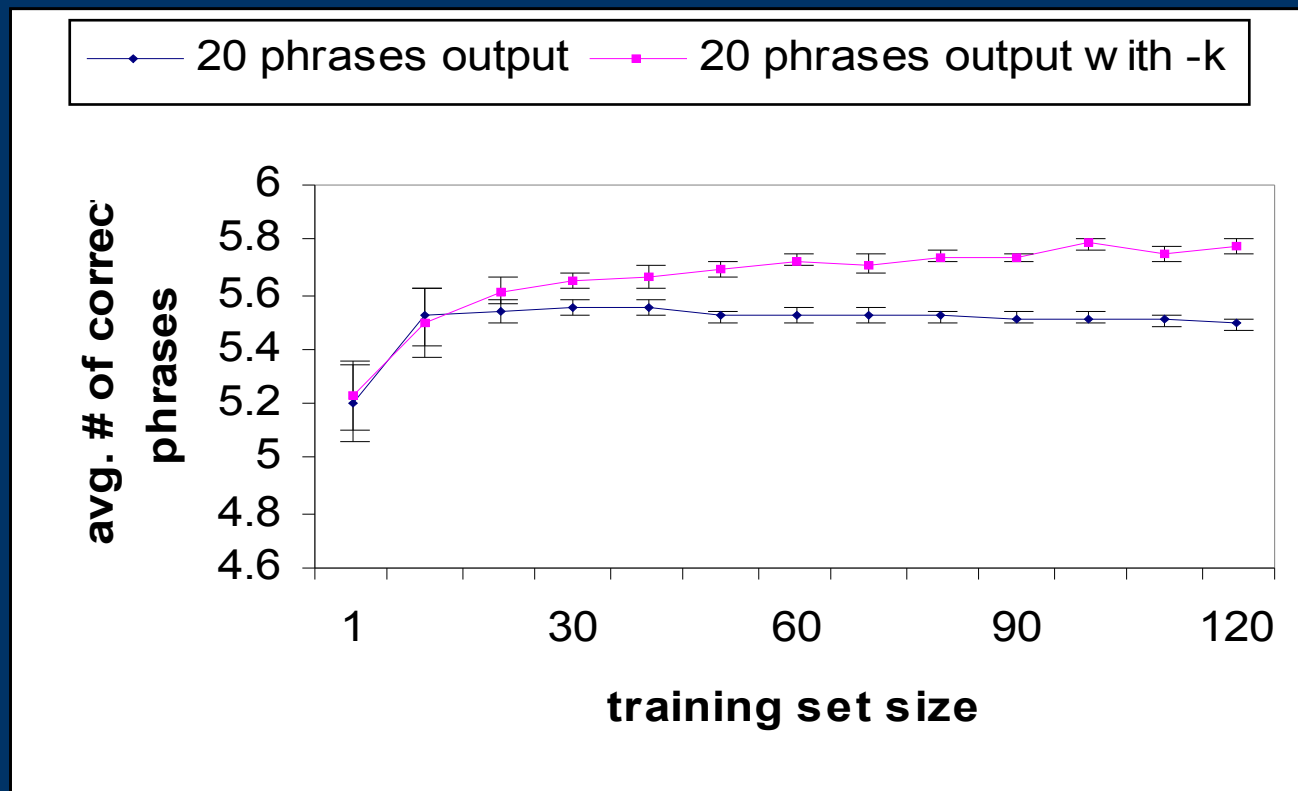


# *Evaluation of Kea—Effect of Training Set Size (cont'd)*

- 30 training documents are sufficient to push Kea's performance to the limit. (It seems odd that training set size is so small!)
  - Remark: The instances for training and testing are the candidate phrases. Therefore, 30 training documents have on average around 6536 candidate phrases for training.
  - A document is still an important entity in the study, since it contains the candidate phrases and provides the feature values of each candidate phrase.
- 
-

# Evaluation of Kea—Effect of Using Keyphrase Frequency Statistic

- Keyphrase frequency  $K$ —the number of times a phrase occurs as human-annotated keyphrase in all training documents
- $\Pr(\text{key}|\text{TFIDF}; \text{pos}; K) = \frac{\Pr(\text{TFIDF}|\text{key}) * \Pr(\text{pos}|\text{key}) * \Pr(K|\text{key}) * \Pr(\text{key})}{\Pr(\text{TFIDF}; \text{pos}; K)}$



# Discussion

- The results show that Kea can be used for automatic named entity tagging.
  - The observation shows that the low quality of automatically generated candidate phrases (the instances for training and testing) might affect Kea's performance significantly. A better information representation should be used in the future study.
  - It is also possible that Kea's performance might be affected by highly imbalanced data (a very small proportion of positive instances—keyphrases).
- 
-