Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5896–5904
Marseille, 11–16 May 2020

Automatic Compilation of Resources for Academic Writing and Evaluating with Informal Word Identification and Paraphrasing System

Seid Muhie Yimam[1], Gopalakrishnan Venkatesh[2], John Sie Yuen Lee[3] and Chris Biemann[1]
Universität Hamburg, Germany[1], International Institute of Information Technology, Bangalore, India[2],
City University of Hong Kong, Hong Kong SAR[3]
yimam@informatik.uni-hamburg.de

Abstract

We present the first approach to automatically building resources for academic writing. The aim is to build a writing aid system that automatically edits a text so that it better adheres to the academic style of writing. On top of existing academic resources, such as the Corpus of Contemporary American English (COCA) academic Word List, the New Academic Word List, and the Academic Collocation List, we also explore how to dynamically build such resources that would be used to automatically identify informal or non-academic words or phrases. The resources are compiled using different generic approaches that can be extended for different domains and languages. We describe the evaluation of resources with a system implementation. The system consists of an informal word identification (IWI), academic candidate paraphrase generation, and paraphrase ranking components. To generate candidates and rank them in context, we have used the PPDB and WordNet paraphrase resources. We use the *Concepts in Context* (CoInCO) "All-Words" lexical substitution dataset both for the informal word identification and paraphrase generation experiments. Our informal word identification component achieves an F-1 score of 82%, significantly outperforming a stratified classifier baseline. The main contribution of this work is a *domain-independent* methodology to build targeted resources for writing aids.

Keywords: academic writing, academic word, academic phrase, informal word identification, academic text paraphrasing

1 Introduction

We present the first approach to building resources for an academic writing aid system automatically. Academic writing aid systems help in automatically editing a text so that it better adheres to the academic style of writing, particularly by choosing a better academic word in a given domain. In the context of academic paraphrasing tasks, the resources are mainly words or phrases, that are more appropriate to use in an academic writing style. Moreover, the academic resources might vary from domain to domain as some words or phrases are extensively used in one domain over the other.

The first step in building an academic writing aid tool is to collect resources that determines whether a given phrase follows the style of writing in academia. This involves analyzing a given sentence and determining if the lexemes of the sentences are well-selected academic words and phrases or not.

To evaluate the resources compiled, we have to build a system, analogous to the lexical substitution and text simplification tasks, for example, (Szarvas et al., 2013; Štajner and Saggion, 2018), that consists of informal word identification, academic candidate generation, and candidate paraphrase ranking components (see Figure 1). While it is possible to follow the same approaches as the lexical substitution and text simplification approaches for academic text rewriting tasks, the main challenge for the academic paraphrasing task is the collection of resources for academic texts.

The following are the main objectives of building academic resources:

1. Identify suitable academic and non-academic datasets that are to be used to build academic resources.
2. Design a generic, *domain-independent*, approach to extract academic resources.
3. Evaluate the quality of the collected resources and use these resources for informal word identification (IWI) and academic paraphrasing systems.

The informal word identification (IWI) component automatically identifies informal words (see Section 4.2) that are going to be replaced with academic paraphrases. The candidate generation and ranking component determine the best academic candidate paraphrase to replace the informal words.

The ultimate goal of this research work is to integrate the informal word identification, candidate generation, and paraphrase ranking components into writing aid tools, for example to word processors or text composing software like latex packages, to automatically assist users in academic text composing.

In this work, we have targeted the following research questions 1) **How to build academic resources (words or phrases), which are used to replace informal or less academic expressions in academic texts?** 2) **How to build a system that can be used to evaluate the collected resources?**

In Section 2, a brief review of related works is presented. In Section 3, we discuss how to build academic resources using reference corpora and evaluate the quality of the resource. In Section 4, we present the approaches that are used to build an informal word identification and paraphrasing system for academic rewriting. Setups of the academic paraphrasing systems and the experimental results are discussed in Section 5. Analysis of system results and conclusion of the research are presented in Section 6 and Section 7 respectively.

2 Previous Work

In this section, we review previous work in lexical substitution, a closely related task, and discuss how the academic text rewriting system potentially differs.

In essence, our system is similar to lexical substitution (LS) and text simplification tasks, in such a way that both focus on the rewriting of an original text towards a given goal. Lexical substitution system mainly focuses on rewriting texts by replacing some of the words or phrases without altering the original meaning (Szarvas et al., 2013; Štajner and Saggion, 2018). The work by Guo et al. (2018) targeted text simplification based on the sequence-to-sequence deep neural network model, where its entailment and paraphrasing capabilities are improved via multi-task learning.

While the complex word identification (CWI) task focuses on identifying lexical units that pose difficulties to understand the sentence (Yimam et al., 2017b; Yimam et al., 2017a; Yimam et al., 2018; Paetzold and Specia, 2016), our informal word identification (IWI) component focuses on identifying words that are not fitting or adhering to the academic style of writing.

The work by Riedl et al. (2014) focuses on the lexical substitution task, particularly for medical documents. They have relied on Distributional Thesaurus (DT), computed on medical texts to generate synonyms for target words.

Existing resources for academic writing are limited to a precompiled list of words such as the Corpus Of Contemporary American English (COCA) (Gardner and Davies, 2013) and the New Academic Word List 1.0 (NAWL) (Browne et al., 2013) vocabulary lists. Regarding phrases (multiword expressions) for academic writing, the only available resources are the academic bi-grams compiled by Pearson[1]. However, these resources are 1) limited to a certain domain and target writers (mostly L2 learners and students), 2) their vocabulary is fixed, thus requiring manual work for an extension, and 3) the resources are limited to uni-gram and bi-gram lists. In this work, we build academic resources that are more generic, which can be built from existing reference corpora. In addition to uni-gram and bi-gram resources, we also design a system that can produce resources up to a length of four words (quad-grams).

As far as we know, the only system available to academic writing is the work of Lee et al. (2018), which addresses a different aspect, which is a sentence restructuring based on nominalizing verbal expressions.

3 Building Academic Resources

In this section, we will first discuss the existing academic resources, how they are built and their limitations. Then, we will present our approach that describes the process of building academic resources from different reference corpora. Finally, we will discuss the quality of the collected resources against two evaluation measures, namely comparing with the existing resources and manually evaluating the academic fitness of resources.

[1]Academic collocation list: https://pearsonpte.com/organizations/resea

3.1 Existing Resources for Academic Writing

In this subsection, we will present the existing academic word lists and phrases, which will be used to evaluate the quality of the dataset we build from reference corpora.

3.1.1 Academic Vocabulary

There are some efforts in building a list of vocabularies or words for academic writing. Some of them are created by analyzing text from academic writing corpora such as journals, theses works, and essays. One such resource is the Corpus Of Contemporary American English (COCA) (Gardner and Davies, 2013) vocabulary list, which contains about 3,000 words (in lemmas) that are derived from a 120 million word sub-corpus of the 560 million words. Similarly, the New Academic Word List 1.0 (NAWL) (Browne et al., 2013) was also built in the same way as the COCA list as a reference resource for second language learners of English, which is selected from an academic corpus of 288 million words.

3.1.2 Academic Phrases

Academic phrases are a list of collocated words (multiword expressions), which are mostly used for academic writing. The list from Ackermann and Chen (2013) comprises of 2,468 bi-gram collocations. The list is compiled from the written curricular component of the Pearson International Corpus of Academic English (PICAE) comprising of over 25 million words. However, the academic phrases, like the academic word lists, are mostly used as a guideline (study material) to practice academic writing.

3.2 Academic and Non-Academic Reference Corpora

The existing resources that are presented in Section 3.1 are prepared mostly as references or study guidelines for academic writers. However, to build automatic writing support, it is required to have more comprehensive and larger resources that can also be updated dynamically. In addition to single word and bi-gram lists, it would be also beneficial if the resource includes longer sequences of words. Hence, we have further extended the academic phrase list that includes up to four-gram phrases. The resource helps the academic paraphrasing or rewriting system in 1) identifying words or phrases in a text that are less academic and 2) providing alternative academic words or phrases that are more relevant to the contexts presented.

To this end, we have compiled a list of academic phrases that are extracted from the ACL Anthology Reference Corpus (ACLAC) (Bird et al., 2008). This corpus contains 22,878 scholarly publications (articles) about Computational Linguistics. To understand the syntactic difference of an academic corpus from a non-academic corpus, we have used the Amazon Review Full Score Dataset (Zhang et al., 2015) as our non-academic reference. The non-academic dataset is constructed by randomly taking 600,000 training samples and 130,000 testing samples for each review score from 1 to 5 (Zhang et al., 2015). In this paper, a review refers to the review text from the training sample.

The above two corpora can be considered to be a good fit as it shows a high match with the existing academic vocabulary or phrase list, as shown in Table 1. From Table 1,

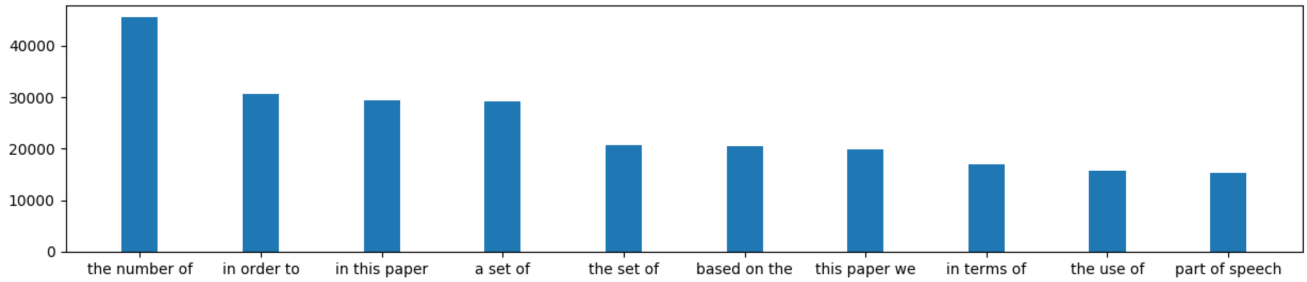

Figure 1: Frequencies of the highest occurring tri-grams collected from the reference corpora based on our approach.

Resource	Size	Coverage (%)
COCA	3,015	95.39
NAWL	963	**99.90**
Academic phrases	2,468	79.34

Table 1: Coverage of the existing resources for academic writing in our reference ACLAC corpus.

we can see that 95% of the academic words from COCA and 99.90% of the academic words from NAWL are represented in the ACLAC corpus. Similarly, around 80% of the bi-grams from the academic phrases (PICAE) are contained in the ACLAC corpus.

3.3 Approach to Build the New Academic Resource

On analyzing the corpora, we noticed that the non-academic corpus is much larger (in terms of the number of words) than the academic corpus. Therefore, we downsampled the non-academic text (to have comparable resources in terms of size) and ensured that the total number of words in both of the corpora are comparable. As a part of the pre-processing step, we clean the corpus (removing special characters) and lower case each word. We have considered a total of 991,798 reviews, which results in 75,184,498 tokens.

Using the NLTK's[2] Bi-, Tri- and Quad-Gram multi-word expression finder, we have extracted phrases from the two corpora (ACLAC and Amazon Review Full Score Dataset) and also compute the frequency distributions of these phrases across both the corpora as it can bee seen in Figure 1. The phrases extracted from both corpora can be used to assess naively the distribution across the two domains.

However, we have followed two different widely adopted approaches to extract representative phrases in a corpus, which is specifically known as keyphrases. The first approach is called Term Frequency-Inverse Document Frequency (TF-IDF), which is one of the most important statistics that show the relative importance of a term in a document in comparison to the corpus. The importance increases proportionally to the number of times a word appears in the document while its weight is lowered when the term occurs in many documents. We used the scikit-learn[3] implementation of TF-IDF to compute the scores of the different n-grams and thereby select the phrases that have maximum TF-IDF scores as keyphrases. In the ACLAC corpus, we have considered an article as one document while for the Amazon Review dataset, a review is considered as a single document.

In the second approach, we explore keyphrase extraction techniques based on part-of-speech sequences. We have employed *EmbedRank*, an unsupervised keyphrase extraction tool trained with sentence embeddings (Bennani-Smires et al., 2018). We consider only those phrases that consist of zero or more adjectives followed by one or multiple nouns (Wan and Xiao, 2008). While using the official implementation[4], we also explored the possibility of using the *Spacy*[5] POS tagger for keyphrase extraction in our corpora, which has a permissive license to redistribute our resource generation system as an open-source project.

As per the heuristic approach followed in the COCA word list compilation, we only retain those phrases that occur at least 50% more frequently in the academic portion of the corpora than would otherwise be expected. In other words, the ratio of the academic frequency of a term (in the ACLAC dataset) to the non-academic frequency (in the Amazon Review Full Score Dataset) should be 1.50 or higher (Gardner and Davies, 2013). Using a similar approach, we have also created the non-academic resources, which are also used to evaluate the quality of the academic resources in the human evaluation experiment (cf. Section 3.5)

3.4 Newly Collected Academic Resources

Based on the two keyphrase extraction approaches discussed in Section 3.3 (TF-IDF and EmbedRank based keyphrase extractions), we have compiled a total of 6,836 academic phrases (5,275 from EmbedRank and 1,900 from the TF-IDF approach). From Table 2, we can see that most of the academic keyphrases are extracted using the EmbedRank approach.

3.5 Manual Evaluation of Resources

From the automatically compiled list of resources (words and phrases), we have randomly sampled 520 words and phrases comprising of **155 uni-grams**, **100 bi-grams** and **5 tri-grams** from each of the compiled academic and non-academic phrase list. We then distributed the word and phrase lists to a total of 9 annotators (Ph.D. and postdoc-

[2] https://www.nltk.org/
[3] https://scikit-learn.org/
[4] https://github.com/swisscom/ai-research-keyphrase-extraction
[5] https://spacy.io/

Newly Collected resources				
Approach	Uni-gram	Bi-gram	Tri-gram	Quad-gram
EmbedRank	1,267	3,848	156	4
TF-IDF	1,090	690	109	11
From Existing Resources				
COCA	3,016	0	0	0
NAWAL	960	0	0	0
PICAE	0	2,468	0	0

Table 2: Academic word and phrases lists from the existing as well as from newly collected resources.

toral researchers) and requested the participants to label each entry as academic or non-academic. The sampled words and phrases are evaluated by two sets of annotators and the annotators were able to label the entries with an inter-annotator agreement of **68.22%**.

3.6 Results and Discussions on the Collected Resources

While analyzing the COCA list, we noticed that it contains a few stop words such as **both** and **above**. Hence, while relying on TF-IDF, we have considered extracting academic resources in different scenarios. First, we remove stop words as a part of the preprocessing step and in the second approach we have used the whole corpus as it is.
The system proposed by us relies on the relative frequencies in the reference corpora which can be computed independently of the language used. Thus the compilation of such an academic resource (through keyphrase extraction) can be considered language agnostic.
While performing the human evaluation, the annotators were asked to classify whether the given phrase is academic or not. The evaluation would have been more rigorous if they had to classify the phrases given the context in which the term had occurred. The annotators have at times labeled an entry as both academic and non-academic. Consider the word **attention**, it was used both in an academic (ACLAC) and non-academic (Amazon Review Full Score Dataset) context, for example as "LSTM with *attention*" and "the kid's *attention* to the game" respectively.

4 Evaluating the Resources for Academic Rewriting System

4.1 Academic Words

We define a word as **academic** or **formal** if it is in one of the following lists of academic phrases 1) keyphrases (up to four-grams) compiled by our system (cf. Section 3.3 – comprises of 6,836 phrases) 2) the COCA list (Davies, 2012) 3) the New Academic Word List (Browne et al., 2013)[6].
Some example academic words are shown in Table 3. The academic word lists are also extended to phrases or multi-word expressions. Pearson has published a set of academic bi-grams[7]. Words like *best*, *almost*, and *way* are not by themselves *academic*, but they can be combined with other words to form academic expressions such as *best described*, *almost identical*, and *appropriate way*.

[6]http://www.newgeneralservicelist.org/nawl-new-academic-word-list

[7]Academic collocation list: https://pearsonpte.com/organizations/resea

Academic words	report, state, claim...
Non-academic words	say, declare, mention, allege...

Table 3: Example of academic and non-academic words based on our academic resources.

4.2 Informal Words

The naive approach is to attempt to rewrite every non-academic word, using our definition above. That is a misplaced goal, however, since even the average document in the BAWE corpus (Alsop and Nesi, 2009) contains a considerable number of words outside the list, including function words and other words commonly used in all English documents.
We define a word as **informal** if it is a non-academic term that can be paraphrased by an academic term. If the term is academic, or it is non-academic but does not have an academic paraphrase, it is termed as **formal**.

4.3 Architecture

As shown in Figure 2, our proposed system consists of four components, which is analogous to the lexical simplification systems (Paetzold and Specia, 2017). The components of our system constituted informal word identification (IWI), paraphrase generation, candidate selection, and paraphrase ranking.

4.3.1 Informal Word Identification

The informal word identification (IWI) component identifies each word as *informal*, or not. The system attempts to paraphrase only the informal words in the rest of the pipeline.
Similar to CWI (Yimam et al., 2017b; Yimam et al., 2017a; Yimam et al., 2018; Paetzold and Specia, 2016), IWI is more accurate when placed in context. The word *big*, for example, may need to be paraphrased to *major* in the context of "*This article makes two big contributions.*" It should not be paraphrased, however, when it is part of the expression *big data*.

4.3.2 Paraphrase Generation, Selection, and Ranking

Given an informal word, this step generates a list of substitution candidates. While there are different approaches to generate candidates for target words, such as using existing paraphrase resources like WordNet and Distributional thesaurus (see Yimam et al. (2016)), we depend solely on the CoInCo (Kremer et al., 2014), WordNet (Miller, 1995), and the paraphrase database (PPDB) (Pavlick et al., 2015) resources to generate candidates.
Once the candidates are generated, all of the candidates, which must be academic words are retained for the paraphrase ranking component. Given a list of academic substitution candidates, the paraphrase ranking component finds the one that fits best in the context. The detailed approach is presented in Section 4.4.

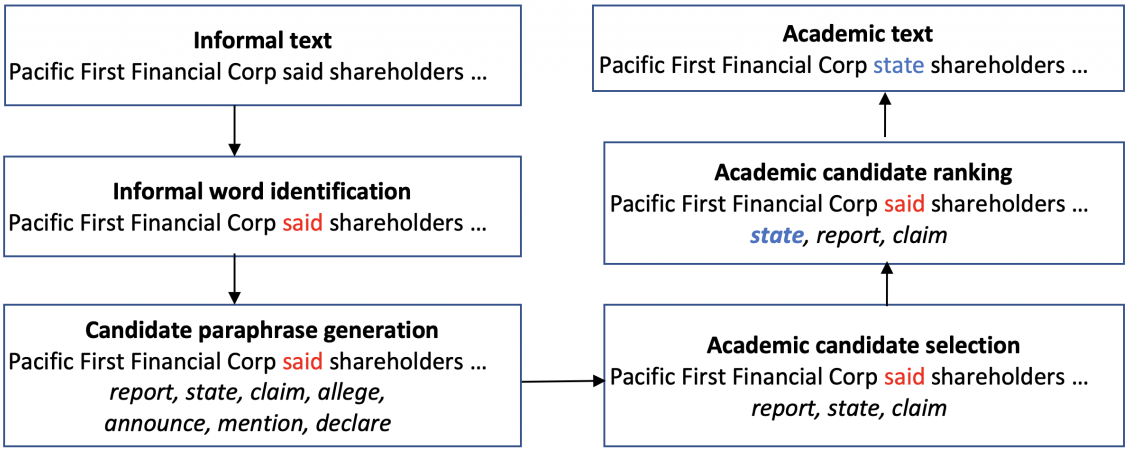

Figure 2: Architecture of the system.

CoInCo annotation	Pacific First Financial Corp said[paraphrases: report, state, detect] shareholders
IWI dataset	Pacific[N] First[N] Financial[N] Corp[N] said[Y] shareholders

Table 4: Transformation of the CoInCo dataset into IWI dataset, with respect to the academic word list in Table 3

4.4 Datasets for IWI and the Paraphrasing Components

For this evaluation, we derive our dataset from a lexical substitution dataset called the Concepts in Context (CoInCo) (Kremer et al., 2014). The CoInCo dataset is an *All-Words* lexical substitution dataset, where all words that could be substituted are manually annotated. The corpus is sampled from newswire and fiction genres of the Manually Annotated Sub-Corpus (MASC) corpus[8]. While the targets (words that are going to be substituted) are used to build the informal word identification dataset, the candidates are further processed to perform the academic paraphrase ranking task.

A total of 1,608 training and 866 test sentences are compiled out of 2474 sentences from the CoInCo dataset. Statistics on the IWI dataset are shown in Table 5.

4.4.1 Building the IWI Dataset

We automatically generated an IWI dataset from CoInCo as follows. For each non-academic target word, we determine if its substitution candidates include at least one academic word. If so, it is labeled as **informal**; otherwise, it is labeled as **formal**. All academic target words and all words without substitution candidates are labeled as **formal**. An example is given in Example 4.1 and Table 3.

> **Example 4.1.**
> **Sentence**: Pacific First Financial Corp said shareholders ...
> **CoInCo annotation**: Target word: said. Paraphrases: report, state, claim, **allege, announce, mention, declare**
> **IWI dataset** ([I]–informal, [F]–formal): Pacific[F] First[F] Financial[F] Corp[F] said[I] shareholders[N]

[8] http://www.anc.org/data/masc/

Dataset	# Tokens		#Types	
	I	F	I	F
IWI training	6,783	3,358	2,266	1,509
IWI test	3,666	1,822	1,577	994

Table 5: Statistics on the IWI dataset. *#Tokens* shows the total number of tokens (formal (*F*) and informal (*I*)) while *#Types* shows the unique occurrences of tokens in the IWI training and test sets. *I* stands for informal and *F* for formal tokens and types resp.

4.4.2 Paraphrase Candidates

To generate **non-academic** to **academic** word pairs for paraphrasing, we used the paraphrases (word pairs) in CoInCo, WordNet, and PDPB as the starting point.

For the CoInCo dataset, we have only included those word pairs where: 1) the target word is non-academic, 2) the substitution candidate is academic, 3) the target word has a higher word frequency than the substitute candidate in our academic resources. Since the academic resource is not exhaustive, some proper academic terms may be mistakenly considered as **non-academic**. This requirement aims to prevent these words from being substituted.

For example, from the sentence in Example 4.1, we obtained the word pairs say:report, say:state, and say:claim. We have collected a total of 23,476 word pairs from the CoInCo Training Set.

The dataset is prepared with 4 candidates for each informal target, where 2 candidates are academic and 2 candidates are non-academic. When we do not have appropriate candidates we extract further candidates from WordNet (Miller, 1995) and PPDB (Pavlick et al., 2015). Table 6 shows the statistics of target words extracted from the CoInCo dataset, where 59% of the informal words have possible candidate paraphrases.

4.5 Academic Paraphrase Corpus

In general, any existing paraphrase or lexical substitution corpus can be converted into an academic paraphrase corpus with the following steps:

1) Discard all academic target words since they do not need

# target words		Paraphrase coverage
Original	Our corpus	in (%)
5,480	3,250	59.30

Table 6: Statistics on our evaluation dataset. The last column shows the percentage of non-academic words in the corpus for which paraphrases can be obtained.

to be paraphrased.
2) Remove all non-academic substitution candidates for the remaining (non-academic) target words.
If no candidate is left after step (2), also remove that target word.

4.6 Informal Word Identification Models

We trained three Support Vector Machine (SVM) classifiers, using Radial Basis Function kernel, from scikit-learn[9] with different feature sets. We use the following features:
Word frequency: We use word frequencies 1) in the Beautiful Data[10] which are derived from the Google Web Trillion Word Corpus, 2) in the general COCA list, and 3) in the ACL anthology corpus (Bird et al., 2008).
Word Embedding: We have used GloVe (Pennington et al., 2014) word embedding to compute the cosine similarity between the word and the sentence[11]. We also explore the option of using Euclidean distance between the word and the sentence as a feature while training the classifier.
Part of Speech Tag (POS): The POS tag of the word obtained from the TreeTagger[12].
Word level features: We use the word length and the number of vowels as features for training the classifier.

4.7 Paraphrase Ranking Models

In order to rank the best candidates for academic rewriting, we have followed the learning-to-rank machine learning approach, where candidates are ranked based on their relevance score. The number of annotators selected the given candidate is considered as a relevance score. The TF-Ranking deep learning model provided by *TensorFlow Ranking*[13] library (Pasumarthi et al., 2019) is used to build the paraphrase ranking model.

5 Experiments

5.1 Informal Word Identification

We trained the IWI classifier on the ColnCo Train Set using SVM. Similar to most of the CWI evaluation metrics, we evaluate the performance of the system on the following evaluation metrics:
Precision: The number of correct informal targets, out of all targets proposed by the system.
Recall: The number of correct informal targets, out of all informal words that should be paraphrased.
F-Measure: The harmonic average of precision and recall.
Table 7 shows IWI precision and recalls on the ColnCo Test Set. We use a simple stratified randomization algorithm from scikit-learn as a baseline system. The proposed algorithm (SVM classifier) achieves a better performance overall in the F-Score of 0.8204. As it can be seen in Table 7, the following features work better for the IWI task: frequencies, cosine similarity, and Euclidean distance.

Method	Precision	Recall	F-score
Baseline	0.6679	0.6787	0.6733
SVM *Fe1*	0.7584	**0.8933**	**0.8204**
SVM *Fe2*	**0.7650**	0.8748	0.8162
SVM *Fe3*	0.7552	0.8912	0.8176

Table 7: Precision and recall on the informal word identification task. The baseline system has been setup using the Stratified classifier from scikit-learn: The stratified classifier in scikit-learn generates predictions by respecting the training set's class distribution. Fe1 = (Frequencies, cosine similarity), Fe2 = Fe1 + (Euclidean distance), Fe3 = All features

Parameters		Ranking metric
Loss	Steps	MRR
Logistic	50	0.8861
	100	**0.8926**
	200	0.8895
Softmax	50	0.8893
	100	0.8895
	200	**0.8914**

Table 8: Academic paraphrasing performance on the CoInCo Test Set using the MRR ranking metric.

5.2 Academic Paraphrasing

We evaluate the system performance on automatically generating academic paraphrases and ranking them. Following standard evaluation metrics in lexical simplification, we report on the **Mean Reciprocal Rank (MRR)**[14] metric.
The model from TF-Ranking (Pasumarthi et al., 2019) library has been trained to re-rank the candidates on the CoInCo test set. The model was trained using the *Adagrad* optimizer with a learning rate of 0.05. Experiments were performed on various loss functions (*pairwise_logistic_loss* and *softmax_loss*) and different *step*[15] (50, 100 and 200) values. Table 8 shows the experimental results.

6 Analysis of Results

For the informal word identification task, our models have a slightly lower precision as our dataset is not balanced (we have more informal words than formal words, as shown in Table 5).
From an error analysis, we find out that even if the term is academic in general, its usage in the test dataset is in-

[9] https://scikit-learn.org/
[10] https://norvig.com/ngrams/
[11] Embedding for the sentence is calculated by averaging the embedding of words in the sentence
[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[13] https://github.com/tensorflow/ranking
[14] https://en.wikipedia.org/wiki/Mean_reciprocal_rank
[15] Steps are the number of training iterations executed.

clined to be informal. For example, in the sentence "*It was last February, after the winter break, that we moved in together.*", *break* is labeled as academic but should be labeled as informal. This issue could be solved by further enhancing the dataset by employing human annotators during the resource compilation process.

Similarly, some of the errors from the system's prediction are to be attributed to the annotation process of the test set. For example, in the sentence "*They included support for marine reserves and money for fisheries management reform.*", *reserves* is annotated as informal while the system identified it as formal.

In general, while bootstrapping the academic resource compilation and the informal word identification tasks, a minimal intervention of human annotators would enhance the overall system. Furthermore, integration of a BERT or other contextualized embedding model (Devlin et al., 2019) could also help to improve the performance of the system. Contextualized word embeddings provide word vector representations based on their context. As the vector representation of words varies as per the context, they implicitly provide a model for word sense disambiguation (WSD).

7 Conclusion and Future Direction

In the realm of academic text writing, we explored how to compile academic resources, automatically identify informal words (words that are less formal for academic writing), and provide better substitutes. We have used a generic approach to compile the academic resources, which can be easily transferred to domains or languages as it only requires text corpus. The academic text rewriting system, analogous to lexical substitution systems, consists of informal word identification, candidate generation, candidate selection, and ranking components. As far as we know, this is the first experiment towards the development of academic writing support for academia, while there might be commercial cases (for example Grammarly[16]) that we do not know how the systems operate.

We envision this system to be embedded into open source academic writing aid tools where the academic sources are used to detect informal terms and propose academic substitutes. For the resource compilation process, it would be nice to extend the EmbedRank approach to extract keyphrases beyond the adjective and noun POS tag patterns, especially to cover verbs used in academic contexts.

Source code and resources of this the paper are released publicly[17] on the Github repository under permissive licenses (ASL 2.0, CC-BY).

Acknowledgments

This work was partially funded by a HKSAR UGC Teaching & Learning Grant (Meeting the Challenge of Teaching and Learning Language in the University: Enhancing Linguistic Competence and Performance in English and Chinese) in the 2016-19 Triennium.

[16] https://www.grammarly.com/

[17] https://github.com/uhh-lt/par4Acad

8 Bibliographical References

Ackermann, K. and Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12:235–247.

Alsop, S. and Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4(1):71–83.

Axelsson, M. W. (2000). USE - The Uppsala Student English Corpus: An instrument for needs analysis. *"International Computer Archive of Modern and Medieval English"*, (24):155 – 157.

Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., and Jaggi, M. (2018). Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229, Brussels, Belgium.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *International conference on Language Resources and Evaluation (LREC 2008)*, pages 1755–1759, Marrakech, Morocco.

Browne, C., Culligan, B., and Phillips, J. (2013). New academic word list 1.0. Accessed December 2019:http://www.newgeneralservicelist.org/nawl-new-academic-word-list.

Cohn, T., Callison-Burch, C., and Lapata, M. (2008). Constructing Corpora for the Development and Evaluation of Paraphrase Systems. *Computational Linguistics*, 34(4):597–614.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4):397 – 423.

Coxhead, A. (2019). An introduction to the academic word list. Accessed December 2019:http://ksngo.org/images/download/LDOCE_AWL.pdf.

Davies, M. and Gardner, D. (2013). A New Academic Vocabulary List. *Applied Linguistics*, 35(3):305–327.

Davies, M. (2012). Corpus of Contemporary American English (1990-2012). Accessed December 2019:http://corpus.byu.edu/coca/.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dong, L., Mallinson, J., Reddy, S., and Lapata, M. (2017). Learning to Paraphrase for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark.

García Salido, M., Garcia, M., Villayandre-Llamazares, M., and Alonso-Ramos, M. (2018). A Lexical Tool for Academic Writing in Spanish based on Expert and

Novice Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan.

Gardner, D. and Davies, M. (2013). A New Academic Vocabulary List. *Applied Linguistics*, 35(3):305–327.

Guo, H., Pasunuru, R., and Bansal, M. (2018). Dynamic Multi-Level Multi-Task Learning for Sentence Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Santa Fe, NM, USA.

Kasewa, S., Stenetorp, P., and Riedel, S. (2018). Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4977–4983, Brussels, Belgium.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What Substitutes Tell Us - Analysis of an "All-Words" Lexical Substitution Corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden.

Lee, J., Saberi, D., Lam, M., and Webster, J. (2018). Assisted nominalization for academic English writing. In *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, pages 26–30, Tilburg, the Netherlands.

McCarthy, D. and Navigli, R. (2007). Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

Michael, M. and O'Dell, F. (2008). *Academic Vocabulary in Use: 50 Units of Academic Vocabulary Reference and Practice ; Self-study and Classroom Use*. Cambridge University Press.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Morley, J. (2014). Academic phrasebank. Technical Report 2014b edition, The University of Manchester. Accessed December 2019: http://www.kfs.edu.eg/com/pdf/2082015294739.pdf.

Oshima, A. and Hogue, A. (2007). *Introduction to Academic Writing*. Pearson Education, Third Edition (The Longman Academic Writing Series, Level 3) (3e) edition.

Paetzold, G. and Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, CA, USA.

Paetzold, G. H. and Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60(1):549–593.

Pasumarthi, R. K., Bruch, S., Wang, X., Li, C., Bendersky, M., Najork, M., Pfeifer, J., Golbandi, N., Anil, R., and Wolf, S. (2019). TF-Ranking: Scalable TensorFlow Library for Learning-to-Rank. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2970–2978, Anchorage, AK, USA.

Pavlick, E. and Callison-Burch, C. (2016). Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany.

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., and Callison-Burch, C. (2015). PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Riedl, M., Glass, M., and Gliozzo, A. (2014). Lexical substitution for the medical domain. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 610–614, Doha, Qatar.

Ruppert, E., Kaufmann, M., Riedl, M., and Biemann, C. (2015). JOBIMVIZ: A Web-based Visualization for Graph-based Distributional Semantic Models. In *The Annual Meeting of the Association for Computational Linguistics (ACL) System Demonstrations*, pages 103–108, Beijing, China.

Sekizawa, Y., Kajiwara, T., and Komachi, M. (2017). Improving japanese-to-english neural machine translation by paraphrasing the target language. In *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, pages 64–69, Taipei, Taiwan.

Štajner, S. and Saggion, H. (2018). Data-driven text simplification. In *Proceedings of COLING 2018, the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 19–23, Santa Fe, NM, USA.

Szarvas, G., Biemann, C., and Gurevych, I. (2013). Supervised All-Words Lexical Substitution using Delexicalized Features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1131–1141, Atlanta, Georgia.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Edmonton, Canada.

Wan, X. and Xiao, J. (2008). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, pages 855–860, Chicago, IL, USA.

Yimam, S. M. and Biemann, C. (2018). Par4sim – adaptive paraphrasing for text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 331–342, Santa Fe, NM, USA.

Yimam, S. M., Martínez Alonso, H., Riedl, M., and Biemann, C. (2016). Learning Paraphrasing for Multiword Expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 1–10, Berlin, Germany.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017a). CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017b). Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822, Varna, Bulgaria.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, LA, USA.

Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Cambridge, MA, USA.