



New trap to catch copy-cats:

Using Natural Language Processing for Automatic Detection of Plagiarism

Miranda Chong*, Lucia Specia, Ruslan Mitkov

Research Group in Computational Linguistics, University of Wolverhampton, UK

* miranda.chong@wlv.ac.uk

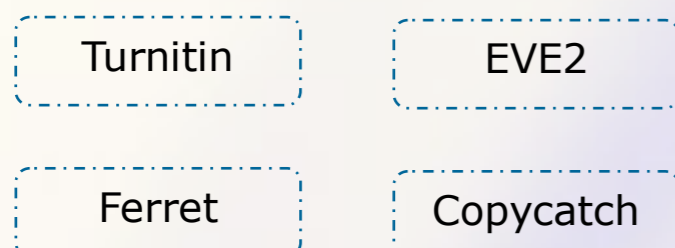
1. Background

Plagiarism and plagiarism detection have increasingly attracted attention amongst pedagogic and many other fields.

Plagiarism detection can be categorised as follows:

- Programming source-code detection
- Written text detection
 - Intrinsic (detect plagiarism within a document)
 - Extrinsic/ External (detect plagiarism between documents)

Examples of existing detection tools for written text (external):



External plagiarism detection

Limitations of detection tools and research approaches:

- String/character level matching
- No complex linguistic treatment
 - Limited syntax (intrinsic detection)
 - Very limited semantics at the word-level only
 - Tentative of synonymy, but challenging with sense ambiguity
- Very limited use of Machine Learning
- Limited cross-lingual detection

2. What is NLP?

The processing of human languages by machines.

Examples of current applications of NLP:

- Computer-assisted language learning
- Multi-lingual translation systems
- Extraction of biomedical information

NLP is still an under-explored area for plagiarism detection!

3. Challenges

- Lexical changes: synonymy, related concepts, etc
- Structural changes: active/passive voice, word order, etc
- Textual entailment: sentence paraphrase and other semantic variations
- Multi-source plagiarism
- Multi-lingual plagiarism

Aims of study:

Improve detection accuracy by:

1. Incorporating NLP techniques
2. Exploiting new comparison methodologies

4. Proposed framework

- 1) Analyse strings, grammar and meaning of text.
- 2) Account for text relations (synonymy, sentence paraphrasing).

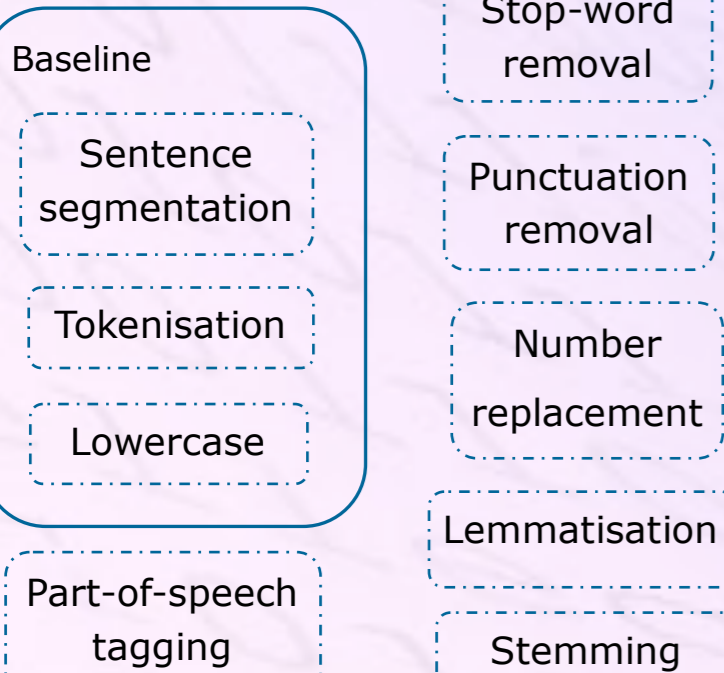
Experimental setup:

Corpus of plagiarised short answers (Clough & Stevenson, 2009):

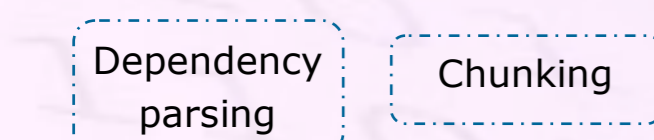
- Original documents (parts of Wikipedia articles)
- Non-plagiarised cases
- Plagiarised cases (with various levels of text overlaps)

Text pre-processing and

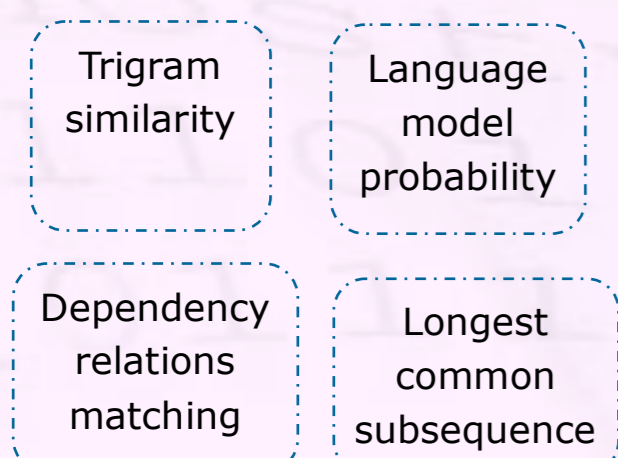
NLP techniques:



Syntactic processing techniques:



Comparison methodologies:

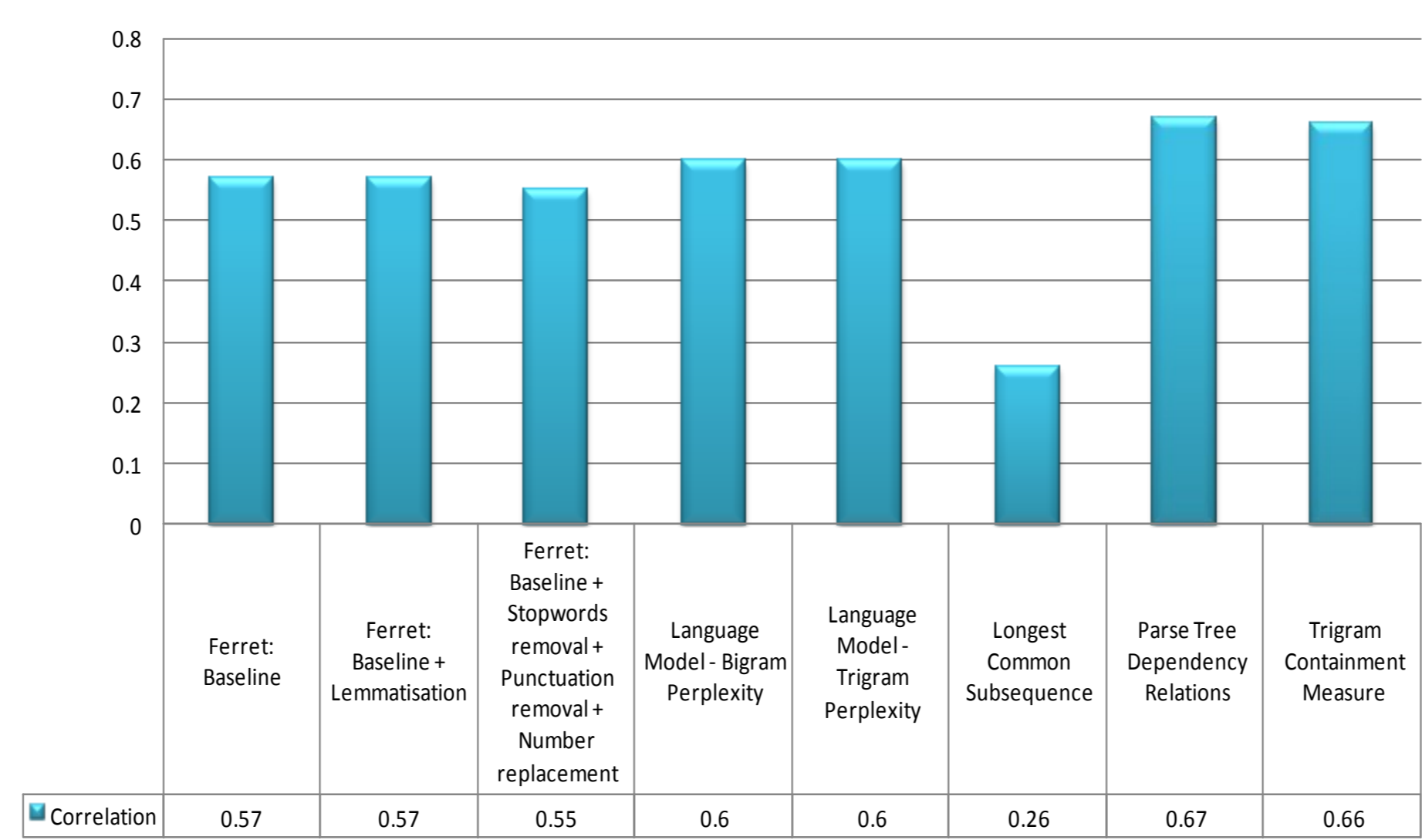


Machine learning algorithm:

Naive Bayes Classifier

5. Experimental results

Correlation coefficient scores of best features



Comparison on accuracy of plagiarism levels classification after using the above features scores to train machine learning algorithm:

Naive Bayes Classifier	Accuracy
Our approach	70.53 %
Trigram comparison approach	66.32 %
Combined approaches	60.00 %

6. Future work

- Investigate language independent detection
- Handle paraphrased word/structure
 - Incorporate thesaurus to determine lexical relations
 - Incorporate textual entailment to determine other relations

7. Social Impact

Detection for potential educational purposes:

- Detection tool acts as a pre-emptive tool
- Helps identify incorrectly referenced texts
- Offer guidance to students rather than punishment
- Automated tools help to identify potential plagiarised cases for further investigation

8. Summary

- Plagiarism detection methodologies can be improved using NLP
- These tools can identify possible plagiarised cases
- Human intervention will always be required to judge plagiarised cases

9. References

Clough, P., & Stevenson, M. (2009). Developing a corpus of plagiarised short answers. Language Resources and Evaluation, LRE 2010.