Machine Translation Evaluation Metric for Text Alignment

Prasha Shrestha, Suraj Maharjan, and Thamar Solorio

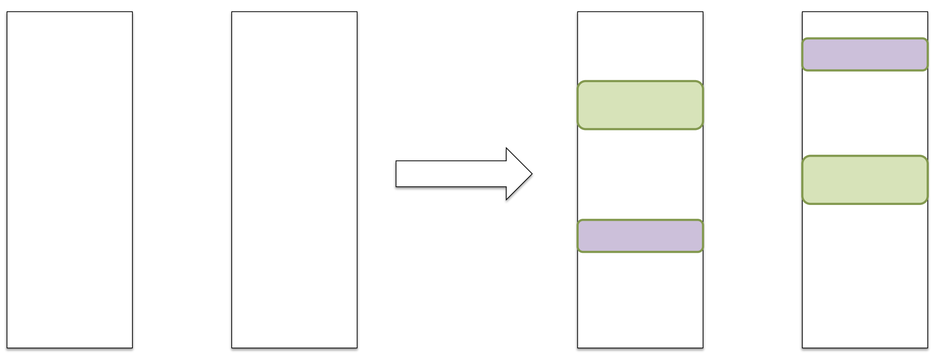
University of Alabama at Birmingham

Department of Computer and Information Sciences

Birmingham, Alabama 35294-1170, USA

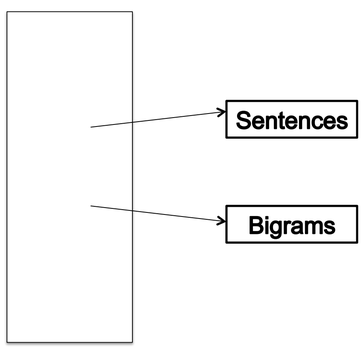
{prasha, suraj, solorio}@cis.uab.edu

Task



System Workflow

Seeding



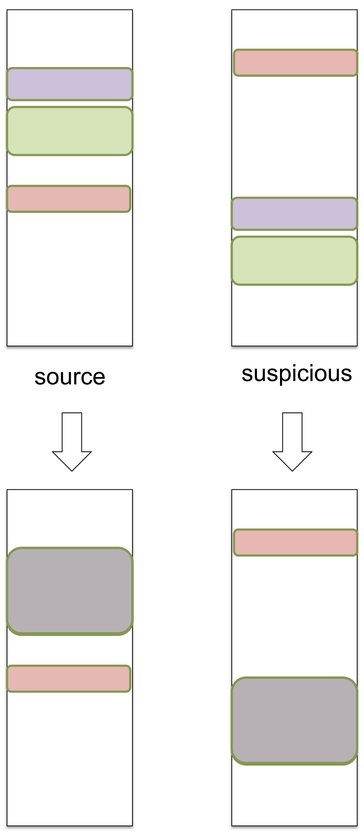
Matching

Sentence pairs with TER-p scores < 0.9

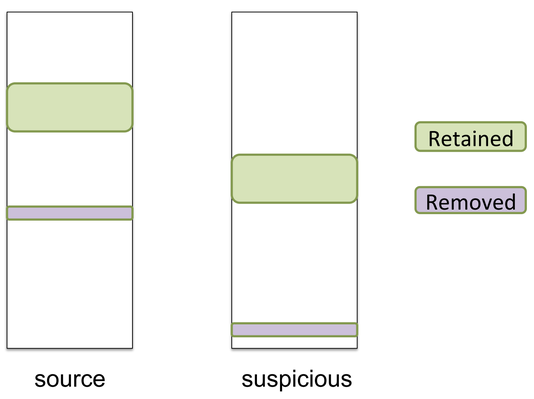
+

Exactly matched bigrams

Merging



Postprocessing



PAN'13 System Comparison

- Features used in our 2013 system: stopword n-grams, named entity n-grams and all-words n-grams
- Used *n* = 10 for stopwords and 8 for named entity n-grams and all words n-grams, now just bigrams
- Postprocessing: Similar in both systems; without postprocessing, the granularity is very high in both systems.

Results

Plagiarism Type	Plagdet	Precision	Recall	Granularity
No Obfuscation	0.88009	0.80287	0.97374	1.00000
Random Obfuscation	0.86150	0.89992	0.83358	1.00642
Translation Obfuscation	0.85427	0.85322	0.86086	1.00447
Summary Obfuscation	0.15853	0.98171	0.08975	1.05263

Table 1: Evaluation results for the training dataset

System	Plagdet	Precision	Recall	Granularity
Our PAN'14 System	0.84404	0.85906	0.83782	1.00701
PAN'14 Best System	0.87816	0.88168	0.87904	1.00344
Our PAN'13 System	0.69595	0.87461	0.73892	1.22072
PAN'13 Best System	0.82827	0.89564	0.77177	1.00140
Baseline	0.42191	0.92939	0.34223	1.27473

Table 2: Evaluation results on the 2013 test dataset

System	Plagdet	Precision	Recall	Granularity
Our PAN'14 System	0.86806	0.84418	0.89839	1.00381
PAN'14 Best System	0.90779	0.92757	0.88916	1.00027
Our PAN'13 System	0.78420	0.85634	0.85340	1.12892
PAN'13 Best System	0.84667	0.89179	0.80590	1.00000
Baseline	0.64740	0.90024	0.52838	1.04005

Table 3: Evaluation results on the supplemental test dataset

Conclusion

- Smaller *n* is better when trying to find matches between two texts as smaller n-gram matches can always be merged to get larger n-grams but the reverse is not possible.
- Matches from TER-p method give higher recall and lower precision whereas bigrams give higher precision and lower recall. Together, they balance each other out.
- Postprocessing decreases recall slightly, but is important in order to improve granularity.
- Machine translation metrics, summarization metrics and textual entailment metrics all measure either similarity or edit distant. So, they can all be used for plagiarism detection.

Partially supported by

NSF

ONR

CLEF'14,
15-18 September,
Sheffield - UK