Using Machine Translation Evaluation Techniques to Evaluate Text Simplification Systems

Sandeep Albert Mathias
Advisor: Prof. Pushpak Bhattacharyya

Department of Computer Science and Engineering, IIT Bombay

Evaluating Text Simplification Systems

- An automatic text simplification system is a system that takes text as input, simplifies it and returns the simplified text as output.
- How usable is the text simplification system that we get?
  - Grammaticality
  - Meaning Preservation
  - Simplicity
  - Overall Usability

LREC 2016 Shared Task Submission

- Evaluate text simplification systems based on:
  - How grammaticality correct is the output?
  - How simple is the output?
  - How much of the meaning of the input is preserved in the output?
  - How good is the overall usability of the system?
- The shared task was modeled as a set of classification problems.
  - Classes: good, ok, bad
  - Training data: 505 sentence pairs. 5 fold cross validation is used in evaluation.
  - Test data: 126 sentence pairs

Grammaticality

- Example:
  - Tell me vs. Say me
  - Tell <something> to me vs. Say <something> to me
- We use language modeling to solve this problem. We mainly tackle usage errors, as compared to syntactic errors.
- We make use of the Simple English Wikipedia dump to train a language model.
- Features used:
  - No. of words in the sentence
  - No. of OOVs (out of vocabulary words)
  - Language model score for the sentence
  - Perplexity of the sentence
  - Average perplexity per word, of the sentence
- We used the SRILM toolkit to train the language model,

Meaning Preservation

- Example:
  - Input: Warsaw lies on the Vistula River, about 240 miles southeast of the Baltic coast city of Gdansk.
  - Good output: **Warsaw** is on the **Vistula** River, about 240 miles southeast of Gdansk. Gdansk is a Baltic coast city.
  - Bad output: **Vistula** is on the **Warsaw** River, about 240 miles southeast of the Baltic coast city of Gdansk.
- We need a way to handle:
  - Exact matching of phrases
  - Matching of stems[1]
  - Matching of synonyms[2]
  - Matching of paraphrases[3]
- The answer – METEOR![3]

Results for Grammaticality, Meaning Preservation and Simplicity

Experiment	Accuracy (G)	MAE (G)	RMSE (G)	Accuracy (M)	MAE (M)	RMSE (M)	Accuracy (S)	MAE (S)	RMSE (S)
Training Set – Baseline	75.64	17.23	36.96	58.21	28.61	46.94	52.67	32.18	49.60
Training Set	76.04	16.63	36.01	66.34	19.50	35.25	48.31	32.87	48.59
Test Set – Baseline	76.19	18.25	22.43	57.94	28.97	35.30	55.56	29.37	31.22
Test Set	72.22	21.43	25.78	63.49	20.63	26.75	47.62	34.13	38.85

Simplicity

- Lexical complexity (Lc(S))
- Corpus complexity (Cc(g))[4] – Ratio of log likelihood of an n-gram (g) occurring in the English Wikipedia to the log likelihood of that n-gram occurring in the Simple English Wikipedia.
  - $Cc(g) = \frac{LL(g|normal)}{LL(g|simple)}$
- Syllable Count (Sc(g)) – Number of syllables in the n-gram (g).
- $Lc(S) = \sum_n W_n \sum_g Sc(g) * Cc(g)$,
- where W_n is the weight of an n-gram of size n (typically $\frac{1}{n}$)

Overall Usability

- We also classify the overall usability of the text. We consider 2 types of experiments here:
  - Overall Classes – Feature set is the output class of grammaticality, meaning and simplicity
  - Overall Values – Feature set includes the values of the individual features used to classify the grammaticality, meaning preservation and simplicity

Experimental Setup

- Each of the problems is viewed as a classification problem, with classes good, ok, and bad.
- The classifier used is REPTree (we used other classifiers, like Naïve Bayes, Multilayer Perceptron, SVM, but they gave the same baseline outputs).
- In all experiments, the majority class (good) is used as the baseline.

Results for Overall Usability

Experiment	Accuracy (O)	MAE (O)	RMSE (O)
Training Set – Baseline	43.76	33.17	46.51
Training Set – Classes	45.74	31.39	44.67
Training Set – Values	56.23	23.56	36.70
Test Set – Baseline	43.65	21.87	40.52
Test Set – Classes	33.33	43.46	47.83
Test Set – Values	39.68	34.92	42.97

Conclusions

- Grammaticality and meaning preservation give the best results and best results above the majority baseline.
- Simplicity is still a challenging area to classify. This is also one of the reasons why the overall output suffers.

References

- [1] Porter, M. F. (2001). Snowball: A language for stemming algorithms.
- [2] Fellbaum, C. (1998). WordNet: An electronic database.
- [3] Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.
- [4] Biran, O., Samuel, B., and Elhadad, N. (2011). Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, volume 2, pages 496 – 501. Association for Computational Linguistics.

Contact: {sam,pb}@cse.iitb.ac.in