LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

International Conference on Similarity Search and Applications

SISAP 2025 - Reykjavík, Iceland

CAU

SCAPE

Semantically Context-Aware Password Generation using Word Embeddings

Nadine-Sarah "Nanni" Schüler, Maximilian von Zastrow, Tobias Vent, Michael Eichberg

Abstract

SCAPE is a novel password-guessing method that extends Probabilistic Context-Free Grammar (PCFG) based approaches by introducing context awareness through word embeddings and similarity search. Evaluated on real-world datasets, including RockYou and various forum leaks, SCAPE consistently outperforms state-of-the-art methods. Our results show that combining classic NLP techniques with semantic similarity search is a powerful and efficient strategy for password guessing in cyber-security applications.

What is Password Guesing?

- Goal: Identify an unknown string used as a password with the help of leaked password lists in an offline setting
- 3 rules for a safe password:
  - ▷ Use sufficient complexity – a strong mix of symbols and adequate length.
  - ▷ Use uniqueness – every service should have its own password.
  - ▷ Use validation – regularly check passwords against recent data breaches.
- Two types of users:
  - ▷ Security-aware users – follow best practices, use password managers, and comply with all three rules ▷ passwords are almost impossible to guess.
  - ▷ Everyday users – prioritize convenience, choosing passwords that are easy to remember or type ▷ our use case.

How Does SCAPE Work?

- SCAPE is based on the "*Pretty Cool Fuzzy Guesser*" by Weir et al.[1], a password-guessing tool that uses probabilistic context-free grammars (PCFGs).
- When training, passwords are generalized into base structures, categorized by type and length: A for letters (alphabetic strings), D for digits, and O for other symbols, punctuation, etc.
- For example: *pass1word234!* is represented as base structure $A_4D_1A_4D_3O_1$.
- The resulting grammar tables are iteratively expanded with semantic information derived from word embeddings.
- This allows SCAPE to generate meaningful new password candidates that include words not seen in the original training set.

Results

	SCAPE	PCFG+ OMEN	SePass	OMEN	Semantic PCFG	Hashcat Best64	PassGPT
Hashmob Medium	21.20%	31.27%	NR†	6.66%	9.51%	* 35.14%	13.59%
HIBP? Subset	1.38%	2.65%	NR†	0.46%	0.20%	* 0.61%	0.36%
RockYou Subset	36.39%	29.15%	18.45%	20.06%	14.29%	16.99%	19.09%
zxcvbn	75.26%	58.01%	56.68%	51.12%	34.69%	49.12%	62.59%
Fotoboom	21.14%	23.47%	10.43%	17.87%	10.05%	16.36%	14.83%
Gunforum	36.96%	33.09%	11.26%	17.47%	9.46%	15.60%	1.50%
Vapersforum	27.60%	19.21%	9.27%	13.43%	10.00%	19.12%	9.97%
Kiteforum	31.81%	27.46%	14.45%	23.86%	11.13%	29.94%	4.50%

Table 1: Comparison of the results of all tested attack methods on seven different data sets. This is the final percentage of hits after 10 Mio. guesses. †NR=No result - the algorithm was aborted after 5 days of runtime. *Results with Hashcat's list longer than 10^7 guesses.

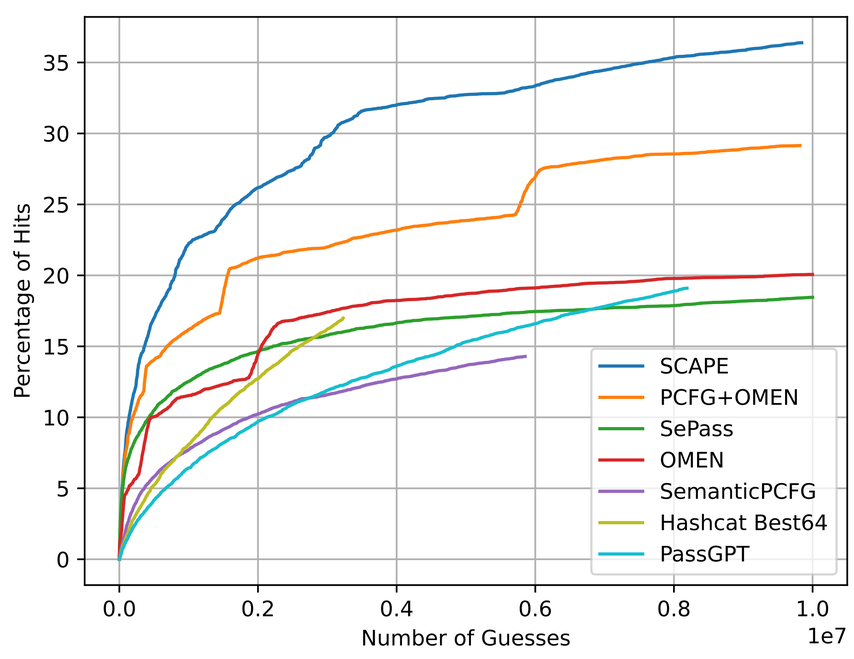

Figure 1: Comparison of all algorithms trained and tested on the RockYou subset. Each graph shows the percentage of correctly guessed passwords after a specified number of attempts.

SCAPE's Unique Password Candidates

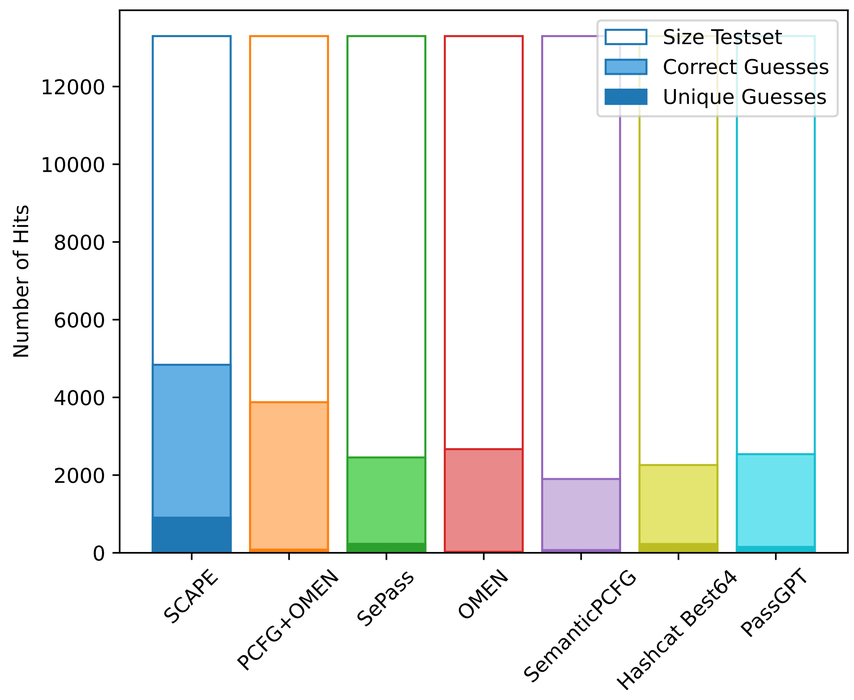

Figure 2: A comparison of all algorithms' ability to generate unique suggestions on the RockYou subset. SCAPE not only produced the highest number of correct guesses overall but also generated nearly 20% of unique and correct candidates that were not produced by any other algorithm.

Conclusions

In conclusion, SCAPE consistently performs well across diverse datasets of varying size and origin, outperforming all tested methods on 5 of 8 datasets. In the remaining cases, it ranks second when Hashcat is restricted to the same 107-guess limit. By combining the statistical efficiency of PCFG with SePass-style semantic expansion using word embeddings, SCAPE improves hit rates while preserving PCFG's comparatively low runtime. This underscores the value of incorporating natural language semantics into password-guessing models. While this work relies on static word embeddings, future research could explore dynamic embeddings and the semantic capabilities of large language models (LLMs), although leveraging LLMs as word embeddings is non-trivial and poses significant integration challenges.

References

[1] Matt Weir, Sudhir Aggarwal, Breno De Medeiros, and Bill Glodek. Password cracking using probabilistic context-free grammars. In *2009 30th IEEE symposium on security and privacy*, pages 391–405. IEEE, 2009.

Contact Me

Email: N.Schueler@lmu.de