From NLP to MLP

Peter Teufl/Udo Payer/Günther Lackner
Graz University of Technology, Austria
What to expect...

- Knowledge mining in various areas:
  - Event correlation, RDF data analysis, WIFI privacy, malware analysis
  - ... and e-Participation
  - Common model based on ML and AI
e-Participation

- Natural Language Processing (NLP)
- Analysis of text
- Clustering
- Relations
- Semantic aware search
Idea...

• In previous work, analysis of polymorphic shellcodes

• Neural networks for the detection of shellcode decryption loops

• Code vs. natural language?
NLP

• Natural Language Processing
• Various techniques are available
• Sentence analysis
• POS tagging
• Word sense disambiguation
MLP

- Machine Language Processing
- In this talk: focus on assembler, but applicable to any other language
- Assuming that machine language has similar properties as real language
- Semantics, grammar, (word/instruction) disambiguation
MLP

• Malware: Signatures? Understanding?

• Fuzzy analysis

• Group code with similar behavior

• Semantic search for similar code

• Semantic relations within code
Goal

Can we map NLP to MLP???

NLP research highly developed, mature frameworks, techniques

Problems in language analysis easier to spot, since we know our language very well
Lexical parser

- **NLP:**
  - Extract sentences, analyze terms, find relations (e.g. Stanford parser)
  - This beautiful_A city_N is_V called_V St. Petersburg_SN.
Lexical parser

- **MLP:**
  - instruction sequences (mov, sub, xor...)
  - relations between typical instructions (modifying the same register, variable)
  - e.g. interrupt preparation, loops e.g.
POS tagging, filter

- **NLP:**
  - This beautiful _A city_ _N is_ _V called_ _V St. Petersburg_ _SN._
  - Tagging terms
  - Filtering (stop words etc.): beautiful, city, St. Petersburg
POC tagging, filter

- MLP:
  - jmp, call, jz... (branch type)
  - add, sub... (arithmetic)
  - xor... (logic)
Lemmatization

• NLP:
  • bought => buy
  • cars => car

• MLP:
  • drop parameters (mov ax,4)
  • group instructions (arithmetic, logic)
Activation Patterns

• Based on semantic networks, spreading activation, machine learning

• Allows us to analyze arbitrary combination of features (symbolic, real values)

• Patterns are the basis for a wide range of analysis methods
Analysis

- Unsupervised learning (clustering)
- Supervised learning (classifiers)
- Semantic search
- Semantic relations
- Anomaly detection
- Feature relevance
Applications

- e-Participation
- Text analysis
- Event Correlation
- RDF Analysis
- Twitter Mining
- Malware Analysis
- User Tracking in WIFI Networks

Intro Idea MLP vs. NLP Lexical Parser POS(C) Tagging Lemmatization Activation Patterns Analysis
MLP Example

- Metasploit shellcodes
- Decoder loops of various decryption engines
- Clustering, semantic search and relations
Clustering

- **NLP:**
  - Group similar documents, extract key concepts from clusters, gain quick overview

- **MLP:**
  - Group similar code (instruction sequences, functions, decryption loops)
Semantic Search

- **NLP:**
  - St. Petersburg is beautiful. The city was founded in 1703.

- **MLP:**
  - Search for similar concepts (decryption loops)
Semantic Relations

- NLP
  - POS tagging
  - Lexical parser
  - Activation patterns

- MLP
  - Lemmatization
  - Analysis
Thoughts...

- Same basic principles for NLP and MLP
- Use the existing knowledge, bring it to another domain
- Apply it to arbitrary language
- Understand malware/programs???
Thank you for your attention!