Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Digital Innovation

Master Thesis

Easy-to-Read (E2R) Adaptation of Figurative Language

Author: **Jelle van Lieshout**

Supervisor: **Dr. Mari Carmen Suárez–Figueroa**

Madrid, MM YYYY

Abstract

Figurative language, such as idioms and metaphors, presents a significant barrier to text accessibility and downstream Natural Language Processing tasks. While Large Language Models (LLMs) have shown promise in general language understanding, their ability to reliably handle figurative expressions in zero-shot scenarios remains under-explored compared to task-specific supervised models. This thesis investigates the effectiveness of agentic LLM pipelines for the detection, interpretation, and Easy-to-Read (E2R) adaptation of figurative language.

We propose a multi-stage agentic workflow that explicitly decomposes the task into detection, explanation, literal replacement, and self-verification. Using the SemEval-2022 Task 2 dataset for idioms and the VU Amsterdam Metaphor Corpus for metaphors, we evaluate the system's performance against both monolithic single-prompt LLM baselines and established supervised benchmarks.

We hypothesize that an agentic decomposition significantly improves the quality of literal replacements compared to standard prompting, particularly for complex metaphors that rely on abstract source-target mappings. Furthermore, we explore the trade-offs between semantic fidelity and readability in the generated replacements. This work aims to demonstrate that structured, observable agentic reasoning enables reliable figurative language interpretation and transformation without the need for extensive task-specific training, thereby contributing to more accessible and inclusive digital content.

Resumen

...

Acknowledgement

...

Contents

List of Tables

List of Figures

List of Equations

1 Introduction

- Motivation: figurative language as a barrier to accessibility, simplification, and downstream NLP
- Limitations of supervised, task-specific approaches
- Rise of LLMs and agentic reasoning
- Contributions
  - Zero/one-shot/few-shot agentic system for figurative language handling
  - Unified treatment of idioms and metaphors (is this the case?)
  - Empirical evaluation of replacement quality [1]

2 Background & Related Work

2.1 Figurative language

- Idioms vs metaphors (linguistic + cognitive distinction)
- Prior work on detection, classification, and simplification

2.2 Datasets

- SemEval-2022 Task 2 (idioms)
- VU Amsterdam Metaphor Corpus
- Supporting datasets (MAGPIE, TroFi, MOH)
- Conceptual resources (MetaNet)

2.3 Large Language Models & Agentic Systems

- Zero-shot, one-shot and few-shot prompting
- Agentic decomposition (planning, reflection, self-verification)
- Observability (LangSmith-style traces)

2.4 Literature Review

Relevant papers to include:

- *First steps in the development of a support application for easy-to-read adaptation* (Suárez-Figueroa et al., 2024) - E2R Methodology
- *Towards an Automatic Easy-to-Read Adaptation of Morphological Features in Spanish Texts* - E2R Methodology
- *Metaphors and Analogies in the Context of Large Language Models* (Dmitrijev et al., 2024)
- *An analysis of language models for metaphor recognition* (Neidlein et al., 2020)
- *Testing the Ability of Language Models to Interpret Figurative Language* (Liu et al., 2022)
- *A Survey on Automatic Generation of Figurative Language* (Lai & Nissim, 2024)
- *Large Language Model Displays Emergent Ability to Interpret Novel Literary Metaphors* (Ichien et al., 2024)
- *Curriculum-style Data Augmentation for LLM-based Metaphor Detection* (Jia et al., 2025)

3 System Design & Methodology

3.1 Problem formulation

- Input: raw text
- Output: literalized, meaning-preserving text + intermediate explanations

3.2 Agentic pipeline

- Figurative span detection
- Interpretation / explanation
- Literal replacement generation
- Self-verification & revision
- Self-learning through addition to RAG?

3.3 Baselines

- Monolithic single-prompt LLM
- Detection-only prompting
- Naïve paraphrasing

4 RQ1: Detection & Interpretation

To what extent can a zero-/one-shot agentic LLM pipeline accurately detect and interpret idiomatic and metaphorical expressions in context?

4.1 Experiments

- Idioms: SemEval-2022 (detection / idiomaticity)
- Metaphors: VUA (token-level metaphor detection)
- Compare:
  - Agentic pipeline
  - Single-prompt LLM
  - Reported supervised benchmarks (from literature)

4.2 Metrics

- Accuracy / F1 (detection)
- Qualitative interpretation correctness

5 RQ2: Agentic Decomposition

Does agentic decomposition improve figurative language handling compared to monolithic prompting?

5.1 Experiments

- Same inputs, different architectures
- Ablation:
  - no explanation step
  - no self-verification
  - full agentic pipeline

5.2 Metrics

- Replacement quality (see Chapter 6)
- Error types
- Failure traceability

6 RQ3: Replacement Quality

How effectively can figurative expressions be replaced with literal, meaning-preserving alternatives?

6.1 Idioms

- SemEval-2022 Subtask B–style evaluation
- Semantic similarity to gold paraphrases

6.2 Metaphors

- Custom evaluation protocol

6.3 Analysis

- Idioms vs metaphors
- Trade-off: readability vs semantic fidelity

7 RQ4: Observability & Error Analysis

How observable and debuggable are agentic systems compared to end-to-end prompting?

- Case studies using LangSmith traces
- Error localization
- Correlation between explanation quality and outcome quality

8 Discussion & Limitations

- What agentic systems can/cannot do
- Where metaphors fundamentally resist literalization
- Limitations:
  - Conceptual: metaphors encoding meaning that cannot be fully literalized; replacement oversimplifying nuance
  - Evaluation limits: Subjectivity of human evaluation; no gold standard for metaphor replacement
  - Model dependence: Results vary across LLM providers
  - Scope limits: Sentence/MWE level focus; no discourse-wide tracking; primarily English

9 Conclusion & Future Work

- Summary of findings
- Implications for NLP accessibility
- Directions: multilinguality, discourse-level metaphors, human-in-the-loop

References

[1] A. Author. Sample article title. *Journal Name*, 1(1):1–10, 2024.

10 Annex

...