UAlberta at SemEval 2023 Task 1: Context Augmentation and Translation for Visual WSD

Michael Ogezi, Bradley Hauer, Talgat Omarov, Ning Shi, Grzegorz Kondrak
mikeogezi@ualberta.ca

amii
UNIVERSITY OF ALBERTA

1. Overview

Task: Given a focus word ***f*** in context ***c***, and a set of candidate images ***I***, determine which image ***i**** ∈ ***I***, best depicts the meaning of ***f***

Our ideas:

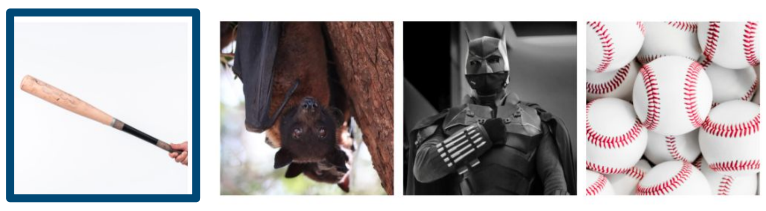
1. Produce extra context to augment the original context
2. Translate non-English samples to English
3. Score images based on similarity to context

Our methods:

1. Use a language model to generate extra context for augmentation
2. Apply a pair-wise image-scoring algorithm to select ***i****

Example:

- **Focus word:** bat
- **Context:** baseball bat



3. Method: Translate & Augment

Translation

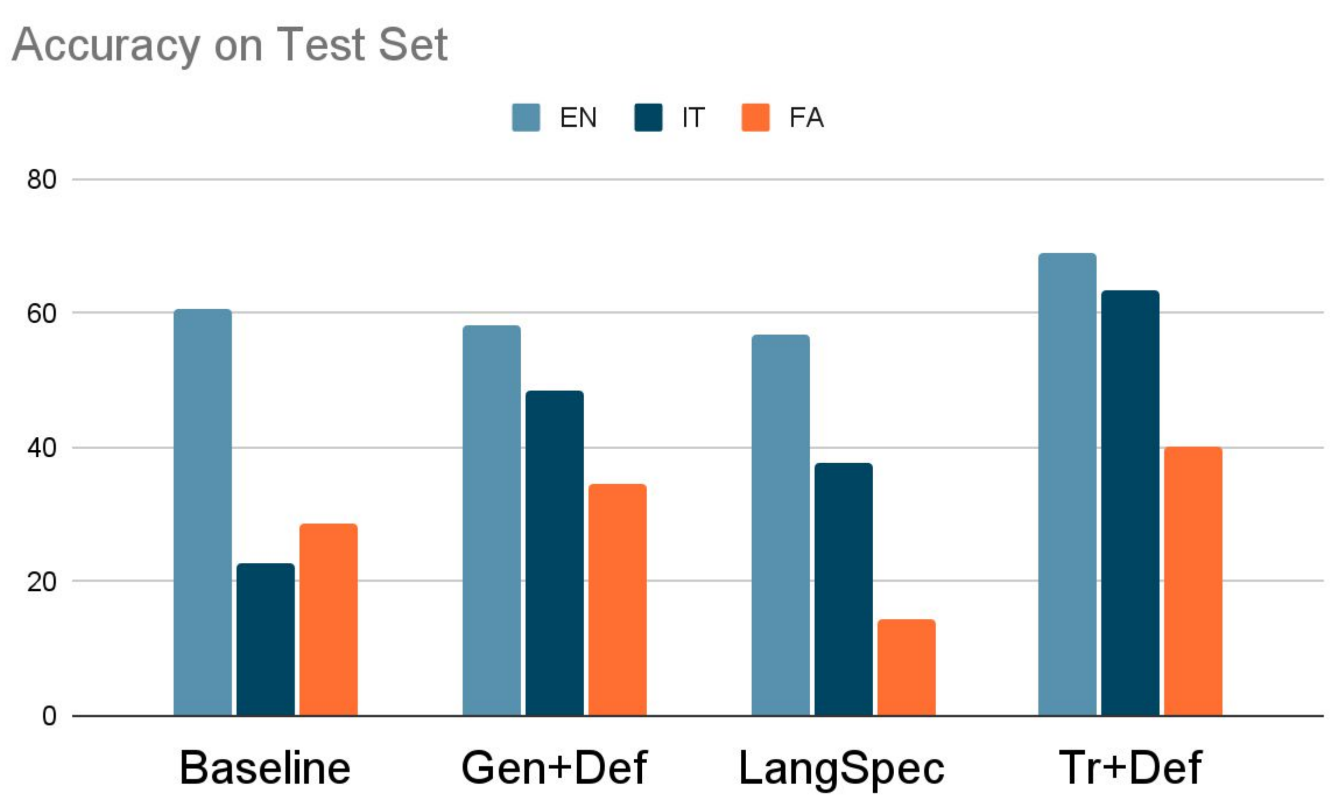
Focus word: gomma
Context: gomma per smacchiare (IT) → eraser (EN)

Context Augmentation

Context: baseball bat → A baseball bat is a cylindrical club used to hit a baseball

5. Experiments & Results

- **Baseline:**
  - We compare the candidate image with the highest similarity to the full context.
- **Image Generation with Context Augmentation (Gen+Def):**
  - We compare the candidate image with the highest similarity to an ensemble of **images** generated from **translated** + **augmented** context.
- **Language-Specific (LangSpec):**
  - We use compare the candidate image with the highest similarity to the full context using **language-specific** (EN|IT|FA) models.
- **Translations with Context Augmentation (Tr+Def):**
  - We compare the candidate image with the highest similarity to the full **translated** + **augmented** context using **English** models.



2. Tools & Resources

- **Dataset:**
  - **SemEval 2023 Task 1 dataset: <*f, c, I, i**>**
- **Lexical resources:**
  - **BabelNet:** focus word senses and glosses
- **Models:**
  - **CLIP:** pre-trained vision-language model for image similarity
  - **BERT:** pre-trained masked language model for text similarity
  - **ChatGPT:** general-purpose language model for translation

4. Method: Score Candidate Images

$c \leftarrow$ the context of the focus word
$G \leftarrow$ list of glosses for the focus word
$I \leftarrow$ list of candidate images
for i in I **do**
  $S_g \leftarrow$ empty list
  for g in G **do**
    $s_{ig} \leftarrow sim^{VL}(i, g)$ ▷ image-gloss score
    $s_{cg} \leftarrow sim^{L}(c, g)$ ▷ context-gloss score
    $S_g.append(s_{ig} + s_{cg})$
  $s_{ic} \leftarrow sim^{VL}(i, c)$ ▷ image-context score
  $scores[i] \leftarrow S_g.max() + s_{ic}$
return $scores$

6. Error Analysis

- **Context Augmentation helps with ambiguity:**
  - ❌ **Original Context:** andromeda tree
  - ✅ **Augmented Context:** An andromeda tree is a Japanese tree with light-colored flowers and green leaves
  - **CLIP needs help to understand nuanced meaning**

7. Conclusion

- **We find that context augmentation is a powerful tool in improving performance**
- **Applying the English-only CLIP model to automatically translated text, yields higher accuracy than language-specific CLIP to original languages.**
- **Standard SOTA WSD systems have difficulty disambiguating short contexts, and are not necessarily effective for V-WSD.**

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Alberta Machine Intelligence Institute (Amii).