When it comes to the interaction between humans and computers, natural language plays an increasingly important role, e.g. in personal assistants such as Alexa. In this project, we aim to analyze natural language and infer semantic properties of words. The recently proposed notion of semantic tagging (Bjerva et al. 2016) aims at labeling words with tags that reveal their properties and are useful in semantic tasks. While a training corpus exists, its coverage is limited. We propose a way to automatically extend the coverage by drawing on large-scale word representation data to derive a large new Semantic Tag lexicon. Our experiments show that we can infer semantic tags for unseen words with high accuracy.

**Motivation**

- Previously, Part-of-Speech tagging and Named Entity classes do contribute to determine lexical semantics but they are not sufficiently informative.
- Semantic tagging incorporate semantic virtues of these two tasks and fill gaps in semantic modeling by adding new categories.
- An example of parsing the sentence The dog attacked the little boy.

The dog attacked the little boy.

The dog

**DEF**

> **CON** concept: dog, person

*EPS* past simple: ate, went

attacked

**DEF**

the

**DEF**

little

**IST** interactive: open, vegetarian, quickly

boy

**CON** concept: dog, person

NIL empty semantics: (), to, of

- Given the usefulness of semantic tags, it is helpful to create a large cross-lingual semantic tagging dataset.

**Methodology**

Generate Semantic Tag Vectors from PMB V2.1.0 Dataset

A "1" means token \( \omega \) is marked once with the semantic tag at this position

\[
\sum_{\omega} \begin{cases} 0 & 1 \\ 1 & 32 \\ \cdot & \cdot \\ 0 & 0 \end{cases}
\]

Resulting semantic tag vector for token \( \omega \)

Each position corresponds to a unique semantic tag

Find k Nearest Neighbors

using Stanford GloVe word vector

Neighbors are often variations of original word

Example: notice \( \rightarrow \) notices, noticed

Lower Prediction Accuracy

Take k nearest neighbors which exists in PMB V2.1.0

Higher Prediction Accuracy

using Sketch Engine word Embeddings (PoS+Lemma)

Neighbors are same or similar POS property

Example: Notice-\( v \) \( \rightarrow \) see-\( v \), spot-\( v \)

Predict Semantic Tag Vector

Combine k vectors by taking average weighted by cosine similarity scores between \( \omega \) and \( N_i \)

\[ N_{1} \cdot 0.02 \cdot 0.03 \cdot 0.01 \]

\[ N_{K} \cdot 0.02 \cdot 0.03 \cdot 0.01 \]

Prediction tag vector for \( \omega \)

\[ \omega \cdot 0 \cdot 0 \cdot 0 \]

\[ \cdot \cdot \cdot \]

\[ \cdot \cdot \cdot \]

\[ \cdot \cdot \cdot \]

\[ Avg \cdot 0 \cdot 46.3 \cdot 1.4 \]

Future Work

- Generating semantic tag datasets for over 370 languages and make it freely available for download.
- Keep our dataset up-to-date as Parallel Meaning Bank is constantly updating their data.

**References**
