

Task and Motivation

*Travel by bus is
expensive , bored and
annoying .*

*Traveling by bus is
expensive , boring and
annoying .*

- ❑ **Grammatical Error Correction (GEC)** is the task of automatically correcting ungrammatical text.
- ❑ GEC systems help learners improve writing skills and allow native speakers spot errors.
- ❑ GEC datasets are scarce, therefore, the research community has developed methods to generate synthetic data.

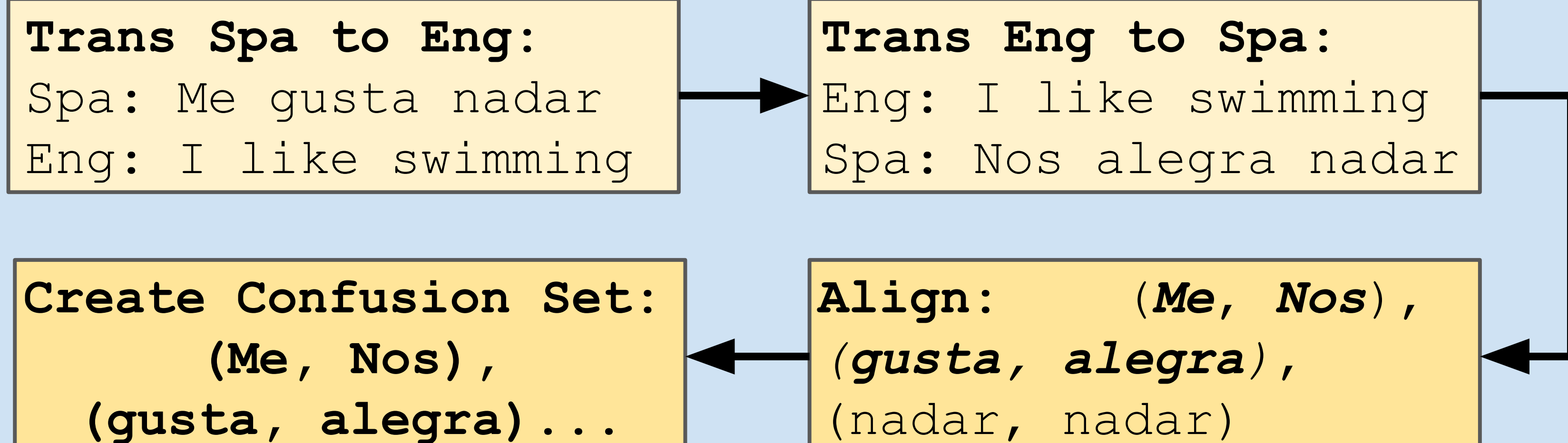
Approach

- ❑ We use **round-trip neural machine translation (NMT)** to generate diverse confusion sets.
- ❑ Confusion sets are groups of words easily confused with each other. *e.g {there, their, they're}, {cite, sight, site}*
- ❑ Use confusion sets to replace words in training corpora to synthetically generate grammatical errors.
- ❑ **Intuition:** Lexical errors, common in language learners, appear in translation systems.

Contributions

- ❑ Propose novel approach for generating confusion sets using round-trip NMT.
- ❑ Evaluate our approach in Spanish, a low resource language in GEC.
- ❑ Compare against known sets e.g Aspell and Unimorph.

Generating Confusion Sets



- ❑ We use monolingual text to create conf. set:
 - ❑ **BT-native:** Translate 100K Spanish sentences written by native speakers.
- ❑ We compare against previously used sets:
 - ❑ **Aspell:** phonologically and lexically similar.
 - ❑ **Unimorph:** database of morphological variants.
- ❑ Ungrammatical sentences are generated by replacing words in Spanish text with confusion set words.

Evaluation

	Precision	Recall	F0.5
<i>Unimorph</i>	56.67	29.15	47.67
<i>Aspell</i>	58.58	35.32	51.76
<i>BT-native</i>	59.09	34.16	51.56
<i>Unimorph + BT-native</i>	58.97	34.94	51.83

- ❑ We train transformer models to translate ungrammatical sentences to their grammatical versions.
- ❑ Combining *Unimorph* and *BT-native* achieves **51.83** F0.5 score, outperforming commonly used confusion sets.
- ❑ *BT-native* yields competitive GEC models compared to manually created confusion sets.

Conclusion

- ❑ Using round-trip NMT is an effective way to automatically generate confusion sets.
- ❑ Round-trip NMT conf. sets are competitive against manually created conf. sets.