Synonym Scoring

CS 505: Natural Language Processing
Final Report

Jacob Brinton
Boston University
Correspondence: jbrin@bu.edu

1 Introduction

Oftentimes when translating between languages, words do not have an exact 1-to-1 mapping. For example, consider the English verbs *to do* and *to make*. In Spanish, both of these are commonly translated to *hacer*. Conversely, both the Spanish verbs *ser* and *estar* mean *to be* in English. This project aims to quantify this discrepancy and give each word-to-word translation a proximity rating—that is, if we select an English word, we can see a list of possible translations along with a percentage of how "close" the two words are, and vice versa.

This will function similar to how a some thesauruses work, where you can see more and less relevant synonyms based on how close their precise meanings overlap. The user will enter a word and select input and output languages just as they would use a translator. They'll be returned a list of the top translations of the word along with a score for each of them showing how often, in equivalent scenarios in the two languages, those words could be translated.

Naturally, there are variations within languages and within contexts in which they are used. Further work could be done to quantify the formality of words and ensure words are not mistakenly placed close to each other when their significance is similar but their use cases are not. For example, a word that would reliably translate between English and French (and be placed next to similar words within sentences) could have more formal connotations in English. On a somewhat different note, this same process could be done but for dialects—words could be assigned scores to how well they fit into different regions. Some words are very dialectallly dependent, and as such, may not be appropriate for certain usages, although they are valid translations.

Another possible direction could be creating a "word web". In the center, the user would enter their query. The program would generate translations for this word and place them in a ring surrounding the entered word, with words more commonly related set in larger font. This could be continued to include translations (back into the original language) of the words in the first ring, and so forth.

A potential application of this general idea could be used when drafting emails or writing copy in one's second language when a native-language equivalent is available. A program could be created to scan the native-language text and then flag words in the new text that are not often seen in the original text's setting. This would work by going through each word in the native-language text and searching for equivalents in the second-language text. However, some deeper understanding of grammar differences would need to be used here, as some function words would require more than just a vocabulary-level understanding of their meaning. Due to this, this option might be more feasible in a future semester.

1.1 Goals

Successful completion of this project could be a valuable tool for human translators. This tool could be used to check translations and determine which words are tricky to work with, helping flag certain words that might be hard for a machine to translate and as such require more care. Furthermore, it would simply be interesting to take a look at how languages are related—one could see which words are shared between languages in not a lexical sense but a semantic sense. Two words—in some cases even spelt the same, such as *actual* in Spanish and English, have evolved in meaning over generations from when they were initially borrowed—the Spanish version means "current" or "of the present", a difference from how English-speaker would most often use it to mean real, existing, or to emphasize

importance.

For language learners it could be used to determine which words are more easily relied on and which are best used with caution. It could be used to find new words to use in lieu of common translations we might get in the habit of making. As for the potential option of assigning words a formality score, this could aid learners in developing a more conversational, friendly tone that many second language learners lack, having learned mostly in a classroom setting. Adding dialectal scores could assist learners who are focused on a particular area or culture to tailor their vocabulary towards that of their chosen dialect.

2 Methods

2.1 Data

For this project, I am using the Tatoeba v2023-04-12 dataset, which is freely available at tatoeba.org and which I accessed at opus.nlpl.eu/Tatoeba. Tatoeba provides a dataset of sentence pairs contributed by volunteers, covering a wide range of languages, which makes it a valuable resource for parallel data. Scanning through possible datasets, this dataset seemed more human-crafted and contained more natural translations to my eye. Additionally it has a good mix of everyday and non-technical language which was lacking from some other datasets. This dataset would hopefully be reasonably applicable to contexts in which the everyday person would find the "synonym ratings" useful or interesting.

The dataset was provided in a `.txt` format, with each language's sentences in separate files. To prepare it for use, I performed the following preprocessing steps:

Loading and Alignment: I read the files for each language, ensuring that each line in the English language file corresponds to the same line in the Spanish language file, and stored these aligned sentence pairs in a pandas DataFrame.

Tokenization: Using `sacremoses` (a Python wrapper for Moses tokenizer), I tokenized the sentences in both languages to handle punctuation, contractions, and language-specific rules. Next, I lemmatized the sentences using `spaCy`. For the English model I used `en_core_web_sm`, and for Spanish I used `es_core_news_sm`. As they are from different types of sources I initially was unsure about using them, but as they are the defaults for `spaCy` and they are only being used for lemmatization, so I'm hoping they won't affect accuracy.

In order to think of the ideal metric, let us first think about what exactly we are trying to measure. We are trying to create a scale to determine the "closeness" of words between languages and go beyond a simple true/false relationship on whether or not two words are synonyms. It stands to reason that a good way to measure this would be the percent of the time that a translating word would fulfill the same role in the translated sentence as it did in the original sentence. That is, we can use the source sentences to create the benchmark by counting the number of times a given word is translated to a certain synonym in the translated sentence and divide by the number of times the given word was present in the original sentence.

2.2 Baseline

As a baseline I can use the dictionary/translation dictionary and assign 100% to words that appear in the dictionary as translations and 0% to words that are not translations. To this end I am using `deep_translator` package which uses Google Translate. Each word returns a single word as its translation, so any synonyms are ignored.

To evaluate the baseline, we first create a list of 100 random English words in the datasest. We can then create a subset of the sentences that include this English word. Now, we may calculate the average percent of the Spanish sentences which contain this word (to which our baseline assigns 100% confidence). Additionally, the same can be done for words which are not considered translations by the baseline and sum up these scores. In testing for words that are considered translations by the baseline, 40.0% of sentences actually contained the Spanish translation of the word.

For example, the token '2013' was only present in 71% of the Spanish translations of sentences containing '2013', indicating that some sentences may be translated in a looser or more creative manner, so our metric may not be perfect here. Additionally, was can look at the case where the word "Tom" appeared in an English sentence, but the Spanish equivalent only appeared 4% of the time in their translated sentences, presumably because the translation kept "Tom" in its original form. There are other cases where single letters, such as "I" are translated (not having more context) by Google to just be "I" in Spanish, although what was probably meant in context was not just the letter but the pronoun.

There is also something to be desired in the fact that certain words are simply more common in a certain language, although they are valid translations for multiple words in the translated language. That is, certain words should be given more "weight" to distribute to synonyms in the target language.

3 Experiments

All of the English words in the dataset were loaded in a dataframe ordered by their first appearance, along with their part of speech as tagged by SpaCy. Then each 10th word in the dataframe was translated to Spanish and each of their synonym scores were calculated. The averages for each part of speech are displayed in Table 1.

The calculation of the synonym score is a conditional probability of the target word appearing in target sentences given that the source sentence contained the source word. Given a list of sentence pairs (s_t, s_s), we sum over all cases where both the source and target word were found in their respective sentences and divide by the number of source sentences containing the source word.

$$p(w_t|w_s) = \frac{\sum_{(s_s,s_t)\in S} \mathbb{I}(w_s \in s_s)\mathbb{I}(w_t \in t_s)}{|C(w_s)|}$$

Part of Speech	English	Spanish
ADJ	53.57	55.39
ADP	81.16	23.03
ADV	37.79	49.23
AUX	16.92	45.76
DET	–	22.99
INTJ	43.04	50.00
NOUN	57.07	58.72
NUM	70.83	90.17
PRON	22.29	46.38
PROPN	58.96	74.28
VERB	34.48	50.87

Table 1: Synonym scores by part of speech for English and Spanish

3.1 Sentence Embeddings

I attempted to implement a scorer using sentence embeddings. However, there were many dependancy issues in the several packages I tried with pretrained multilingual embeddings—i.e. sentence embeddings for Spanish are in the same vector space as English, so sentences that are translations of each other, despite being from different languages, are close to each other. Another technique is to create separate feature spaces for both languages and then learn an alignment matrix to map them neatly onto each other.

Once I am able to get the sentence embeddings functioning, I will be able to create another metric using the cosine distance between respective sentences. First, the dataset will be filtered by all sentences that contain the source word. Next, the embeddings of all target sentences in the filtered dataset containing the target word will be averaged. So that it coincides with our previous metric as far as scale, the synonym score assigned to the source and target word pair will be a rescaling of the cosine distance between this average target sentence embedding and the original embedding of the source sentence.

3.2 Examples

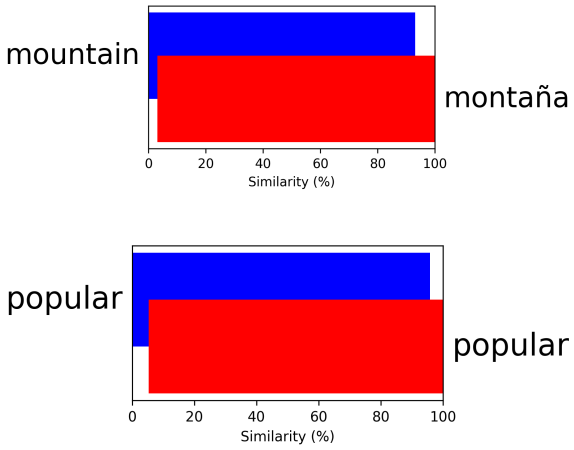


Figure 1: Cognates

Figure 1 shows how cognates commonly demonstrate high similarity scores. However, not all cognates show a 2-sided relationship: many words have more formal pretenses when used in English, thus leading them to be more common in Spanish. Figure 2 demonstrates the pair (indicate, indicar) in which the Spanish equivalent is broader than the English word. In Spanish, indicar can be used to mean "to point out" or even "to show", whereas its English equivalent has a narrower denotation.

Oftentimes, it's difficult to take into account dialectal differences, especially when corpora are organized by language. In Figure 3, we see how

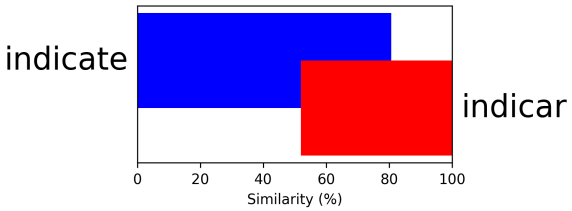


Figure 2: Commonality shift

the English word "juice" is split between translations into Spanish. Despite the fact that the Spanish equivalents nearly always translate back to "juice", about a third of the time, we get the word commonly used in Spain ("zumo"), and most of the rest of the time the Latin American word, "jugo", is seen.

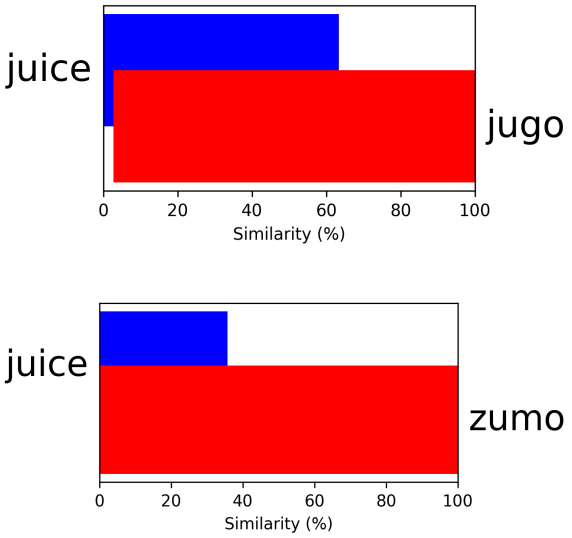


Figure 3: Dialectal variations

One particular domain in which synonym scores struggled was pronouns. An aspect that contributed to this was the method used for lemmatization, especially when considering that Spanish is a pro-drop language. In general, but even more so for third-person pronouns and certain tenses, pronouns must be disambiguated through setting. Additionally, they may appear as clitics attached to the word, which is difficult to catch when tokenizing, as they might mean another, unrelated word given the context.

In Figure 4 (On the next page), two general trends can be noted. First, English includes pronouns with greater frequency (summing up the blue bars on the left-hand side, about a four-fold increase). About 75% of the time, when an English sentence contains a second-person pronoun (I suppose we are ignoring cases where "yall" or even "thou" might have appeared), its Spanish lemmatization did not include a second-person pronoun. We also see that the dataset uses the informal second-person pronoun nearly 8 times more than its formal equivalent. This metric would surely change depending on the data: for example, government documents would contain more formal language.

4 Conclusions

We can see that synonym scores can be a useful way to compare words between languages—to see which words are more commonly translated and which ones are "wider", or can take on more meanings than others. Overall, we were able to align well with what we would expect synonyms to score given a dictionary.

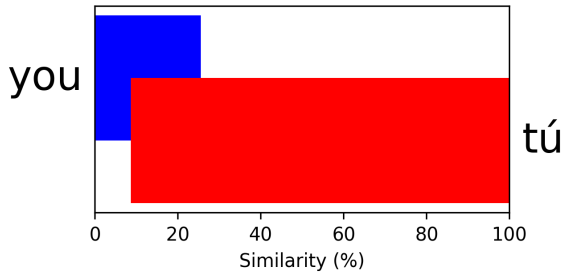
Further work on this subject, such as trying different embeddings, could be used to refine my approach and balance its drawbacks. In the future I also hope to add sentence and part of speech context to translations, to make them more accuract for evaluating over the whole corpus. On that note, I do think it would be interesting to apply this to other corpora and see if there are significant differences in the formality or dialectal features, and if they would affect the synonym scores.

5 Replicability

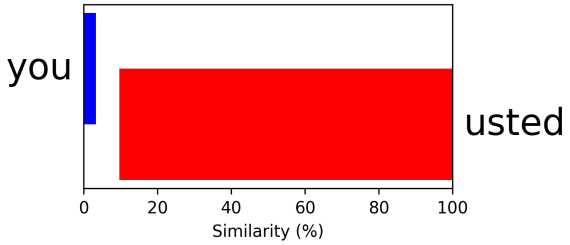
My code can be found at the following link:
https://github.com/jebrinton/nlp-final-project

6 Citations

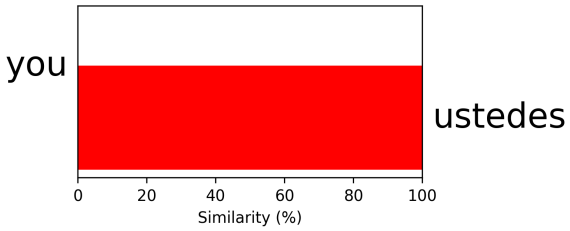
Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In Proceedings of the Fifth Conference on Machine Translation, pages 1174–1182, Online. Association for Computational Linguistics.



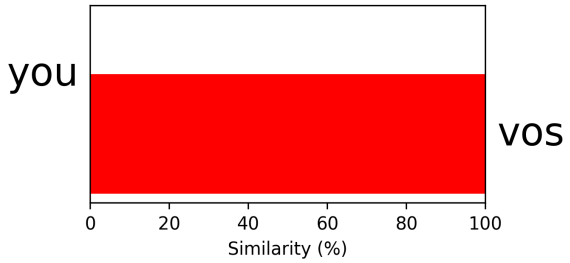
(a) Most common informal singular form



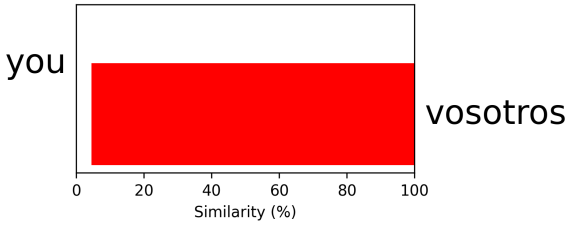
(b) Formal singular form



(c) Plural form used in Latin America; formal connotations in Spain and Africa



(d) Informal singular form used in some areas of Latin America



(e) Informal plural form found only in Spain and Africa

Figure 4: Second-person singular and plural pronouns