Spring 5-2018

Proverbial Machine Translation: Translating Proverbs between Spanish and English Using Phrased Based Statistical Machine Translation with the Grammatical Category Based Approach

Teneala Spencer
University of Southern Mississippi

The University of Southern Mississippi

Proverbial Machine Translation: Translating Proverbs between Spanish and English Using Phrased Based Statistical Machine Translation with the Grammatical Category Based Approach

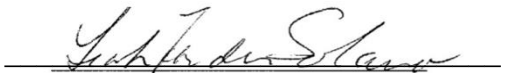
by

Teneala N. Spencer

A Thesis
Submitted to the Honors College of
The University of Southern Mississippi
in Partial Fulfillment
of the Requirement for the Degree of
Bachelor of Science
in the Department of Computer Science

April 2018

Approved by



Leah Fonder-Solano, Ph.D., Thesis Adviser Professor of Spanish

Tom Rishel, M.S., Thesis Adviser Instructor of Computer Science

Christopher Miles, Ph.D., Chair Department of Spanish

Ellen Weinauer, Ph.D., Dean Honors College

Abstract

The focus of the research presented in the paper is translating the non-literal interpretation of proverbs from Spanish to English without changing their intended meaning. Proverbs have limited variation in comparison to slang, poetry, and metaphors which tend to differ significantly within the context that they are used. Although there are many other approaches for machine translation (MT), the one most suitable for the research project is Phrased Based Statistical Machine translation used in conjunction with the Grammatical Category- Based approach.

Key Words: Machine translation, phrase based, statistical, grammar

Dedication

Jesus, Randy Spencer, Felecia Spencer, and Rande Spencer:

Thank you for your constant spiritual, physical, and emotional support!

No one can make it on their own!

I have the greatest support system any college student could ever ask for!

There aren't enough words in my vocabulary to express my gratitude towards God and my family. They were essential and absolutely necessary to my success.

Could never have completed undergrad without you guys! Gracias a Dios!

Acknowledgements

I want to thank both of my thesis advisors, Dr. Leah Fonder-Solano and Tom Rishel, for all of the support and guidance they gave to me through the duration of my research experience. Their insight into the research process was vital in forming my ideas and completing the program. They are without a doubt the most influential professors I've met here at USM, and have impacted my life in ways I was not expecting.

Table of Contents

List of Illustrations

List of Tables

List of Abbreviations

MT	Machine Translation
SMT	Statistical Machine Translation
GCB	Grammatical Category Based
NLP	Natural Language Processing

Proverbial machine translation: correcting translation of proverbs From Spanish to English using phrase based statistical machine translation In conjunction with the grammatical category-based approach

Teneala N. Spencer
Department of Computer Science and Spanish Linguistics
University of Southern Mississippi
Long Beach, USA

Abstract— The focus of the research presented in the paper is translating the non-literal interpretation of proverbs from Spanish to English without changing their intended meaning. Proverbs have limited variation in comparison to slang, poetry, and metaphors which tend to differ significantly within the context that they are used. Although there are many other approaches for machine translation (MT), the one most suitable for the research project is Phrased Based Statistical Machine translation used in conjunction with the Grammatical Category-Based approach.

I. Introduction

Machine translation (MT) is an area within computer science that deals with how to computationally translate language based on many different linguistic approaches that have been developed over time. The rule-based approach uses linguistic rules in order to translate from the source to target language. The approach takes into account word order, meaning, semantics, syntax, and other linguistic entities. The problem here is that all of those linguistic components vary significantly from language to language. Another approach used for MT is the dictionary based approach which translates languages word for word based on their meaning which doesn't take into account linguistic rules.

A separate approach that focuses solely on MT errors is the taxonomical approach. It solves linguistic errors such as orthographic, lexical, grammatical, semantic, and discourse errors by classifying the type of error that has occurred and from those classifications, a computer program can be used to detect the error in order to produce the appropriate output from source to target language. The taxonomic approach is designed to be used with the statistical machine translation (SMT) approach to see how often these errors occur and from the statistical analysis, the probability of, for example, an orthographic error could be computed. SMT refers to translations that are produced by using statistical models which are derived from an analysis of bilingual text corpora.

II. Errors in MT

MT has many types of errors that have yet to be corrected. One area in particular is the area of translating symbolic language. Symbolic language refers to the use of colloquialisms, proverbs, metaphors, similes, hyperboles and other linguistic structure that are not translated or interpreted literally. Such structures cause ambiguity amongst the meanings of non-literal phrases which serves as the main reason why machine translation of symbolic language is erroneous. A second reason is because a majority of the research that has been conducted within the realm of machine translation has mostly focused translations of literal phrases. This is problematic because language, typically, is composed of both literal and non-literal expressions. A computer cannot correctly translate a symbolic statement unless proper instructions are written to accurately translate the statement. If a symbolic phrase were used in a sentence, the computer would take the literal translation of that

particular phrase and translate it into the target language.

While there are many forms of symbolic language, the focus of my research is translating the non-literal interpretation of proverbs from Spanish to English without altering their intended meaning. I chose proverbs because they have limited variation in comparison to other symbolic linguistic structures which tend to differ significantly within the context that they are used. Language is composed of literal and nonliteral speech so in translating it, an emphasis needs to be placed on both components. Since a lot of focus has been placed on literal speech, as it is the most common, studying the problem of translating proverbs is justified in that not a lot of emphasis has been placed on nonliteral speech. The following is an example to illustrate how erroneous the translation of a Spanish proverbial expression into English is, using Google translate.

SPA: *El que no llora, no mama*.

Google Translate ENG: *The one who does not cry does not.*

Literal ENG: *The one who does not cry does not suck.*

Figurative ENG: *A closed mouth doesn't get fed.*

SPA: *¿Por qué no me preguntó? Usted debe saber que* ***el que no llora, no mama.***

Google Translate ENG: *Why did not he ask me? You should know that* ***the one who does not cry, does not breast.***

Literal ENG: *Why didn't you ask me? You should know that* ***he that does not cry, does not suck.***

Figurative ENG: *Why didn't you ask me? You should know* ***a closed mouth doesn't get fed.***

Although there are many other approaches for MT, the one most suitable for my research project is Phrased Based Statistical Machine translation used in conjunction with the Grammatical Category-Based approach. I will build upon current research using my preferred approach by modifying existing machine translation software to more accurately translate Spanish proverbs regardless of how they are used in a sentence.

III. Literature Review

In his book *Statistical Machine Translation*, Phillip Koehn talks about how SMT studies languages in parts by separating strings through tokenization. Tokenization is the process whereby a string or set of words is individualized by phrases, keywords, or symbols. A part of the process of SMT is using parallel text corpora to pair tokens or full sentences from the source language with their equivalent in the target language (Koehn). A bilingual text corpus is a lexicon of phrases that have been paired together from source to target language. SMT is also useful also in predicting the correct translation from one language into another by studying the frequency of word patterns. If the words of a particular phrase follow the same sequence every time they appear in a bilingual text corpus, then the phrase is less likely to change its meaning regardless of context. However, if that particular phrase is missing a word or even rearranged differently then the probability of that phrase meaning something else is higher. This is why SMT accounts for these instances by using mathematical probability functions that maximize likelihood estimations (Koehn).

$$t = argmax\{\sum_{m=1} \lambda_m h_m(t,s)\}$$

- The equation above is used to maximize the probability of a log-linear sequence of words or phrases. More generally, the equation is used to predict the probability of a phrase following the same sequence every

time it appears in the text corpus. (Farrús, et al., 2011)

$$t_1^i = argmax\{p(s_1^j, t_1^j)\} = \cdots$$

$$= argmax\{\prod_{n=1}^{N} p((s,t)_n|(s,t)_{n-x+1}, \ldots, (s,t)_{n-1}\}$$

The equation above is used to approximate the occurrences of tuples at the sentence level. (Farrús, et al., 2011)

SMT uses phrase based models which have proven to be more successful as opposed to word based models because translating language word for word can lead to significant linguistic and grammatical errors.

Since SMT uses phrases instead of individual words to translate between languages, it is important to keep in mind how a phrase can have many different variations when word choice is taken into account. In the paper by Thomás, Loret, and Casacuberta, "Phrase-Based Statistical Machine Translation using Approximate Matching," research has been conducted to see how closely SMT can predict the meaning of a modified phrase despite its variance from the original source phrase found in the text corpus. The authors research the problem of how to overcome the effects of generalizing a phrase and the implication it has on its meaning. The research also focused on being able to identify a source phrase that has been modified by either reordering, word substitution, insertion, or deletion. The goal was to get the computer to recognize a phrase that is very similar to a preexisting source phrase in the text corpus in order to adequately translate into the target language. The paper states that if a phrase does not appear in the training corpus then the computer cannot translate the phrase without first identifying if it is similar to a preexisting phrase within the corpus. If the source phrase matches the phrase found in the corpus word for word, then it can be translated into its target equivalent. If the phrase does not appear in the training corpus, the computer will look for a phrase that matches most directly even if it varies by one word. There is a solution that suggests using word classes as a learned way to perform an unsupervised method from a bilingual text corpus to translate the phrase. The approach, according to the article, is inaccurate due to overgeneralization of a word. Their conclusion about overgeneralization makes sense because vocabulary can alter the meaning of a phrase even if the words appear to be synonymous whether in meaning or in function.

In order to determine how closely the phrases match, computer scientists use two approximation methods: long distance and short distance phrase-based reordering. The focus of the research in the article is on short distance phrase-based reordering which refers to how much the modified phrase varies from the original source phrase. In order to determine how much a phrase varies from the one found in the training corpus, the authors use an edit distance as a source of measurement. The edit distance accounts for all the occurrences of substitutions, insertions, and deletions of words. In order to figure out which word has been altered by insertion, substitution, or deletion, an algorithm is used to determine the function of each word in the sentence and match each word of the modified phrase to each word of original phrase. If a substitution is found, then the target word is replaced by its word or phrase equivalent. If a deletion occurs, then the target word is deleted. Likewise, if an insertion is found then the new word is added to the target phrase.

A constraint on the variation of a phrase is one when ($e_{max} = 1$). This comes from the edit distance formula:

$$\forall \frac{s', t'}{p(t'|s')} \geq p_{min} \wedge EditDistance(s', s) \leq e_{max}$$

(Thomás, Loret, & Casacuberta, 2007)

The research and technique is most suitable for studying literal phrases, however, when it comes to word choice despite its function and part of speech, the meaning of a non-literal expression can change drastically.

Research on linguistic ambiguity has been done before, more specifically at the grammatical level when observing word meaning differences. In the paper, "Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair," the authors look at one of the ways to properly translate polysemic and homonymic expressions. Polysemic expressions refer to words that have more than one meaning in the source and target language. Homonymic expressions are those that have the same spelling but different meanings depending on the context of the sentence and its grammatical use of the word. Phrase Based Statistical Machine translation used in conjunction with the grammatical category-based approach (GCB) is one of the most effective and proven methods of machine translation that deals with ambiguity that arises from polysemic and homonymic expressions.

The GCB approach allows for the computer to identify the meaning of an ambiguous word depending on its grammatical function. The work talks about how in polysemic expressions, the ambiguity that lies between translating *perqué* as *porque* or *para que* is resolved using SMT along with the grammatical category-based approach. *Perqué* is the Catalán word for either the Spanish equivalent of *porque* or *para que*. In order for the computer to know which to use, it has to identify how the conjunction *perqué* functions in the sentence. In the instance *perqué* is proceeded by a verb that is in the subjunctive, the computer will translate it into *para que*. If the verb is preceded by *perqué* and it is in the indicative, then the computer will translate it as *porque*. SMT is useful in this case and others like it because the authors were able to determine the probability of the meaning of the word based on a bilingual text corpus. The text corpus used in the research contained 1.7 million sentences and of that, *para que* was only proceeded by an indicative verb 0.5% of the time. This means that the probability of *perqué* meaning *para que* when followed by a verb in the subjunctive has a 99.5% chance of accuracy.

To show homonymic words the researchers used the Spanish adverb *sòlo* vs the Spanish adjective *solo*. By using the GCB approach, the computer is able to distinguish when *solo* functions as an adverb or adjective within a sentence according to its grammatical function and with what frequency it falls in line with the words that precede and proceed it. Using a statistical approach in studying ambiguity amongst words and phrases has been proven in MT to be most effective because a dictionary based approach or even a rule based approach does not allow for exceptions like *perqué* and *solo*.

Looking at the problem of translating and distinguishing between literal and non-literal phrases through the lens of SMT with the grammatical category-based approach works well because of how colloquial phrases function grammatically and how the computer can use grammar to recognize and interpret the phrase based on how frequently the meaning of the phrase varies, in what sequence the words in the phrase occur, and what words are statistically associated with that phrase. In the paper, "That is so cool: investigating the translation of adverbial intensifiers in English-Spanish dubbing through a parallel corpus of sitcoms," the focus of the research is to see how grammar affects colloquial expressions from English to Spanish with an emphasis on emphatic language with the use of adverbs. Adverbial intensifiers are commonly associated with colloquial language in respect to emotionally-loaded phrases (Baños, 2012). The author studied translated scripts of the sitcom *Friends* and how adverbs affect their meaning. Using the script from *Friends* in both Spanish and English, the author was able to identify how colloquialisms are formulated by using the equation: intensifier + adjective, the adjective being a colloquialism itself. When translating from Spanish to English it is more common to use the superlative as an intensifier such as the suffix *–ísimo*. Being able to identify colloquialisms based on grammar and the words associated with their grammatical function is useful in translating colloquial and proverbial expressions from Spanish to English.

Research shows that grammar can be used to identify symbolic language when studying non-literal phrases. The difference between a literal and a

non-literal phrase can be the omission or insertion of a word in respect to its meaning and grammatical function. So, using the Phrase Based Statistical Machine Translation approach along with the Grammatical Category-based approach would be most appropriate in translating a non-literal phrase and in determining whether or not the deletion, substitution, or insertion of a word modifies a non-literal phrase when translating from Spanish to English. It is important to be aware of overgeneralization of a particular word within a given phrase by looking at the phrase holistically and not just on a lexical level. Grammar, word choice, and sequence the phrase is ordered should be considered when translating and neither should be preferred over the other. The method is most suitable for translating proverbs because in general, proverbs only vary by one word making it easier to detect whether or not it has new meaning.

IV. Methodology

Machine translation of figurative speech is not always correct because instead of translating the figurative meaning of an expression, the computer translates it literally. Language is not only composed of literal phrases, expressions, and speech; therefore, it is important to study non- literal speech when improving machine translation in general. The problem I intend to solve is translating proverbs from Spanish to English.

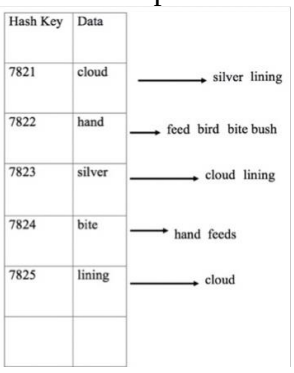
The end result of my project is Spanish and English proverbial translation application, TransVerb, with a user interface that will take in user input. There is a separate application to handle data analysis.

I have collected over 4000 Spanish and English proverbs and their equivalents and have stored them into a database. For the design process of the application, the programming languages I interfaced Java and C++ for the Android application, and Swift, Objective C, and C++ for iOS.

The C++ implementation is responsible for three basic functions: store proverbs, keywords, and their English equivalent in a hash table, return English equivalent as translation, prompt user to improve the translation and store recommendation in database for later use.

Swift and Java are used to implement natural language processing (NLP) libraries. The libraries are used in order to stem words. After removing stop words from the proverbs store in the database, the computer will treat the remaining words as keywords to generate keys for the hash table. This same process will occur during translation again to remove stop words from user input to identify keywords in the phrase. The keywords will be used to search the associated proverb. For example, if the user inputs “Las estrellas inclinan, pero no obligan” the computer will return “estrellas”, “inclinan”, “obligan” as keywords.

If a sentence contains three or more keywords, it will be identified as potentially containing a proverb. Each location where the keywords are stored in the hash table will have pointers to other keywords. The pointers are so that when identifying one keyword in a sentence, it can also do a check for other keywords that are associated with the keyword on which the computer performed the hashing operation. The hash table design avoids the importance of word order when identifying a proverb because regardless of which keyword is hashed first, the computer will still be able to search for the other keywords that are associated with the proverb.



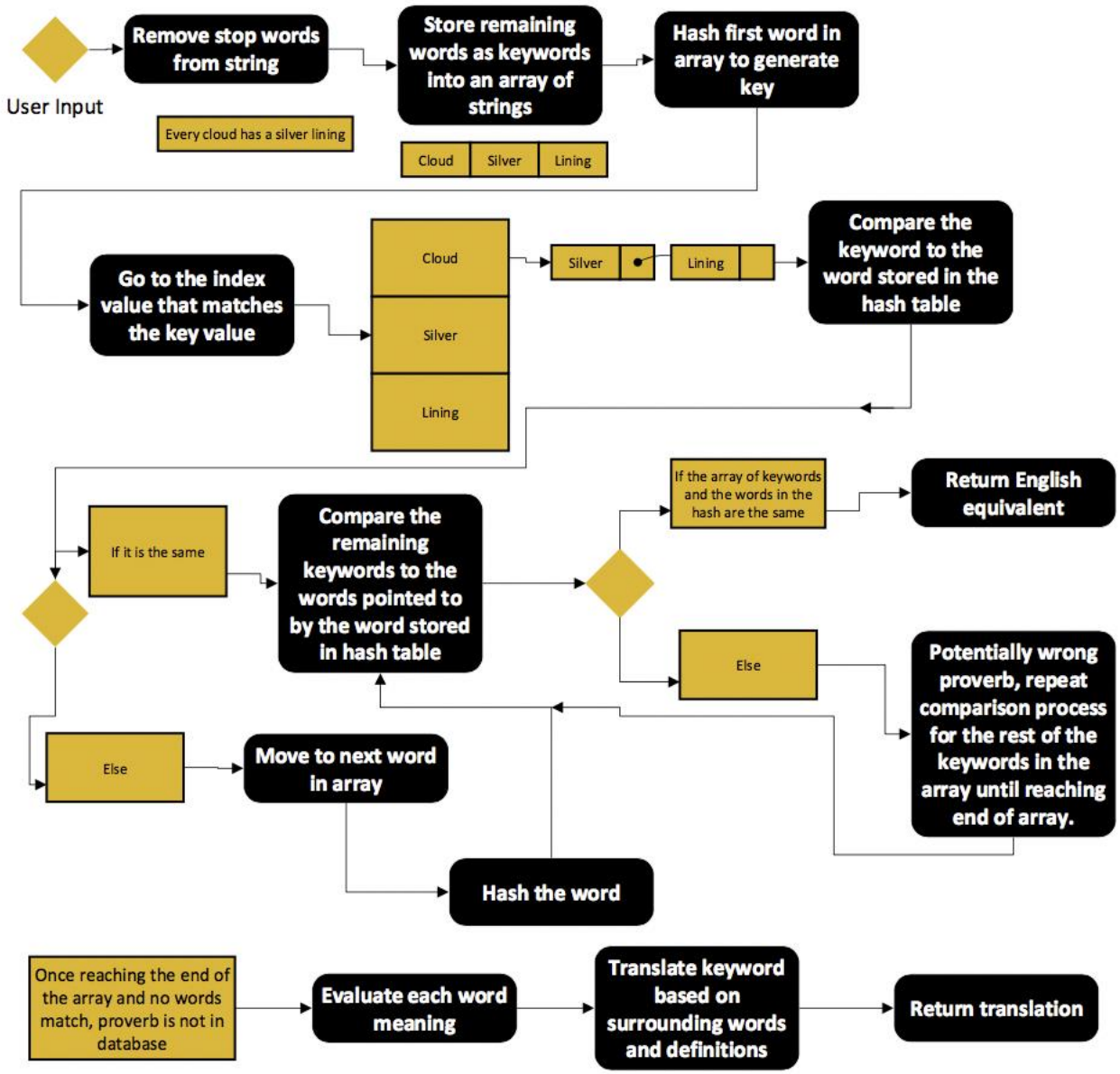
The above example illustrates how if upon hashing the word silver first instead of cloud in a

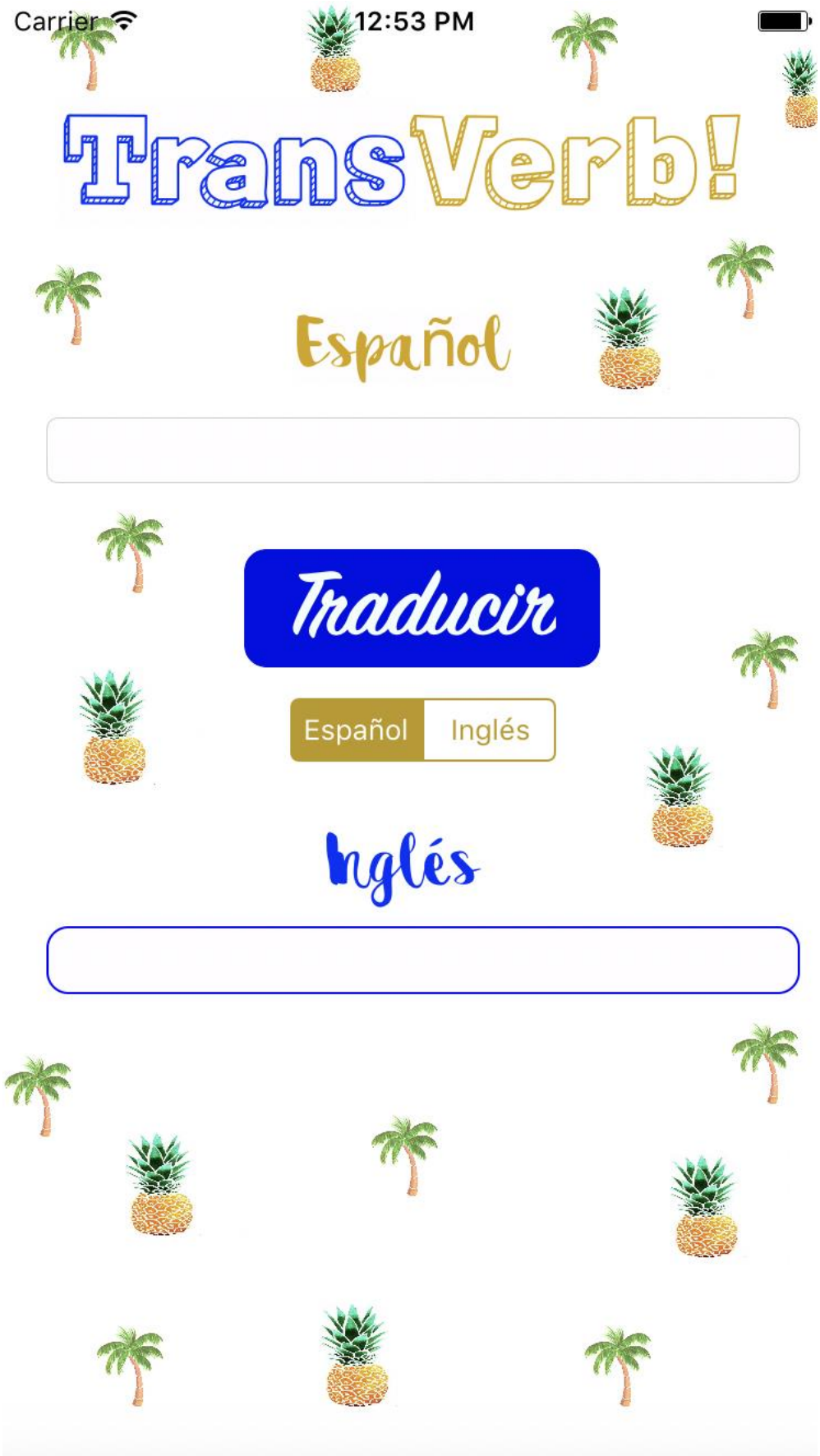
variation of the proverb “Every cloud has a silver lining” (i.e. “Every silver lining has a cloud.”) where the word order is different, the computer is still able to identify whether or not the sentence contains the other keywords associated with the proverb.

The hash table will be an array of structs that is able to hold a keyword, its associated keywords, the Spanish proverb, and the English equivalent. This so that upon finding a keyword, the computer can return the translation without having to perform an additional search through another hash table or array.

In C++, the function that prompts the user to improve the translation will store the new phrase in another hash table along with the Spanish proverb that was translated. After collecting a lot of responses, the translation will be improved using statistical NLP models in Python to decide which improvement phrase appear most.

On the subsequent pages, the first diagram illustrates the algorithm that will be responsible for the translation of the proverb. The next illustration is what the user interface will look for the app. The third is a results table comparing Google Translate and TransVerb.


User Input
Remove stop words from string
Store remaining words as keywords into an array of strings
Hash first word in array to generate key
Every cloud has a silver lining
Cloud
Silver
Lining
Go to the index value that matches the key value
Cloud
Silver
Lining
Silver
Lining
Compare the keyword to the word stored in the hash table
If it is the same
Compare the remaining keywords to the words pointed to by the word stored in hash table
If the array of keywords and the words in the hash are the same
Return English equivalent
Else
Potentially wrong proverb, repeat comparison process for the rest of the keywords in the array until reaching end of array.
Else
Move to next word in array
Hash the word
Once reaching the end of the array and no words match, proverb is not in database
Evaluate each word meaning
Translate keyword based on surrounding words and definitions
Return translation


Carrier
12:53 PM
TransVerb!
Español
Traducir
Español
Inglés
Inglés

Results

Phrase	Google Translate	TransVerb
Le dieron el gato por liebre.	They gave him a cat for a hare.	Mixing apples and oranges.
Nunca digas de esa agua no beberé.	Never say that water I will not drink.	Never say never.
Si no es Juan, es Pedro.	If it's not Juan, it's Pedro.	If it's not one thing, it's another.
Cuando tú vas, yo vuelvo.	When you go, I go back.	I get where you're coming from.
A mi plin y a la madama dulce de coco.	To my plin and to the sweet coconut madam.	I couldn't give a hoot.
Más hace el que quiere que el que puede.	That one who wants will make more that the one who can.	Genius is ten percent inspiration and ninety percent perspiration.
Más pelado que un chucho.	More skinned than a pooch.	Flat broke.

*The table above demonstrates the difference between Google Translate and TransVerb translations of colloquial Spanish phrase. Here, Google give back to the user literal translations while TransVerb give back the figurative meanings.

V. Conclusion

Demonstrated per the results section of the paper, the technique, phrase based SMT with the GCB approach used in TransVerb's design, gives more accurate translations of colloquial phrases. Terminology shapes how well an individual can express themselves. If an individual when speaking colloquially in a second language, uses wrong phraseology then what they are saying can easily be lost in translation. Vernacular language should not be translated word for word justifying the need for machine translation techniques that are used in TransVerb.

TransVerb, however, does not translate literal speech. It is designed to only translate sayings. Nevertheless, for future research, it could be developed to handle both literal and figurative speech.

References

[1] R. Baños, "'That is so cool': investigating the translation of adverbial intensifiers in English-Spanish dubbing through a parallel corpus of sitcoms," *Perspect. Stud. Transl.*, vol. 21, no. 4, pp. 526–542, 2013.

[2] M. R. Costa-Jussà, M. Farrús, J. B. Marino, and J. A. R. Fonollosa, "Study and comparison of rule-based and statistical catalan-spanish machine translation systems," *Comput. Informatics*, vol. 31, no. 2, pp. 245–270, 2012.

[3] Â. Costa, W. Ling, T. Luís, R. Correia, and L. Coheur, "A linguistically motivated taxonomy for Machine Translation error analysis," *Mach. Transl.*, vol. 29, no. 2, pp. 127–161, 2015.

[4] M. Estellés-Arguedas, "Expressing evidentiality through prosody? Prosodic voicing in reported speech in Spanish colloquial conversations," *J. Pragmat.*, vol. 85, pp. 138–154, 2015.

[5] L. Fernández Sánchez, "Learning Spanish sayings in the Spanish as a foreign language class," *J. Soc. Sci. Humanit.*, vol. 2, no. 2, pp. 861–876, 2015.

[6] F. G. Herrera and L. G. Luna, "Using translation paraphrases from trilingual corpora to improve phrase-based statistical machine translation: A preliminary report," *Proc. - 2007 6th Mex. Int. Conf. Artif. Intell. Spec. Sess. MICAI 2007*, pp. 163–172, 2008.

[7] A. M. Júnior and L. S. García, "WIKLANG – A DEFINITION ENVIRONMENT FOR MONOLINGUAL AND BILINGUAL DICTIONARIES TO SHALLOW-TRANSFER MACHINE TRANSLATION," pp. 159–168, 2010.

[8] P. Koehn, *Statistical machine translation*. .

[9] J. Lloret and F. Casacuberta, "Phrase-Based Statistical Machine Translation using Approximate Matching," pp. 1–8.

[10] H. Somers, "Review Article: Example-based Machine Translation," *Mach. Transl. Kluwer Acad. Publ. Print. Netherlands*, vol. 14, no. 2, pp. 113–157, 1999.