Team Members: Rupa Shankar, Jujhaar Singh

PA7 - Machine Translation from Spanish to English

About Spanish

We chose Spanish as our foreign language. Spanish has a two-gender noun system and is primarily an SVO (Subject Verb Object) language. However, it is quite common to leave out the subject entirely if the subject can be deduced from the sentence's context or the verb conjugation e.g "Vi a Rupa" means "I see Rupa," but the direct translation is "See Rupa." That is particularly confusing when translating from Spanish to English because the translator has to infer the subject. Another exception to the SVO structure is that when pronouns are used as objects, Spanish uses the SOV (Subject Object Verb) order e.g "Rupa lo vi" means "Rupa see him," but the direct translation is "Rupa him see."

In addition, Spanish is an inflectional language; the forms of its words vary depending on their relationship to other words in a sentence. For example, Spanish verbs have multiple conjugated forms depending on the subject and the time frame, and adjectives change depending on the gender and quantity of the noun. When translating to English, this language feature poses a challenge because one has to distinguish between these different forms.

The Original Test Document

En un futuro no muy distante, recibirás un diagnóstico y cura completa desde tu smartphone, incluso antes de notar que estás enfermo.
Aunque esto parece ciencia ficción, está a punto de volverse una realidad.
La tecnología digital está lista para transformar radicalmente al sector de la salud y el bienestar.
En el camino, nos ayudará a superar algunos de los retos más significativos que enfrentamos.
A medida que las personas mayores representan una mayor parte de la población, la prevalencia de problemas de salud de largo plazo aumentará.
Esto causará una mayor carga de costos y presionará a los sistemas de salud a acomodar a una mano de obra que envejece.
Por otra parte, los problemas de salud crónicos relacionados con el estilo de vida, incluidos la obesidad y la diabetes, aumentan con implicaciones dramáticas para los presupuestos de los servicios de salud.
El costo de mantener estas tendencias demográficas es insostenible, pero los servicios digitales podrían ser parte de la solución que busca la sociedad.
Una tendencia que ha capturado la imaginación de muchos es el "hackeo del cuerpo" o el entendimiento del "ser cuantificado".
Ya sea un laboratorio que analice tu genoma, una aplicación que rastree tu ingesta alimentaria o una banda portátil que registre la actividad física que haces todos los días, estos dispositivos te ofrecen las herramientas para entender inmediatamente tu salud con base en los datos que

proporciona tu cuerpo.
Muchos moldes de letras para imprimir que te servirán para hacer carteles o utilizar en logos, letreros o cualquier otro motivo para utilizar moldes de letras.
Nicolás Maduro critica a Israel y a la canalla mediática que lo apoya.
Cuba de duelo por el Chávez que acompañó a Fidel "como un hijo verdadero"

Our System Output

not very far in a future will receive a diagnosis and priest complete from your smartphone even before of notice that be sick
although this appear science fiction be at come dot back a reality
the technology finger be list for transform radically to sector of the health and the well-being
in the route help us at lead some of the challenge more meaningful that put face to face
at measure that the persons ancestors represent an older part of the settlement the health problem prevalence of long period grow
this cause an older cost freight and put pressure on at the health system at seat at a work hand that grow older
for other part the health chronic related problem with the life style include the obesity and the diabetes grow with dramatic implications for the budget of the health service
the maintain cost be demographic tendency is unstable but the service finger be able to being part of the solution that search the society
a tendency that capture the imagination of many is the body hack or the understanding of being quantified
now being a laboratory that analyze your genome an application that track your consumption feed or a portable band that register the active state physics that do all the days these regulatory offer you the tool for understanding immediately your health with basis now immediately in that provide your body
many letter molds for print that serve you as for do posters or use in logo signs or any other reason for use letter molds
Nicolás Maduro criticism Israel and at the despicable media that which is support
Cuba of duel for Chavez that accompany Fidel like a son true

Our Rules

1. NOUN1 ‘of’ NOUN2 → NOUN2 NOUN1

One way of creating a compound noun in Spanish is to put two nouns together using some form of the word ‘de,’ which means ‘of.’ For example, “problemas de salud” literally translates to “problems of health,” though it simply means “health problems.” However, in English, compound nouns are created by just putting the two parts together without anything in between. The first part is typically a descriptor and the second part usually identifies the object itself e.g “health sector.” Thus in order to fix this, we decided to rearrange non-proper nouns that have the word

‘of’ in between, so in our text, “hack of body” became “body hack.” Although this rule works for a lot of cases, there are some instances for which it does not work, such as “pocket of sunshine”, “plate of food”, or “bale of hay”. When NOUN2 refers to what’s contained in NOUN1, the of stays.

2. ‘be’ VERB → VERB

In Spanish, there are many verb conjugations that use the verb “hacer,” or “to be.” Some examples of these conjugations, translated to English and using the verb “to capture,” are “is capturing” and “was captured.” However, when doing a one-to-one translation, it is very difficult to accurately translate the conjugation correctly without context of the subject. Therefore, a phrase such as “Una tendencia que ha capturado”, which actually means “A tendency that has captured,” literally translates to “A tendency that be capture.” In order to improve the fluency of sentences like these, we wrote a rule to remove the word “be” when it occurs before a verb. Thus “a tendency that be capture” becomes “a tendency that capture.” We thought that the slight reduction in faithfulness is offset by the gain in fluency.

3. NOUN ADJECTIVE → ADJECTIVE NOUN

In Spanish, adjectives typically come after the noun that they describe, while in English, adjectives come after the noun. This rule fixes this order by switching nouns and adjectives e.g “implications dramatic” becomes “dramatic implications.” It works for the majority of cases, but would not for a few Spanish adjectives that do come before the noun, such as “primera” (first).

4. ‘in’ WORD+ ADVERB+ → ADVERB+ ‘in’ WORD+

An adverb clause is usually introduced by a subordinating conjunction. In Spanish, the description of a particular noun or verb usually comes afterwards, not before. To fix this, we wrote a rule that moves the adverb clause before the subordinating conjunction and its subsequent clause. For example, looking at the phrase “in a future not very far,” the subordinating conjunction is “in,” the following clause is “a future,” and the adverb clause is “not very far.” Based on our rule, this phrase becomes “not very far in a future.” We initially tested this rule with any subordinating conjunction, but that did not work well as a heuristic, so we narrowed the focus to just “in.”

5. VERB ‘at’ PROPER_NOUN → VERB PROPER_NOUN

The word “a,” in Spanish, has a number of different meanings. On one hand, it is a preposition whose primary meaning is “at.” Alternatively, when used in between a verb and a proper noun, it is usually a filler word that does not align with an English word in the actual correct translation. Therefore, we wrote a rule to remove “at” when in between a verb and a proper noun. For example, “Nicolás Maduro criticism at Israel” (which should be “Nicolás Maduro criticizes at Israel,” but for the multiple definitions of “critica”) becomes “Nicolás Maduro criticism Israel” (or

“Nicolás Maduro criticizes Israel”). For our actual implementation, we ended up removing the restriction that the previous word must be a verb because our POS tagger often mistagged verbs as nouns, which might lead to mistranslations in other contexts. However, this is the logic upon which our rule is based.

6. ARTICLE PROPER_NOUN → PROPER_NOUN

A common practice in Spanish is to refer to proper nouns, usually people or places, using a definite article e.g “por el Chavez” literally translates to “for the Chavez,” which is unwieldy in English. So, we decided to remove the definite article before a proper noun, transforming “for the Chavez” to “for Chavez,” which is a much more fluent English phrase. Although there are situations for which this rule does not work (e.g we would want to keep the “the” in “the United States”), the net gain in faithfulness and fluency is greater than the net loss.

7. Fixing a/an

For a particular noun in Spanish, which singular indefinite article to use depends on the noun’s gender whereas in English, the choice of which singular indefinite article to use depends on whether or not the noun begins with a vowel. However, “una” and “un”, Spanish’s singular indefinite articles, both directly translate to “a,” which sometimes results in grammatically incorrect phrases such as “a application.” To fix this error, we implemented a rule that checks each instance of “a” or “an,” its corresponding noun, and corrects the indefinite article if necessary. With this rule, “a application” becomes “an application.”

8. Removing consecutive words that are the same

Sometimes, the output of a direct translation accidentally contains consecutive words that are the same, such as “for for.” Exactly why this occurs depends on the specific text and context. One possible situation is that for a particular verb, the preposition that follows it is included in the dictionary entry but then is also included in the actual text e.g “is used for for.” Most of the time, the second instance of the same word is extraneous and should be removed. Although we do not have any consecutive words that are the same in our text, we implemented this rule anyways because it is a situation that we have both often encountered while translating.

9. VERB ADVERB → ADVERB VERB

In Spanish, adverbs often come after the verb that they describe, while in English, adverbs tend to come before the verb. Therefore, we wrote a rule that swaps verbs and adverbs. For example, “understanding immediately” should become “immediately understanding.” In general, based on Spanish grammar, this rule is a good one. However, this rule did not actually have any effect on our text because our POS tagger often mistagged verbs as nouns e.g “understanding” was tagged as a noun although in context, it was definitely a verb.

10. PERSONAL_PRONOUN VERB → VERB PERSONAL_PRONOUN - reflexive verbs

Reflexive verbs in Spanish are common, and usually refer to the object that is having the verb done to it. For example, lavarse is the verb to bathe, and when used in context, it says who is bathing who: “me lavo” means I wash myself. This is different from English, because we usually say the verb and then the object, “I pushed Sally”, “she loves chocolate”, etc. This rule remedies the difference by switching the order of the words to reflect the standard in English.

Error Analysis of Our System

The errors we ran into while translating fell into four categories: those caused by a bad dictionary, those caused by POS tagging, those caused by differences in Spanish and English that we weren’t able to account for, and those caused by oversimplified rules.

The first place our translation had some errors was in the dictionary. Looking up the words was solely based on the first translation that showed up in WordReference, so some words that could’ve been translated better in context had glaring problems. For example, ‘cura’ was translated to ‘priest’, even though it was obvious that the intended meaning was ‘cure’. Another example was ‘digital’, which was translated to ‘finger’ instead of the correct translation of ‘technology’, and ‘duelo’, which was translated to ‘duel’ instead of ‘hurt’. All in all, we can’t do much about the dictionary unless we take into account context or use our own knowledge of the context to write the word meanings.

The second category of errors was caused by bad POS tagging (we used the NLTK POS tagger), and generally made our rules less efficient, or included words that we did not want to be affected by the rule. For example, one of the rules we tried to implement was replacing ‘of’ + a verb with the “ing” form of the verb (e.g. of notice → noticing). Unfortunately, the POS tagger gave us some difficulty - many words that were verbs in context were marked as nouns (e.g. notice). What we then decided to do was ignore the original part of speech of the word, and instead add “ing” to the end of the word, and check if that was a verb. Unfortunately, adding “ing” apparently makes lots of things verbs. We had ‘beinging’, ‘healthing’, and ‘lifing’ as verbs, and some of the nouns that we didn’t want to change became verbs (e.g. long → longing). Overall, if the POS tagger was a little better, we could’ve used this rule, since it applied in many cases.

A second effect of a variable POS tagger was that some of our rules were diluted. For example, rule #9 for switching adverbs and verbs had a few examples that weren’t picked up: “transform radically”, for example, didn’t make the cut because “transform” was marked as a noun by the POS tagger.

Besides the problems mentioned above, there are several rules and text parts that we didn’t get right. Some examples of rule based text that we didn’t get a chance to change are:
- “sector of the health and the well-being” (Google: the health sector and welfare)
- “put face to face” (Google: we face)

- “a work hand that grow older” (Google: aging workforce)
And a lot of other awkward sentences. These fixes would be much more complex, and finding rules to solve them would be hard.

The last subset of problems were those caused by exceptions to our rules, which was a bit too simplistic in some cases. For example, the first rule (NOUN1 of NOUN2 → NOUN2 NOUN1) doesn’t apply in cases where noun1 contains noun2 (e.g. bale of hay, plate of food, pocket of sunshine). Another overly basic rule was “be VERB” → “VERB”, because it fails to take into account tense - I could “be running” or Jamie “has been running”, and in those cases our code would just drop the first form of be.

Output of Google Translate:

In a not too distant future, you will receive a diagnosis and complete cure from your smartphone before you even realize you are sick.
While this sounds like science fiction, is about to become a reality.
Digital technology is poised to radically transform the health sector and welfare.
On the road, help us overcome some of the most significant challenges we face.
As the elderly represent a larger portion of the population, the prevalence of health problems long term increase.
This will cause a greater burden of costs and pressure on health systems to accommodate an aging workforce.
Moreover, chronic health problems related to lifestyle, including obesity and diabetes increase with dramatic implications for the budgets of health services.
The cost of maintaining these demographic trends is unsustainable, but digital services could be part of the solution for society.
A trend that has captured the imagination of many is the "body hacking" or understanding of "be quantified".
Whether a laboratory to analyze your genome, an application that tracks your food intake or a portable band record the physical activity you do every day, these devices give you the tools to understand your health immediately based on the data provided by your body.
Many molds to print letters that will serve to make posters or used in logos, signs or any other reason for using molds of letters.
Nicolas Maduro criticized Israel and rogue media that supports it.
Cuba mourning the accompanying Chavez Fidel "as a true son"

Comparison Analysis for Google Translate vs Our System

Our system, understandably, seems to be a worse translator than Google, so there are many places where Google Translate does a better job. However, there are also several places where our code ends up giving the same result, which is a good sign! There is also at least two places where our translation is better than Google’s!

Where Google Translate Wins:
Google Translate seems to do well with some of the areas of text mentioned above in the error analysis, mostly because we haven't implemented as many features as they have. Specifically, they end up getting rid of a lot of awkward phrasing with too many prepositions and/or articles, and reordering and consolidating different phrases. Google Translate also seems to have better rules. One example is when Spanish has the word "be" + verb in it. Currently, our system just takes out the be, but Google Translate figures out how to convert the phrase into a present perfect one (e.g. "be capture" in our system is "capture" and in Google is "has captured").

Where Google Translate and Our Code Tie:
Our system's rules seems to match up to Google's pretty well. Here are examples of each rule where both Google and we got around the same result:

Format(
rule#. explanation:
"original output" → "our output" ("Google output")
*note: if necessary
)

1. NOUN1 'of' NOUN2 → NOUN2 NOUN1:
"style of life" → "life style" ("lifestyle")

2. 'be' VERB → VERB:
"be capture" → "capture" (" has captured")

3. NOUN ADJECTIVE → ADJECTIVE NOUN:
"implications dramatic" → "dramatic implications" ("dramatic implications")

4. 'in' WORD+ ADVERB+ → ADVERB+ 'in' WORD+:
"in a future not very far" → "not very far in a future" ("In a not too distant future")

5. 'at' PROPER_NOUN → PROPER_NOUN:
"criticism at Israel" → "criticism Israel" ("criticized Israel")

6. ARTICLE PROPER_NOUN → PROPER_NOUN:
"the Chavez" → "Chavez" ("Chavez")

7. Fixing a/an:
"a application" → "an application" ("an application")

8. Removing consecutive words that are the same*:
"for for" → "for" ("for")
*note: this rule didn't occur after we changed the dictionary, so this is an older case.

9. VERB ADVERB → ADVERB VERB*:
“transform radically” → “radically transform” (“radically transform”)
*note: this is dependent on how accurate our POS tagger is.

10. PERSONAL_PRONOUN VERB → VERB PERSONAL_PRONOUN - reflexive verbs
“you serve” → “serve you” (“will serve”)

Where Our Code Wins:
There are two main place where our code is more clear than Google’s. The first is in the last line, which we translate as:

> “for Chavez that accompany at Fidel like a son true”

and Google Translates as:

> “the accompanying Chavez Fidel "as a true son" “

Our translation is better because it differentiates Fidel and Chavez. It is more readable, since Google has two proper nouns right next to one another. It also doesn’t have an extraneous “the”.

The second translation that we performed better on than Google is “being quantified” (they had “be quantified”) because it makes much more sense in the context.

Team Responsibilities

Both: got the text, parsed it, made the dictionary
Individually:
Rupa - 7 rules, Jujhaar 3 rules
Rupa had first half of write up (language and rules), Jujhaar had second half of write up (error analysis, comparison to Google)