Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 53-64 recibido 25-11-2020 revisado 25-01-2021 aceptado 03-02-2021

Using Dependency-Based Contextualization for transferring Passive Constructions from English to Spanish

Contextualización basada en dependencias sintácticas para transferir construcciones pasivas de inglés a español

Pablo Gamallo,[1] **Gorka Labaka**[2]
[1]Centro de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2]IXA, Universidad del País Vasco (UPV/EHU)
pablo.gamallo@usc.gal; gorka.labaka@ehu.eus

Abstract: We hypothesize that parallel corpora as well as machine translation outputs contain many literal translations that are the result of transferring the constructions of the source language to the target language. When translating passive expressions from English to Spanish, there are several constructions available, however, both automatic and human (if of low quality) translations tend to select the periphrastic structure, which is the literal construction. The objective of this article is to make use of strategies trained on monolingual corpora to translate English passive expressions into Spanish so as to verify whether unsupervised translation with monolingual corpora benefits syntactic diversity. Special attention will be given to the monolingual-based strategy relying on dependency-based contextualization. The results of the experiments carried out show that the methods relying on monolingual corpora tend to offer more non-literal translations (middle-voice) than those trained on parallel corpora.

Keywords: Semantic Contextualization, Similarity, Unsupervised Machine Translation, Passive Voice.

Resumen: Formulamos la hipótesis de que los corpus paralelos así como los resultados de la traducción automática contienen muchas traducciones literales que son el resultado de la transferencia de las construcciones del idioma de origen al idioma de destino. Cuando se traducen expresiones pasivas del inglés al español, hay varias construcciones disponibles, sin embargo, tanto las traducciones automáticas como las humanas (si son de baja calidad) tienden a seleccionar la estructura perifásica, que es la construcción literal. El objetivo de este artículo es hacer uso de estrategias entrenadas a partir de corpus monolingües para traducir las expresiones pasivas del inglés al español, a fin de verificar si la traducción no supervisada con corpus monolingües beneficia la diversidad sintáctica. Se prestará especial atención al método semántico que se apoya en el proceso de contextualización en el marco de la sintaxes de dependencias. Los resultados obtenidos en los experimentos muestran que los métodos basados en corpus monolingües tienden a generar más traducciones no literales (voz media) que los entrenados con corpus paralelos.

Palabras clave: Contextualización semántica, Similaridad, Traducción automática no supervisada, Voz pasiva.

1 *Introduction*

Passive voice is a type of linguistic construction shared by most world's languages (Keenan, 1985). It is the result of a detransitivizing process that reduces the verbal valence promoting the object (or a similar function) to the subject position, which becomes the topic of an depersonalized sentence. Languages may differ with regard to formal codification of passive constructions, such as word order, case assignment, or verb morphology, but two elements are fundamental from a cross-linguistic point of view: the existence of an active counterpart and the prominence of a non-agent participant in the syntactic

ISSN 1989-7553. DOI 10.26342/2021-66-4

encoding of the passive clause (Siewierska, 1984).

There are remarkable differences in the passivization process between English and Spanish. Whereas in English passive constructions are mainly encoded by means of the periphrastic passive (*be+pp*), in Spanish there are several ways of encoding passive constructions, including the periphrastic strategy but also middle voice constructions containing the non-anaphoric pronoun *se*. Linguistic studies claim that periphrastic passive is less frequent in Spanish than in English as Spanish tends to use middle-voice constructions with *se* (Rodríguez-Vergara, 2017). In addition to that, it is important to take into account literal translation, which is a translation strategy that follows closely the surface form of the source language. By producing a structurally close translation of the source text, the translation process reaches the final result in a very efficient and fast way, and with the minimal processing effort. For this reason, literal translation is by far the most frequent method in all translation types (Scarpa, 2020). The well-known translation model defined by Vinay and Darbelnet (1995) considers that literal translation is the most suitable strategy if it does not modify the meaning of the source text or is not possible because it changes the structure of the source language. According to these ideas on literal translation, we assume that the periphrastic Spanish passive would be the most used translation of the corresponding English construction, leaving out the other non-literal alternative, i.e., middle-voice, even though this construction could be even more appropriate than the periphrastic one in many situations. Considering these facts, we propose the following hypothesis:

> **Periphrastic Passive Bias in English-Spanish Translation:** The Spanish periphrastic passive is over-represented in English-Spanish parallel corpora as it corresponds to the literal translation, which leads to the fact that English-Spanish automatic translations based on parallel corpora (supervised machine translation) tend to massively generate this structure in the Spanish texts; by contrast, the alternative passive constructions (middle-voice), which are very common structures in all Spanish domains and genres, are scarce in parallel corpora and then underused by supervised machine translation.

To check if the hypothesis is true, we propose using unsupervised methods trained on monolingual corpora (non-parallel texts) in both English and Spanish. As middle-voice constructions are more natural and frequent in untranslated texts, they are supposed to emerge more frequently with strategies relying on monolingual corpora than with supervised methods based on parallel corpora. So far, the main reason for abandoning dependence on parallel corpora is that they are scarce especially in the case of low-resource languages. However, we now propose a linguistic reason based on the above-mentioned linguistic bias. If it is verified, we can admit that there is a general bias towards translations that are too literal (especially if the translator is not professional, is poorly paid or has little time), and this has repercussions on the quality and structural diversity of the results of machine translation.

This hypothesis is in accordance with recent findings (Vanmassenhove, Shterionov, and Way, 2019; Toral, 2019), which conclude that machine translation gives rise to lower lexical diversity than human translations. In addition, machine and human translations have lower lexical diversity than monolingual texts that are naturally written in the target language.

In the present work, we will make use of strategies trained on monolingual corpora to translate English passive expressions into Spanish. In addition to unsupervised machine translation techniques (Artetxe et al., 2018; Artetxe, Labaka, and Agirre, 2019), special attention will be paid to a dependency-based approach to perform contextualized translations from monolingual corpora (Gamallo et al., 2019). We will configure a system that follows this approach in order to select the most appropriate passive construction in Spanish given an English periphrastic passive expression. The specific objective is to permit the dependency-based strategy to transfer passive constructions from one language to another.

The rest of the article is organized as follows. In the next section (Sec. 2), we deal with passive constructions in English and, above all, in Spanish from a linguistic point of view. Section 3 introduces two unsupervised approaches to translation from monolingual corpora, paying special attention to the strategy of syntactic contextualization (Subsection 3.2). Section 4 applies the latter strategy to the transfer of passive constructions from English to Spanish. Experiments are reported in Section 5 and final conclusions are addressed in Section 6.

2 *Passivization in English and Spanish*

Semantically, the passive construction implies, on the one hand, defocalization of the agent, which is encoded in oblique case or even suppressed from the sentence and, on the other hand, topicalization of other participant, which can be the patient, experiencer, theme, beneficiary, etc. So, passivization can be seen as the shift in focus from the agent (or similar role) to a non-agentive participant (usually the patient) in an event. Siewierska (1984) points out that the necessity for the use of the passive varies from language to language, and the differences are not only in the way the constructions are syntactically encoded, but also in how they express semantic-discursive functions related to topicalization, impersonalization and detransitivisation.

2.1 Types of Passive Constructions in English and Spanish

From the syntactic point of view, English and Spanish share the periphrastic passive which consist of transforming the active verb into the periphrasis *be/ser + past participle*. Semantically, the periphrastic passive construction in both languages allows the conceptualization of the agent, although it remains a marginal conceptualization as it can only be encoded with an optional oblique case (*by/por + agent*) (Fernández, 2007). Table 1 shows some examples of expressions encoded with the periphrastic passive in English and Spanish. All English periphrastic expressions can be translated into Spanish using the same construction. Even though the agent can be expressed in all of them, only the first example contains an explicit agent coded with the oblique case: *by the Jesuits/por los Jesuitas*. In the other three periphrastic expressions, it has not been expressed as it is optional.

As Table 1 also shows (see the two columns on the right), there are two other types of passive constructions in Spanish that do not exist in English: both reflexive and impersonal passives, which are usually called *middle-voice* in opposition to (periphrastic) passive voice. Syntactically, they are constructed with the active verb along with the insertion of the non-anaphoric pronoun *se*. They differ in the way of encoding the non-agentive topicalized participant. In the reflexive construction the topicalized participant is the subject of the clause (there is agreement with the verb), whereas in the impersonal one it is encoded as direct object preceded by the preposition *a*. The two middle-voice constructions are semantically very similar but they are in complementary distribution in some cases: when the topicalized participant is inanimate, the preferred encoding is the reflexive passive. By contrast, if the topicalized participant is animate or human (and then a potential agent), the preferred encoding is the impersonal construction because in those cases reflexive passives are formally very close to active reflexive/reciprocal clauses. For instance, the translation of *The workers were threatened* into a middle-voice construction must be the impersonal passive *Se amenazó a los trabajadores* since the reflexive passive (*Se amenazaron los trabajadores*) may be confused with its corresponding active clause with a reflexive/reciprocal meaning: *the workers threatened each other*. As the non-agentive participant, *the workers/los trabajadores*, are human beings, they can be interpreted as agents and patients at the same time giving rise to the active and reciprocal construction. Thus, to prevent the confusion with active reflexive/reciprocal clauses, the reflexive passive is not allowed with this type of agentive participants. In fact, in those cases it is very common in Spanish to use a hybrid (but ungrammatical) structure that mixes both reflexive and impersonal passives: (**Se amenazaron a los trabajadores*) (Sánchez-López, 2002).

Semantically, unlike the periphrastic passive, the two middle-voice constructions prevent the conceptualization of the agent of the active clause from which it derives. In the first example of Table 1, the agent *by the Jesuits/por los Jesuitas* cannot be syn-

	periphrastic passive	**reflexive passive**	**impersonal passive**
en	*The church was founded in 1850 by the Jesuits*	-	-
spa	*La iglesia fue fundada por los Jesuitas*	**La iglesia se fundó por los Jesuitas*	**Se fundó a la iglesia por los Jesuitas*
en	*The church was founded in 1850*	-	-
spa	*La iglesia fue fundada en 1850*	*La iglesia se fundó en 1850*	**Se fundó a la iglesia en 1850*
en	*The treaty was signed in Lisbon*	-	-
spa	*El tratado fue firmado en Lisboa*	*El tratado se firmó en Lisboa*	**Se firmó al tratado en Lisboa*
en	*The workers were threatened*	-	-
spa	*Los trabajadores fueron amenazados*	*?Se amenazaron los trabajadores*	*Se amenazó a los trabajadores*

Table 1: Passive English sentences and their Spanish translations with different passive constructions: periphrastic, reflexive and impersonal.

tactically encoded in the reflexive passive as it is not semantically conceptualized within the depersonalized scene designed by the verb (Fernández, 2007).

Periphrastic and middle-voice constructions in Spanish are not in complementary distribution. In many cases, both options are allowed to translate the same passive in English (if this one does not contain an explicit oblique agent). The second, third and fourth examples in Table 1 are encoded in Spanish in at least two constructions: periphrastic and either reflexive or impersonal passive. However, there are aspectual and lexical restrictions that tend to favor one construction or another. Periphrastic passive tends to be used with verbs expressing singular events with an external object and agent. By contrast, the use of verbs with a habitual, repetitive (iterative) or generic lexical aspect favors middle-voice passives (de Miguel, 1999). It was also found that the middle-voice is used more with material and relational events (77% of cases) than with mental, existential and behavioural processes (Lourdes Díaz Blanca, 2008).

There are serious problems to carry out quantitative studies to compare the use of periphrastic and middle-voice constructions by using automatic approaches. It is not possible to automatically identify passivizations with the non-anaphoric pronoun *se* as this pronoun also co-occurs with verbs in the active form to build many other syntactic constructions as it is reported in García-Miguel (1985). Alarcos (1978) counted up to 9 types of uses of *se*.

2.2 Frequency of Use of Different Passivization Types in Spanish and English

The experiments reported in Jisa et al. (2002) show that the periphrastic passive constructions are used significantly more in Dutch, English, and French than in Hebrew or Spanish. In fact, Spanish shows very little reliance on this type of construction across narrative and expository texts. However, this does not mean that passivization (either with peripheral or middle-voice constructions) is not as common in Spanish as in English.

A recent study analyzed and counted the number of different constructions found in a parallel English-Spanish text (Rodríguez-Vergara, 2017). The parallel text consists of an scientific article on the medical field written in English and its translation into Spanish. The authors found 52 periphrastic passive constructions in English and 48 passive translations in Spanish, 15 being periphrastic and 33 being reflexive/impersonal (i.e., middle-voice). Despite the small size of the corpus, this study shows two trends: (i) most of the English passive constructions are translated into passivized structures in Spanish (rather than in active constructions), and (ii) middle-voice constructions are more common in Spanish than periphrastic ones. This trend should be much more pronounced in other genres and domains: literary texts, informal language, etc.

To the best of our knowledge, there are no NLP-based studies on the use of passivization in English and Spanish.

3 Machine Translation from Monolingual Corpora

In the Introduction, we have claimed that there is a periphrastic passive bias in English-Spanish translation. This bias consists of the fact that Spanish periphrastic passive is overrepresented in English-Spanish parallel corpora at the expense of the other passivizations because of the influence of the source language (English) in the translation process. This also leads to MT systems based on supervised training (i.e. parallel corpora) producing a bias in favour of the periphrastic structure in their results.

To check whether this claim tends to be true, we will carry out experiments with translation strategies trained on monolingual corpora, which are not biased in favor of one type of construction, but represent natural text without the influence of the source language.

We distinguish between two types of monolingual-based strategies: unsupervised machine translation and dependency-based contextualized translation. In this section, we will briefly explain these two strategies. In the next one, we will propose an improvement for the second that will allow us to apply it to transfer passive constructions from one language to another.

3.1 Unsupervised Machine Translation

Unsupervised MT was born to minimize the dependency on parallel data by training machine translation systems using only monolingual corpora. These strategies started with neural sequence-to-sequence systems which consist of training a dual model combining back-translation and denoising autoencoding (Artetxe et al., 2018; Lample et al., 2018a). The training process is initialized with cross-lingual embeddings, which can be also generated using an entirely unsupervised method by automatically learning the mapping between two vector spaces without the support of an external bilingual dictionary (Artetxe, Labaka, and Agirre, 2018a).

These neural-based systems were recently overtaken by a more traditional phrase-based statistical MT approach also provided with an unsupervised strategy (Lample et al., 2018b; Artetxe, Labaka, and Agirre, 2018b). The new approach leverages the modular architecture of statistical MT: a phrase table is induced through cross-lingual embedding mapped from monolingual corpora, this table is combined with a n-gram language model, and the system is improved through iterative back-translation.

More recently, Artetxe et al. (2019) described a hybrid approach with state-of-the-art results for unsupervised MT. It is based on the above-mentioned statistical MT approach reported in Artetxe (2018b), which is used to initialize an unsupervised neural system improved through on-the-fly back-translation.

3.2 Dependency-Based Contextualized Translation

In a recent work, Gamallo et al. (2019) describe a compositional distributional method to generate contextualized senses of words in a cross-lingual space that is aimed at selecting the most appropriate translations in the target language using monolingual corpora. It is a dependency-based strategy inspired on previous work reported in Erk and Padò (2008; 2010) and applied to a cross-lingual space. The dependency-based translation strategy consists of the following steps:

(1) Generation of candidates: The source expression is syntactically analyzed using syntactic dependencies and the resulting construction is associated with a set of candidate translations in the target language with an equivalent syntactic construction. An example will help us illustrate this. The English phrase *catch a ball* is syntactically analyzed as a direct object dependency: $< obj, catch, ball >$. Then, by making use of a bilingual dictionary, each English constituent term is expanded with their Spanish translations giving rise to a list of candidate expressions encoded with the same construction. For instance, if the Spanish translations of *catch* found in the dictionary are the verbs *coger* and *contraer*, and those of *ball* are the nouns *pelota* and *baile*, then, we can generate four possible Spanish combinations (see Table 2).

Even though four candidates were generated, only the first one is the correct Spanish translation of the source expression. The other cases are weird combinations produced by the polysemy of the words constituting the English expression. We must point out that this is a toy example as the number of equivalent translations per word has been reduced

$< obj, coger, pelota >$	($catch\ a\ ball$)
$< obj, coger, baile >$	($*catch\ a\ dance$)
$< obj, contraer, pelota >$	($*contract\ a\ ball$)
$< obj, contraer, baile >$	($*contract\ a\ dance$)

Table 2: Spanish translation candidates for the dependency $< obj, catch, ball >$.

to facilitate explanation. Using real bilingual dictionaries, *catch* may have up to ten different Spanish translations and *ball* five, giving rise to 50 (10x5) Spanish candidates. In a related paper by the same authors (Gamallo and Garcia, 2019), they used a totally unsupervised strategy. Instead of external bilingual dictionaries, they made use of unsupervised learned cross-lingual embeddings to generate the target language candidates. The next two steps are designed to automatically select the correct translation(s) from the generated candidates.

(2) Contextualized senses: Once the candidates in the target language have been generated, the next step is to build the distributional meaning representation of both the source expression and translation candidates. The distributional meaning of each expression stands for the contextualized senses of its constituent words (Gamallo, 2019). Let us continue with the previous example. The meaning of *catch a ball* consists of two contextualized senses: the sense of the verb *catch* given the noun *ball* in the direct object position, and the sense of *ball* as direct object of *catch*. Each contextualized sense is built in two sequential processes. In the case of *catch*, the first process is to build the selectional preferences of *ball*. Intuitively, they correspond to the most relevant verbs that can be combined with the noun *ball* in the direct object position. Formally, they are defined by the vector resulting of adding the vectors of those relevant verbs. The second process consists of combining the vector of *catch* with the resulting vector representing the selectional restrictions imposed by *ball*. This combination, implemented by means of vector multiplication, represents the contextualized sense of *catch*. Similar processes are carried out with the other word *ball*. The final meaning of the expression are thus two contextualized vectors, noted $\mathbf{catch}_{obj\uparrow}$ and $\mathbf{ball}_{obj\downarrow}$, where $_{obj\uparrow}$ and $_{obj\downarrow}$ represent the *head* and *dependent* roles of *catch* and *ball*, respectively, in the direct object relation. Sense elaboration is thus the result of bi-directional operations: the head word restricts the sense of the dependent one in the same way as the latter restricts the head. The same contextualization process is applied to all the Spanish candidates so as to create their corresponding contextualized vectors. We should note here that there is a great conceptual parallelism between this contextual strategy and the recent Transformers models based on bidirectional contextualized word embeddings (Devlin et al., 2019). However, while the latter are mainly based on syntagmatic relationships (word co-occurrences in context), the former mainly relies on paradigmatic relationships established between optional words potentially occurring in the same syntactic functions.

(3) Selection by Similarity: Finally, the distributional meanings (defined as contextualized senses) of the generated candidates are compared pairwise by means of cosine similarity with the English sentence. The generated Spanish sentence associated with the most similar meaning is selected as the best Spanish translation of the English sentence. More precisely, given a specific dependency $< obj, catch, ball >$ in the source language, its contextualized translation, CT, in the target language is computed as follows:

$$CT(< obj, catch, ball >) = \quad (1)$$
$$\underset{< obj, w_1, w_2 > \in \phi}{\arg\max} \frac{1}{2} S(\mathbf{catch}_{obj\uparrow}, \mathbf{w_1}_{obj\uparrow}) + S(\mathbf{ball}_{obj\downarrow}, \mathbf{w_2}_{obj\downarrow})$$

where (obj, w_1, w_2) is any target dependency belonging to the set of translation candidates, ϕ (see an example of this set in Table 2). The first S computes the similarity between the two contextualized vectors associated to the head words in the source and target languages. The second S computes the similarity between the vectors derived from the dependent words. So, the overall similarity between two composite expressions is the mean of the similarity scores obtained by comparing their head-based and dependent-based contextualized vectors. The resulting translation is, thus, the expression belonging to ϕ with the highest CT score.

4 Applying Dependency-Based Contextualized Translation to Passivization

Given the syntactic nature of the phenomenon of passivization, we think that the dependency-based strategy defined in Subsection 3.2 is perfectly suited to tackle the complexity of the phenomenon. For this purpose, two new requirements are needed: to define specific dependencies for the different passive constructions in English and Spanish, and to expand the set of Spanish candidates with those syntactic constructions by making use of syntactic translation equivalents.

4.1 Passive Dependencies

In order to build the contextualized senses of passive constructions, it is necessary to identify them with the appropriate syntactic analysis. Although periphrastic passives are relatively easy to identify in both English and Spanish, to the best of our knowledge, there is no syntactic parser capable of analyzing the middle-voice constructions in Spanish. In Table 3, we show the passive expression in English and Spanish (first column) along with the syntactic dependency we are looking for (second column).

Notice that we need specific dependencies that are not even defined in the Universal Dependencies (UD) project (Nivre and others, 2017). Neither *nsubj_PP* (nominal subject of a periphrastic passive), *nsubj_RP* (nominal subject of a reflexive passive) nor *obj_IP* (direct object of an impersonal passive) are functions defined in UD. In the case of *nsubj_PP*, it is relatively simple to derive this dependency from the analysis: given the verb to *be* followed by a verb in past participle, its *nsubj* must be of type *PP*. However, the identification of *nsubj_RP* and *nsubj_IP* is much harder. Expressions with similar surface form (*se+verb+np* and *se+verb+pp/a*) represent very different constructions. Table 4 shows Spanish expressions with similar surface form (first column), their functional analysis (second column), and the type of construction (third column). Only two of the six expressions are passive constructions: the first one (*RP*) and the fourth (*IP*).

In order to identify passive dependencies, we defined a set of syntactic rules provided with lexical restrictions on verbs that were implemented with DepPattern formalism.[1] Restrictions on verbs were learned by taking into account the syntactic-semantic classes compiled in the ADESSE database (García-Miguel, Vaamonde, and Domínguez, 2010). The resulting grammars are the basis for rule-based parsing dependencies adapted to the analysis of passive constructions both in English and, especially, in Spanish.

4.2 Syntactic Translation Equivalents between Languages

The dependency-based approach described in Subsection 3.2 is based on the generation of candidates by making use of bilingual lexical information provided by an external dictionary or a cross-lingual lexical model. In the example reported above (*catch the ball*), only one type of construction (direct object) have been used in both English and Spanish. The Spanish candidates have been generated using the same construction as in English, by combining the lexical translation equivalents of the constituent words (head and dependent) of the source expression. However, in addition to the lexical translation equivalents, we also need to consider syntactic translation equivalents.

In order to generate the set of candidates in ϕ, we combine both the set of lexical translation equivalents of the two source words with the set of syntactic translation equivalents of the source dependency by means of the Cartesian product of three sets as follows:

$$\begin{aligned}\phi = ST(r_s) \times LT(w_{s1}) \times LT(w_{s2}) = \\ \{< r_t, w_{t1}, w_{t2} >: r_t \in ST(r_s), \\ w_{t1} \in LT(w_{s1}), w_{t2} \in LT(w_{s2})\}\end{aligned} \tag{2}$$

where $ST(r_s)$ is the set of syntactic translation equivalents of the source dependency r_s; $LT(w_{s1})$ is the set of lexical translation equivalents of the source head word w_{s1}; and $LT(w_{s2})$ is the set of lexical translation equivalents of the source dependent word w_{s2}. So, each $< r_t, w_{t1}, w_{t2} >$ is an ordered triple belonging to ϕ. Notice that Equation 1 defining CT above must be generalized by considering the more generic set of candidates ϕ defined in Equation 2, which includes now syntactic translation equivalents.

To deal with passivization in English-Spanish translation, we propose the set

[1] https://github.com/citiususc/DepPattern

passive expressions	**dependencies**
The house was sold	$< nsubj\_PP, sell, house >$
La casa fue vendida (*The house was sold*)	$< nsubj\_PP, vender, casa >$
La casa se vendió (*The house was sold*)	$< nsubj\_RP, vender, casa >$
Se despidió a los trabajadores (*The workers were fired*)	$< obj\_IP, despedir, trabajador >$

Table 3: Passive expressions and their dependency-based analysis. PP means periphrastic passive, RP reflexive passive, and IP impersonal passive.

Spanish expressions	**functions**	**constructions**
Se vendió la casa (*The house was sold*)	PRED-NSUBJ	reflexive passive (*RP*)
Se comió la manzana (*She/he ate the apple*)	PRED-OBJ	transitive active
Se cayó el lápiz (*The pencil fell down*)	PRED-NSUBJ	intransitive active
Se despidió a los trabajadores (*The workers were fires*)	PRED-OBJ	impersonal passive (IP)
Se comió a los niños (*(The monster) ate the children*)	PRED-OBJ	transitive active
Se arrodilló a tu lado (*She/he knelt beside you*)	PRED-OBL	intransitive active

Table 4: Spanish expressions with similar surface form to RP ($se+verb+np$) and IP ($se+verb+pp/a$) constructions.

$ST(nsubj\_PP)$ to be defined by the following elements: $\{nsub\_PP, nsub\_RP, obj\_IP\}$.

5 Experiments

5.1 Test Dataset and Evaluated Systems

In order to check to what extent the hypothesis set out in the *Periphrastic Passive Bias in English-Spanish Translation* stated in the Introduction is correct or not, we created a test dataset with 240 English passives (PP constructions). With different degrees of exigency, all the expressions of the dataset could be transferred to middle-voice constructions in Spanish (RP or IP constructions), even though most of them can also be transferred to periphrastic passives (PP), keeping so the same construction of the source language.

Table 5 quantifies the distribution of the types of constructions used by different systems for translating into Spanish the 240 English passive expressions. In this evaluation, we do not focus on the quality of the translation concerning the lexical choices, but just on the ability of the system to diversify the transfer of different passive constructions into Spanish. On the top of the table, we show four state-of-the-art commercial machine translators (supervised strategies), namely Bing,[2] DeepL,[3] Google Translator,[4] and Yandex,[5] which mainly use parallel corpora for training (all consulted in January 2020).

The test dataset was also processed with three unsupervised systems: A phrase-based SMT system (Artetxe, Labaka, and Agirre, 2018b), consisting of a log-linear combination of several statistical models learned from monolingual corpora; a hybrid SMT+NMT system (Artetxe, Labaka, and Agirre, 2019), consisting of the improved SMT system that initializes an unsupervised NMT model, which is further fine-tuned on the basis of on-the-fly back-translation; and ContextTrans, which is the enhanced version of the system based on the CT measure described above in Section 4. These three systems were trained using the same monolingual corpora, namely English and Spanish wikipedias (dump files of December 2018), with 21 and 5 billion words, respectively.

For ContextTrans, all texts were syntactically analyzed with LinguaKit (Gamallo et al., 2018), a multilingual suite which also in-

[2] `https://www.bing.com/translator`

[3] `https://www.deepl.com/translator`

[4] `https://translate.google.com/`

[5] `https://translate.yandex.com/`

supervised systems	***PP***	***RP***	***IP***	**% middle**
Yandex	219	21	-	8.75
GoogleTrans	218	22	-	9.16
DeepL	186	52	2	22.25
Bing	180	60	-	25.00
unsupervised systems	***PP***	***RP***	***IP***	**% middle**
Phrase-based_SMT	209	31	-	12.91
Hybrid_SMT+NMT	184	56	-	23.33
ContextTrans	132	94	14	**45.00**

Table 5: Distribution of the three types of passive constructions (PP, RP and IP) across the output returned by both supervised and unsupervised systems.

cludes the dependency-based parser, DepPattern (Gamallo and Garcia, 2018). The syntactically analyzed corpus was the basis for the elaboration of the salient lexico-syntactic contexts with which we constructed selectional preferences and contextualized vectors. Only lexical units occurring more than 100 times in each monolingual corpus were considered. As the lexical translation equivalents are not in the focus of the evaluation, we created a new input file for CT derived from the original test dataset. In this file, the English expressions were lemmatized and each constituent lemma was translated manually into Spanish. For instance, from the English PP expression “the aspirant was defeated”, we just kept the pair of lemmas “aspirant” and “defeat”, which were associated with their corresponding Spanish translations: “aspirante” and “derrotar”. Each Spanish pair of lemmas represents the Cartesian product of $LT(w_{s1}) \times LT(w_{s2})$, which was combined with $ST(w_s)$ to generate all the translation candidates of each English passive expression. Notice that $LT(w_{s1})$ and $LT(w_{s2})$ are sets with cardinality 1 since we only consider variation across the set of syntactic translation candidates: $ST(w_s)$. So, after combining $ST(w_s)$, $LT(w_{s1})$ and $LT(w_{s2})$, each set of candidates, ϕ, is constituted by three dependencies, one per passive construction. Notice that as lexical units were lemmatized, ContextTrans do not consider information on aspect and tense of verbs or noun number.

5.2 Analysis of the Results

As has been said before, Table 5 shows how many times the passive constructions were used by supervised and unsupervised systems. In the case of supervised translators, the diversity in the use of different constructions is poor, as their translations are mostly done with PP construction. Besides, the use of IP is practically non-existent. Only DeepL uses it twice. The rest of the translators never use it, even though there are at least 60 expressions that could be translated that way in our dataset. Among the supervised systems, the two that return more syntactic diversity are DeepL (186 PP, 52 RP and 2 IP) and Bing (180 PP and 60 RP). Google Translator and Yandex behave very similarly with very poor diversity.

Concerning the unsupervised systems, Phrase-based_SMT is the least diverse, but it stands above the two least diverse supervised systems. Hybrid_SMT+NMT is between the two most diverse supervised systems, while ContextTrans is the system with the greatest diversity by large: 132 PP, 94 RP, 14 IP. The last column of Table 5 shows the percentage of middle-voice constructions with regard to the total number of examples in the dataset. The higher the percentage value, the less literal the translation is with respect to syntactic constructions. In the case of ContextTrans, 45% of constructions are not literal (middle-voice), which is almost twice as many cases as the most diverse supervised system. The few number of IP constructions returned by unsupervised approaches suggest that naturally written texts in Spanish contain fewer expressions with this type of construction than with RP.

The results seems to confirm the *Periphrastic Passive Bias* hypothesis, as the methods relying on monolingual corpora tend to offer more non-literal translations (middle-voice) than those trained on parallel corpora. The average of middle-voice translations with the four supervised systems is 16,29%, whereas the average for the three unsupervised systems reaches 27,08.

6 Conclusions

We have carried out an experiment to compare the syntactic diversity between supervised and unsupervised approaches to translation on one dataset consisting of English passive expressions. We have tuned a dependency-based translation strategy trained on monolingual corpora and verified, on the basis of its application to the dataset, that its syntactic diversity is greater than that of commercial translators relying on supervised techniques. These results confirm the hypothesis made at the beginning of our work in which we stated that both supervised systems have a bias towards more literal translation (PP constructions in English are translated by PP constructions in Spanish), and monolingual corpora allow learning a greater diversification of passive structures. So, natural text seems to be more syntactically diverse than MT output and parallel corpus. It should be noted that the syntax-based strategy, ContextTrans, can be considered a hybrid approach that integrates symbolic-syntactic knowledge in statistical-neural learning systems.

As the methodology can be applied to other linguistic phenomena and transferred to different language pairs, in future work we will seek to extend the experimentation towards other types of syntactic constructions taking into account the linguistic studies of Construction Grammar and its application to cross-lingual construction transfer (Boas, 2010). Experiments will also be conducted with a wider variety of linguistic genres, from literary to spoken corpora.

The code for the generic version of ContextTrans (without passive tuning), called compMT, is available at GitHub (`https://github.com/gamallo/compMT`). The dataset with the English passive expressions are availabe at `https://gramatica.usc.es/pln/resources/en_sentences240.txt.zip`.

Acknowledgments

This work has received financial support from DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE), eRisk (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08, Groups of Reference: ED431C 2020/21) and the European Regional Development Fund.

References

Alarcos Llorach, E. 1978. Valores de 'se'. In *Estudios de gramática funcional del español*. Madrid, Gredos, pages 156–165.

Artetxe, M., G. Labaka, and E. Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2018b. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, October-November. Association for Computational Linguistics.

Artetxe, M., G. Labaka, and E. Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July. Association for Computational Linguistics.

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR-2018)*, April.

Boas, H. 2010. *Contrastive Studies in Construction Grammar*. John Benjamins Publishing Company.

de Miguel, E. 1999. El aspecto léxico. In I. Bosque and V. Demonte, editors, *Gramática descriptiva de la lengua española, vol. 2*. Madrid: Real Academia Española; Espasa Calpe.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Erk, K. and S. Padó. 2008. A structured vector space model for word meaning in context. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008*, pages 897–906, Honolulu, HI.

Erk, K., Sebastian, Padó, and U. Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Fernández, S. S. 2007. *La voz pasiva en español: un análisis discursivo*. Frankfurt am Main: Peter Lang.

Gamallo, P., M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Gamallo, P. 2019. A dependency-based approach to word contextualization using compositional distributional semantics. *Language Modelling*, 7(1):53–92.

Gamallo, P. and M. Garcia. 2018. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management*, 54(6):1244–1261.

Gamallo, P. and M. Garcia. 2019. Unsupervised compositional translation of multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 40–48, Florence, Italy, August. Association for Computational Linguistics.

Gamallo, P., S. Sotelo, J. R. Pichel, and M. Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics*, 45(3):395–421.

García-Miguel, J. M., G. Vaamonde, and F. G. Domínguez. 2010. ADESSE, a database with syntactic and semantic annotation of a corpus of Spanish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Garcia-Miguel, J. M. 1985. La voz media en español: Las construcciones pronominales con verbos transitivos. *Verba: Anuario galego de filoloxia*, 12:307–343.

Jisa, H., E. Baruch, J. Reilly, E. Rosado, L. Tolchinsky, L. Verhoeven, and A. Zamora. 2002. Passive voice constructions in written texts: A cross-linguistic developmental study. *Written Language and Literacy*, 5(2):163–182.

Keenan, E. L. 1985. Passive in the world's languages. In T. Shopen, editor, *Language Typology and Syntactic Description. Vol. I.* Cambridge: Cambridge University Press.

Lample, G., A. Conneau, L. Denoyer, and M. A. Renzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the Sixth International Conference on Learning Representations (ICLR-2018)*, April.

Lample, G., M. Ott, A. Conneau, L. Denoyer, and M. A. Ranzato. 2018b. Phrase-Based &amp; neural unsupervised machine translation, April.

Lourdes Díaz Blanca, C. L. D. 2008. Los verbos en las pasivas con se: un intento de clasificación. *Letras [online]*, 50(76).

Nivre, J. et al. 2017. Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, `http://hdl.handle.net/11234/1-1983`.

Rodríguez-Vergara, D. 2017. A systemic functional approach to the passive voice in english into spanish translation: Thematic development in a medical research article. *Open Linguistics*, 3(1).

Scarpa, F., 2020. *Translating Specialised Texts*, pages 187–290. Palgrave Macmillan UK, London.

Siewierska, A. 1984. *The Passive: Comparative Linguistic Analysis.* Routledge, Croom Helm Linguistics Series, London.

Sánchez-López, C. 2002. Las construcciones con se. estado de la cuestión. In C. Sánchez López, editor, *Las construcciones con se.* Madrid: Visor, pages 18–163.

Toral, A. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.

Vanmassenhove, E., D. Shterionov, and A. Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.

Vinay, J. and J. Darbelnet. 1995. *Comparative stylistics of French and English* Benjamins, Amsterdam.