Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 235-246 recibido 04-04-2024 revisado 03-05-2024 aceptado 16-05-2024

Characterizing Spans for Sequence Labeling: A Case on Anglicism Detection

Caracterización de spans en tareas de etiquetado de secuencias: el caso de la detección de anglicismos

Elena Álvarez Mellado, Julio Gonzalo
Universidad Nacional de Educación a Distancia
{elena.alvarez, julio}@lsi.uned.es

Abstract: We propose a set of formal dimensions to characterize spans in sequence labeling evaluation. We apply them to a dataset and model results obtained for anglicism detection in Spanish. Results show that the best performing system is outperformed by other models on certain types of spans. Our methodology can uncover limitations in performance that go unnoticed with standard evaluation.
Keywords: Span identification, sequence labeling, evaluation, anglicism detection.

Resumen: Presentamos un conjunto de dimensiones para caracterizar spans en la evaluación de etiquetado de secuencias y las aplicamos a la tarea de detección de anglicismos en castellano. Los resultados muestran que las dimensiones ayudan a desenmascarar limitaciones que pasaron desapercibidas en la evaluación estándar.
Palabras clave: Etiquetado de secuencias, evaluación, detección de anglicismos.

1 Introduction

Span identification tasks are a type of sequence labeling tasks that consists of retrieving spans of interest from text (Papay, Klinger, and Padó, 2020). Named entity recognition, chunking or multiword expression processing are prime examples of span identification tasks.

Span identification models are usually evaluated over a given test set using holistic metrics, such as F1 score. As useful as it may be, this type of evaluation provides little insight about the capabilities of these models (what they excel at, what they struggle with), and even less about their ability to generalize to new data (Fu, Liu, and Neubig, 2020).

In this work we propose a set of dimensions that may help characterize spans in span identification tasks. We motivate the dimensions for the task of detecting anglicisms, but they can be applied to other span-based problems, such as NER.

2 Related work

2.1 Span-id evaluation

Recent work has pointed out the need to dive deeper into the results produced by NLP models in order to get a better understanding of what models are capable and (more importantly) not capable of, and thus anticipate how models may generalize to new data (Zhou et al., 2023). Prior research has questioned the suitability of evaluating systems testing solely on the standard test set, and some voices have advocated for evaluating models on alternative splits, such as random, heuristics or adversarial splits or on multiple test sets (Gorman and Bedrick, 2019; Søgaard et al., 2021).

In terms of fine-grained evaluation in span-based tasks, Bernier-Colborne and Langlais (2020) proposed calculating the percentage of mislabeled tokens in the test set that were *unseen* (never seen dur-

ISSN 1135-5948 DOI 10.26342/2024-73-17

ing training) or *diff* (whose label was not the most frequent label in training) and found out that diff instances were more error-prone than unseen tokens.

Similarly, Tu and Lignos (2021) proposed evaluating NER models on what they call *tough mentions*, i.e. entities that were unseen (because they were not seen during training, or because they were seen but with a different label), or type-confusable (mentions that appeared in the test set with multiple labels).

Papay, Klinger, and Padó (2020) investigated the impact that four formal attributes (label frequency, span length, span distinctiveness, boundary distinctiveness) can have on model performance in several span identification tasks (NER, chunking and quotation detection) and concluded that frequent spans are easier to learn, while longer spans are more difficult to retrieve.

Fu, Liu, and Neubig (2020) proposed an interpretable approach to NER evaluation by partitioning the test set into a series of buckets based on a set of entity attributes (entity length, sentence length, token frequency, entity frequency, token label consistency, etc.), and analyzed how different models performed on instances with similar attributes.

2.2 Anglicism detection

Anglicism detection is the task of retrieving English lexical borrowings (or *anglicisms*) from non-English texts. Anglicisms can be single-token (*app*) or multitoken (*machine learning*, *fake news*). The task of automatically retrieving lexical borrowings from text has proven to be useful both for lexicographic purposes and for NLP downstream tasks in various languages (Furiassi and Hofland, 2007; Alex, 2008; Andersen, 2012; Losnegaard and Lyse, 2012; Tsvetkov and Dyer, 2016; Serigos, 2017) and has previously been framed as a span-identification task (Alvarez Mellado, 2020; Alvarez Mellado et al., 2021; Chiruzzo et al., 2023).

In Spanish, Alvarez-Mellado and Lignos (2022) recently proposed an annotated corpus and a suite of sequence labeling models for the task of retrieving anglicisms from Spanish text, with models scoring from 55.44 to 84.22 on F1 score. As we will see in section 3, these results, however, are insufficient to truly understand what these models can and cannot actually do.

3 Rationale

Not all anglicisms are created equal. Words such as *streaming*, *online* or *pie* may look equally unremarkable to a native speaker of English, but as anglicisms they all appear very different to the eye (and the ear) of a monolingual speaker of Spanish. Some borrowings display a combination of letters that looks unusual in the recipient language, while others could pass for a native word.

For instance, an anglicism like *streaming* violates the graphotactical expectations that Spanish speakers have, because words in Spanish are not supposed to begin with *str-* or end in *-ing*. *Streaming* could never pass for a word in Spanish.

On the other hand, a word like *online*, while still being an anglicism in Spanish, displays a combination of letters that does not look suspicious in Spanish. It is still a borrowing because it is used as such: the pronunciation of the word follows the English rules of pronunciation, not the Spanish ones. But even when we know that *online* is a borrowing in Spanish, there is nothing in its shape that makes it foreign or that prevents it from being a word in Spanish. A monolingual speaker of Spanish that encountered the word *online* for the first time in written text could easily think that it is a word in Spanish that is unknown to them, not

necessarily assume that it must be a borrowing, unlike *streaming*.

Finally we have a word like *pie*, which is a fully Spanish word when it means "foot", but that may also be an anglicism when it refers to the dessert, as in "un *pie* de limón". When used as a borrowing, the word *pie* will be pronounced according to the English pronunciation rules, but there is nothing preventing the combination of letters p+i+e from being a word in Spanish, and, in fact, it is also a word in Spanish.

All these nuances can make the task of detecting anglicisms in text harder or easier for NLP models. A borrowing like *streaming* will be easier to spot, because just the combination of letters makes it a very good candidate as a foreign word. A model that incorporates the frequency of character n-grams (such as the rule-based model proposed in Serigos (2017) or the CRF model from Alvarez Mellado (2020)) will probably be able to detect it.

A word like *online*, on the contrary, would probably be harder to spot based on character combination only, because there is nothing unusual about the combination of characters in the sequence. A model would need to have seen *online* labeled as a borrowing during training or rely on contextual information to correctly detect it as a borrowing.

Finally, the word *pie* will be the hardest case of all: not only could it easily escape any character-based model, but also the fact of having seen it as a Spanish word or as a borrowing during training does not guarantee that the model will be capable of making the right choice when confronted with *pie* in the test set.

Similarly, certain contexts will make the detection task harder or easier. According to the orthotypographic rules of Spanish, in standard text unassimilated borrowings are expected to be written surrounded by quotation marks (Real Academia Española, 2011). Consequently, anglicisms surrounded by quotation marks will stand a better chance of being retrieved, because the quotation marks may serve as a cue to the models. Capitalization may also affect the detection process: anglicisms in upper case (such as *Big Data*) or in sentence-initial position may have a higher chance of being mistaken with proper names, and are therefore more likely to be missed. It is doubtful that anglicism detection models will perform evenly across all these different scenarios.

What lies at the crux of the issue is that in any span identification task (such as anglicism detection), the *retrievability* of the span will be affected by certain aspects, such as whether the span was seen in training, its shape or the context it appears in. However, none of these issues are taken into account when evaluating the performance of span identification models. How good are models at retrieving sentence initial position spans? Are they exclusively learning to detect certain formal cues, such as quotation marks? Are graphotactical non-compliant spans (odd-looking borrowings) being privileged over compliant ones?

4 Dimensions: proposal

Following the reasoning from section 3, we now define a set of linguistically-motivated dimensions that seek to characterize the anglicisms in a test set.

1. Intrinsic dimensions: dimensions that affect the formal shape of the span itself, regardless of context.

   – **Single token vs multitoken**: the number of tokens that the span comprises (*online* vs *machine learning*).

   – **Graphotactic compliance**: whether the span complies with

the graphotactic patterns of the recipient language. For example, an anglicism such as *healthy* could never be a word coined in Spanish, because we would never expect a Spanish word to have a *-th-* digraph or to end in *-thy*. On the other hand, an anglicism such as *prime time* fully complies with the graphotactics rules of Spanish.

2. Contextual dimensions: dimensions that reflect the context of the span.
   – **Sentence initial**: whether the span appears in the first position of a sentence. Sentence-initial anglicisms can be mistaken with proper names and can be difficult to detect.
   – **Quotations**: whether the span is surrounded by quotation marks or in italics. Quotations may function as a cue and make detection easier.
   – **Capitalization**: whether the span is capitalized. Some borrowings may sometimes be written capitalized, even if they are not proper nouns (for ex., *Big Data*), which may make detection harder, as systems can mistake them with proper nouns.
   – **Adjacent spans**: whether the span is preceeded by another span. For example, the sequence *look total black* is comprised of two adjacent anglicisms (*look* and *total black*) that happen to be collocated. These nuanced combinations may lead to segmentation errors.
3. Novelty dimensions: dimensions that characterize the span in relation to its presence in the training set.
   – **Seen vs unseen**: whether the span was seen during training.
   – **Type confusable**: the span was seen in training with at least two different tags. For example the word *post* may be seen during training both as the Spanish prefix of Latin orgin (as in *post guerra*), and as the anglicism that refers to an online text (*un post de Facebook*). The model will have to decide which of the labels seen in training is the appropriate one when labeling *post* on the test set.
   – **Unseen tagging**: the span was seen in training, but not with the tag that it bears in the test set. For example, let's say that the Spanish word *red* (which means "net") was seen during training, and that the test set contains the previously unseen anglicism *red carpet*. The model will have to apply a label to the word *red* that never saw applied to *red* in training.

These dimensions are not mutually exclusive, and they may combine in various ways. In fact, these dimensions may help us analyze how different dimensions interact with one another. A given sentence may contain an anglicism that is multitoken, unquoted, with a non-compliant shape and that was never seen during training.

5 A practical case on anglicism detection

We now propose to apply the dimensions introduced in section 4 to the dataset and modeling results presented in Alvarez-Mellado and Lignos (2022) for the task of detecting anglicisms from Spanish text. The aim is to investigate if by applying the dimensions to the dataset and to the models' results we can gain some insight into the composition of the dataset (how well represented are different types of spans in the corpus?) and into the models' performance (how good are the models at detecting different types of spans?).

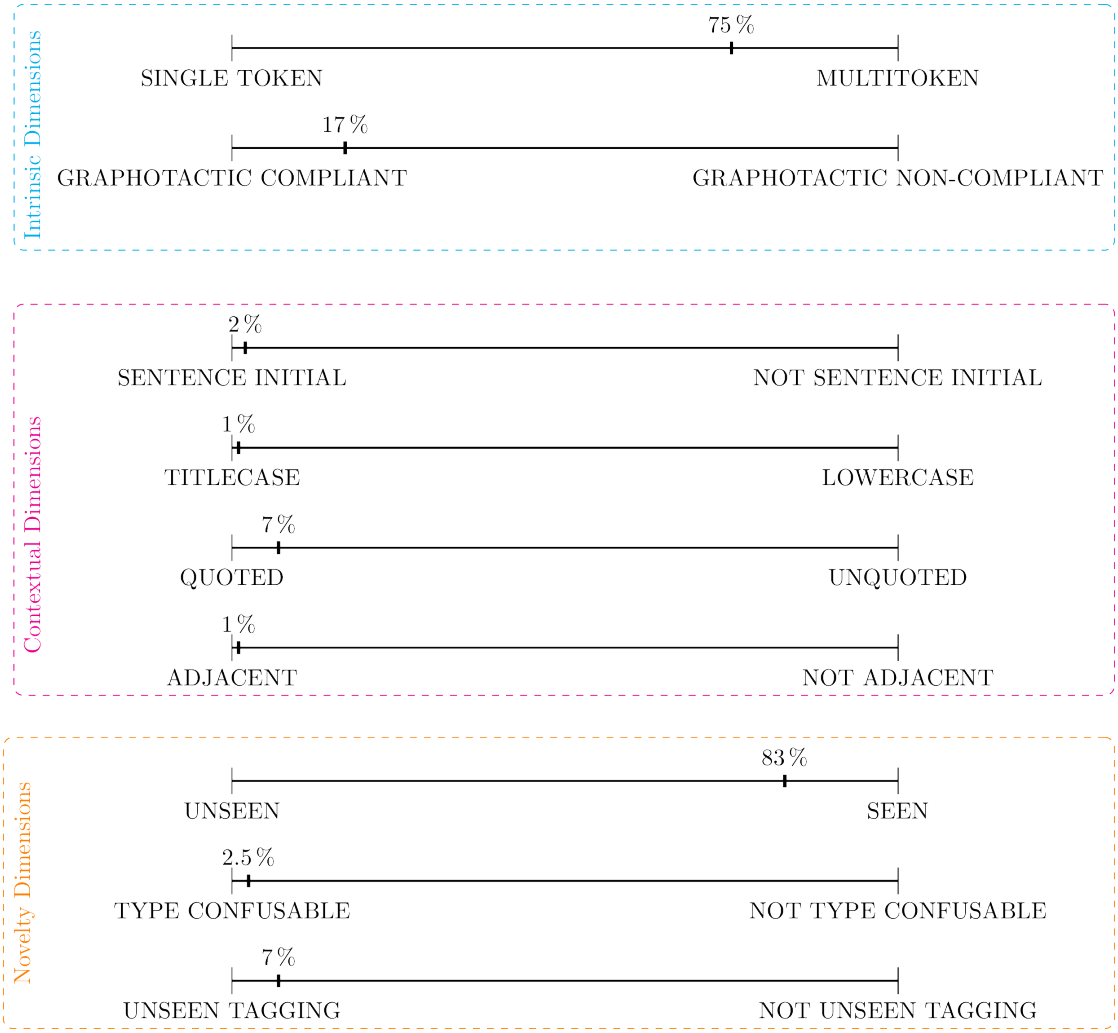

Figure 1: Percentages per dimension for spans in the test set.

5.1 Analysis of the dataset

COALAS (Corpus of Anglicisms in the Spanish Press) is a publicly available dataset of Spanish journalistic text annotated with anglicisms and other unassimilated lexical borrowings (Alvarez-Mellado and Lignos, 2022). The corpus contains 370,000 tokens and its test set is rich in out-of-vocabulary words, with a large majority of the spans in the test set being OOV (unseen during training).

Figure 1 displays the distribution of spans in the test set per dimension. A look at the percentages can give us an idea of what the typical anglicism in the test set looks like: the prototypical anglicism in the COALAS test set is a span of length 1 that does not comply with the graphotactical rules of Spanish spelling, that was not seen during training, that appears mid sentence, not collocated with other borrowings and that is written in lower case, without quotations.

These numbers also show that, as OOV rich as the test set may be, certain phenomena are under-represented

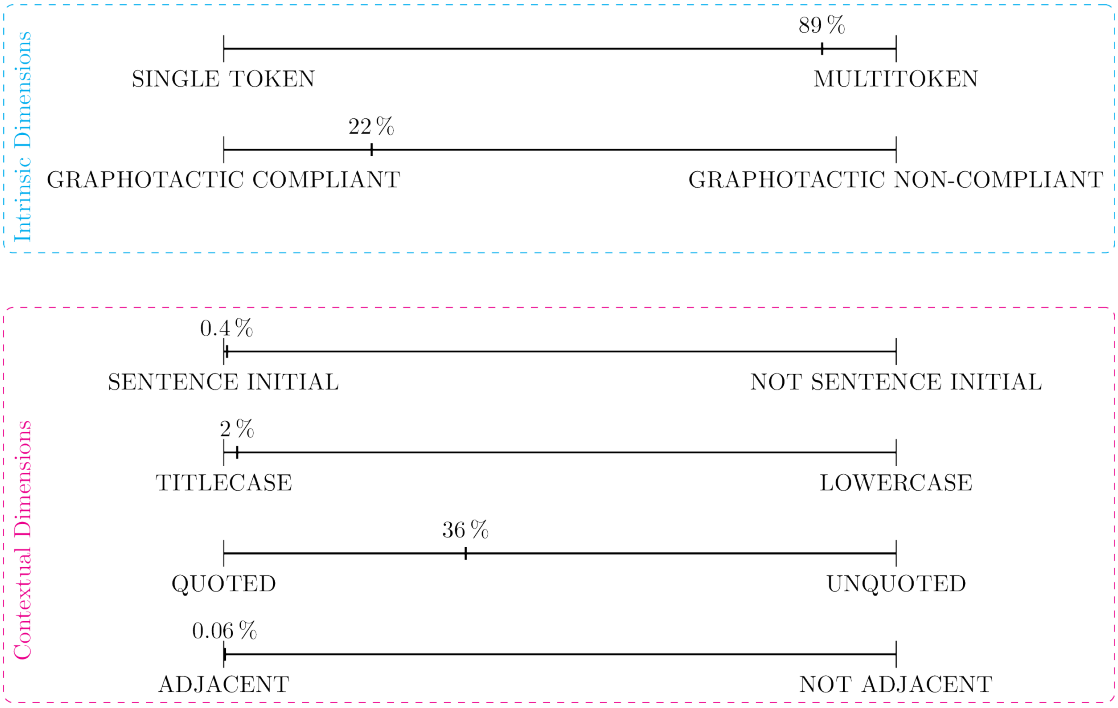

Figure 2: Percentages per dimension for spans in the training set.

in the COALAS corpus. The test set is clearly biased towards graphotactical non-compliant anglicisms, which may very well explain why in the results reported by Alvarez-Mellado and Lignos (2022), models fed with subword embeddings produced better results. The numbers of anglicisms in sentence initial position, capitalized or collocated with other borrowings are also minimal. This raises the question of how good models are at detecting anglicisms in these challenging contexts, as in the hypothetical case that the models performed poorly on those contexts, there would be so few cases that the results would easily go unnoticed on the aggregated evaluation metrics.

Although the novelty dimensions can only be applied to spans from the test set, we may also apply the intrinsic and contextual dimensions to the spans in the training set, and see how much the dimensions from the test set align with the dimensions from the training set. Figure 2 shows that the spans in the training set share many of the properties of the spans in the test set, but some of them are slightly different.

The main difference lies on the amount of quoted spans: 36% of the spans in the training set had quotation marks, but only 7% of the spans in the test set had them. This opens the door to the question of whether the models tended to perform better on the spans of the test set that had quotation marks or if they overgeneralized and inappropiately marked as a borrowing other quotated spans of non-borrowing text.

Other subtle differences between the dimensions from the test set and the training set are sentence initial spans: although spans in the first position of the sentence were infrequent in both splits,

they were four times more frequent in the test set (2%) than in the training set (0.5%).

5.2 Analysis of the models

With the percentages we have just provided, we can rightfully ask ourselves some questions: given that the majority of spans in the test set were single token, how well did the models reported by Alvarez-Mellado and Lignos (2022) perform on multitoken spans? How good was the performance on graphotactic compliant anglicisms? Was the unbalance in quotations between the test and the training set and issue for the models?

To answer these questions, we re-evaluated the five publicly available anglicism detection models produced by Alvarez-Mellado and Lignos (2022): a CRF with handcrafted features, Spanish Transformer-based BETO, multilingual BERT, a BiLSTM-CRF model fed with bilingual Transformer-based embeddings and a BiLSTM-CRF model fed with embeddings pretrained on codeswitched data. We performed the evaluation on the COALAS corpus (the same corpus that the models were originally evaluated on) but we splitted the test set into different subsets according to the different dimensions defined in section 4. For each dimension, results were calculated only on the spans concerned with the dimension in question. In other words, for the sentence initial dimension results, only the spans that satisfied the property of being in sentence initial position were taken into account. If other non-initial borrowings were present in the sentence, they were skipped. Table 1 reports overall results of the five models on the COALAS test set. Tables 2 and 3 display recall and precision per model on the different subsets of the test set defined by our dimensions.

The results obtained per dimension paint a very different picture to the results portrayed by aggregated precision and recall. The BiLSMT-CRF model with codeswitched embeddings (which obtained the highest F1 score overall) ranks second to last across type-confusable, sentence initial and quoted spans. In these dimensions, the BiLSTM-CRF model only outperforms the very modest CRF model. It even ranks last on adjacent spans, with a recall of just 12.50. The BiLSTM-CRF model still outperforms all models across other dimensions, such as unseen spans, single token spans, capitalized spans, unquoted spans and graphotactic non-compliant spans, but the results across dimensions enable to detect some of its shortcomings.

The dimensions also allow to discover some strengths that had gone unnoticed in previous evaluations. The BETO model, which obtained a modest position in the original ranking (second to last, only outperforming the CRF model) seems to obtain surprisingly good results in terms of recall across some of the dimensions where other models fail: BETO ranks first across type-confusable spans, sentence initial, quoted spans (with a spectacular R=97.98) and graphotactic compliant spans (the type of spans that could easily be missed by character-based models). BETO also ranks second in terms of recall across unseen spans, multitoken spans, single token spans, unseen tagging, unquoted spans and graphotactic non-compliant spans. Multilingual BERT was better than other models at retrieving multitoken spans and spans with unseen tagging, while the BiLSTM-CRF model fed with bilingual Transformer-based embeddings was the best at correctly identifying previously seen spans and adjacent spans.

The dimensions also show some of the weakness in the models: the CRF model got passable results at identifying previ-

ously seen or quoted spans, but its predictions are totally unreliable for sentence initial or capitalized spans.

The results across dimensions also confirm our concerns about the unbalanced phenomena in the dataset: all models found quoted spans easier to retrieve than unquoted spans, as well as non-compliant spans were easier than compliant ones. Scores for sentence initial spans were catastrophic across all models, but if we had to choose a model for detecting sentence initial anglicisms, BETO and the BiLSTM-CRF with bilingual embeddings would probably be the safest bet, while the BiLSTM-CRF model with codeswitch embeddings would probably be the best call for capitalized spans.

6 Discussion

6.1 Benefits of the dimensions

There are three main areas where the proposed dimensions may be of use:
– Model evaluation: Evaluation is the *raison d'être* of these dimensions. Their main purpose is to help characterize spans, discover systematic blind spots in models' performance, and identify error-prone instances in the test set. In that regard, these dimensions satisfy the three requirements that Fu, Liu, and Neubig (2020) defined for an ideal evaluation methodology, namely being fully automatic, allowing comparison across different datasets and allowing users to dig deeper into the strengths and weaknesses of the model. Results across dimensions can also help us decide which model is the most appropriate for a given data-specific scenario (capitalized data, previously seen spans, etc).
– Corpus documentation. Recent research has emphasized the need to thoroughly document the type of data contained in NLP corpora (Bender and Friedman, 2018). Data documentation in sequence labeling usually focuses on reporting the number of spans, unseen tokens or provenance of the data, which does not tell us much about the characteristics of the data that the models will have to learn. Applying the dimensions to the COALAS corpus provided valuable information on the type of spans in the corpus and uncovered some of the limitations and under-represented phenomena that this dataset has. Reporting the number of spans per dimensions in sequence labeling datasets can help us get a better understanding of the type of phenomena contained in the test set and allows us to check how well-aligned the different splits are.
– Data curation: on a similar note, span dimensions may be of use during the process of creating a corpus, as it may help to decide the type of instances to allocate to the different splits. Dimensions may help identify that certain types of spans are ill-represented in a given split or, on the contrary, can be helpful when building adversarial splits to test generalization.

6.2 Differences with prior work

The dimensions we have just introduced draw from some of the previous work presented in section 2.1. Our unseen dimension, unseen tagging dimension and type-confusable dimension build from the work in Bernier-Colborne and Langlais (2020), Fu, Liu, and Neubig (2020) and Tu and Lignos (2021).

Our proposal however substantially diverges from prior work and complements some aspects. Bernier-Colborne and Langlais (2020) and Tu and Lignos (2021) exclusively focused on whether a span or its labeling was previously seen or not. Although it has been shown that this aspect heavily corre-

lates with model performance (Bernier-Colborne and Langlais, 2020; Fu, Liu, and Neubig, 2020), the fact that a span was seen in training can hardly be the sole aspect that makes span detection challenging. In fact, we know that models are capable of correctly retrieving previously unseen borrowings. And we know this because models reported by Alvarez-Mellado and Lignos (2022) obtained reasonably good results when tested on a test set that was extremely rich in previously unseen spans. But we also know that, while being able to retrieve *some* previously unseen borrowings, those models were not capable of retrieving *all* previously unseen borrowings, which implies that some previously unseen borrowings were more challenging to spot than others. A previously unseen span may or may not be challenging to retrieve because of other aspects, such as its shape or the context it appears in.

Fu, Liu, and Neubig (2020) did consider some sentence attributes, such as sentence length or span density, but this stills overlooks the fact that certain properties of the span itself (such as casing) or of the context in which the span appears (such as position within the sentence or token surroundings) will affect the retrievability of the span. And that is the blind spot that our proposal seeks to address: spans in certain contexts will be easier to spot than others (thus, the sentence-initial dimension); the presence of typographic cues may help models flag certain spans (thus, the quotation marks dimensions); casing may contribute to some spans going unnoticed (thus the capitalization dimension). Being unseen may be an important attribute, but it is far from the only one and disregarding all other aspects will prevent us from understanding which models are good at what.

6.3 Applicability to other span identification tasks

We have built our case on anglicism detection, but this methodology can be applied to other span identification tasks, such as NER, where polysemous and adjacent spans are pervasive and span retrievability is also context-sensitive.

All dimensions are directly applicable to other tasks, except for the graphotactic compliance dimension. However, this dimension points to a crucial issue that can also be applied to other tasks: the degree of distinctiveness in the spans. In borrowing detection, non-compliant borrowings will be more salient (and therefore, easier to spot), while compliant borrowings may go unnoticed. In NER, research has shown that models trained on data that is rich in English proper names will find it difficult to spot entities in other languages (Lin et al., 2021; Vajjala and Balasubramaniam, 2022). In the case of lexical borrowing, we chose to implement this feature as a binary dimension using formal constraints based on linguistic graphotactics, but continuous representations based on the Kullback–Leibler divergence have also been used in other span identification tasks (Papay, Klinger, and Padó, 2020).

7 Conclusion

We have defined a set of dimensions to characterize the spans in span identification tasks and applied them to the task of anglicism detection. Results show that the dimensions can uncover limitations in the dataset and systematic errors in the models' performance that had gone unnoticed in the standard evaluation.

Acknowledgements

This research is funded by a FPI-UNED 2019 predoctoral contract.

Model	Precision	Recall	F1 score
CRF	77.89	43.04	55.44
BETO	85.03	81.32	83.13
mBERT	88.08	79.38	83.50
BiLSTM w/ BETO+BERT	**90.35**	80.16	84.95
BiLSTM w/ Codeswitch	90.14	**81.79**	**85.76**

Table 1: Precision, recall and F1 score results obtained on the five models released by Alvarez-Mellado and Lignos (2022).

Model	**Overall recall**	**Unseen**	**Seen**	**Multi-token**	**Single-token**	**Type confusable**	**Unseen tagging**	**Sentence initial**	**Capitalized**	**Quoted**	**Unquoted**	**Adjacent**	**Graph. n/ compl.**	**Graph. compl.**
CRF	43.04	34.75	81.86	36.39	45.30	51.61	24.73	8.33	12.50	74.75	40.39	25.00	46.09	26.26
BETO	81.32	79.51	89.82	83.79	80.48	93.55	66.67	41.67	25.00	97.98	79.93	18.75	83.72	68.18
mBERT	79.38	77.43	88.50	84.10	77.77	90.32	74.19	33.33	31.25	96.97	77.91	12.50	82.98	59.60
BiLSTM w/ BETO+BERT	80.16	77.71	91.59	82.87	79.23	74.19	62.37	41.67	18.75	95.96	78.84	31.25	83.26	63.13
BiLSTM w/ Codeswitch	81.79	79.79	91.15	83.79	81.11	70.97	63.44	33.33	50.00	93.94	80.78	12.50	85.37	62.12

Table 2: Recall obtained by each model on different types of spans. For each model, the results obtained by the best performing random seed were chosen. Both BiLSTM-CRF models included BPE embeddings and character embeddings.

Model	**Overall precision**	**Unseen**	**Seen**	**Multi-token**	**Single-token**	**Type confusable**	**Unseen tagging**	**Sentence initial**	**Capitalized**	**Quoted**	**Unquoted**	**Adjacent**	**Graph. n/ compl.**	**Graph. compl.**
CRF	77.89	75.26	83.71	73.01	79.34	53.33	58.97	25.00	13.33	81.32	77.38	40.00	78.65	71.23
BETO	85.03	83.20	93.55	84.57	85.19	93.55	71.26	76.92	36.36	92.38	84.34	21.43	86.75	75.00
mBERT	88.08	87.98	88.11	86.75	88.48	62.22	75.82	61.54	55.56	92.31	87.58	18.18	89.75	76.62
BiLSTM w/ BETO+BERT	90.35	89.26	94.95	88.27	91.12	88.46	79.45	90.91	50.00	90.48	90.34	45.45	91.05	85.62
BiLSTM w/ Codeswitch	90.14	89.42	93.21	88.10	90.88	68.75	76.62	72.73	72.73	93.00	89.87	50.00	90.71	86.01

Table 3: Precision obtained by each model on different types of spans. For each model, the results obtained by the best performing random seed were chosen. Both BiLSTM-CRF models included BPE embeddings and character embeddings.

Bibliografía

Alex, B. 2008. *Automatic detection of English inclusions in mixed-lingual data with an application to parsing.* PhD Thesis, University of Edinburgh.

Alvarez Mellado, E. 2020. Lázaro: An extractor of emergent anglicisms in spanish newswire. Master's thesis, Brandeis University.

Alvarez Mellado, E., L. Espinosa Anke, J. Gonzalo Arroyo, C. Lignos, and J. Porta Zamorano. 2021. Overview of ADoBo 2021: Automatic Detection of Unassimilated Borrowings in the Spanish Press. *Procesamiento del Lenguaje Natural*, 67(0):277–285, September. Number: 0.

Alvarez-Mellado, E. and C. Lignos. 2022. Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3868–3888, Dublin, Ireland, May. Association for Computational Linguistics.

Andersen, G. 2012. Semi-automatic approaches to Anglicism detection in Norwegian corpus data. In C. Furiassi, V. Pulcini, and F. Rodríguez González, editors, *The anglicization of European lexis.* John Benjamins, pages 111–130.

Bender, E. M. and B. Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604. Place: Cambridge, MA Publisher: MIT Press.

Bernier-Colborne, G. and P. Langlais. 2020. HardEval: Focusing on Challenging Tokens to Assess Robustness of NER. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1704–1711, Marseille, France, May. European Language Resources Association.

Chiruzzo, L., M. Agüero-Torales, G. Giménez-Lugo, A. Alvarez, Y. Rodríguez, S. Góngora, and T. Solorio. 2023. Overview of GUASPA at IberLEF 2023: Guarani-Spanish Code Switching Analysis. *Procesamiento del Lenguaje Natural*, 71(0):321–328, September. Number: 0.

Fu, J., P. Liu, and G. Neubig. 2020. Interpretable Multi-dataset Evaluation for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online, November. Association for Computational Linguistics.

Furiassi, C. and K. Hofland. 2007. The retrieval of false anglicisms in newspaper texts. In *Corpus Linguistics 25 Years On.* Brill Rodopi, pages 347–363.

Gorman, K. and S. Bedrick. 2019. We Need to Talk about Standard Splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy, July. Association for Computational Linguistics.

Lin, B. Y., W. Gao, J. Yan, R. Moreno, and X. Ren. 2021. RockNER: A Simple Method to Create Adversarial Examples for Evaluating the Robustness of Named Entity Recognition Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–

3737, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Losnegaard, G. S. and G. I. Lyse. 2012. A data-driven approach to anglicism identification in Norwegian. In G. Andersen, editor, *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*. John Benjamins Publishing, pages 131–154.

Papay, S., R. Klinger, and S. Padó. 2020. Dissecting Span Identification Tasks with Performance Prediction. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4881–4895, Online, November. Association for Computational Linguistics.

Real Academia Española. 2011. *Ortografia de la Lengua Española*. Planeta Publishing, April.

Serigos, J. R. L. 2017. Applying corpus and computational methods to loanword research : new approaches to Anglicisms in Spanish. August.

Søgaard, A., S. Ebert, J. Bastings, and K. Filippova. 2021. We Need To Talk About Random Splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online, April. Association for Computational Linguistics.

Tsvetkov, Y. and C. Dyer. 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research*, 55:63–93.

Tu, J. and C. Lignos. 2021. TMR: Evaluating NER Recall on Tough Mentions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 155–163, Online, April. Association for Computational Linguistics.

Vajjala, S. and R. Balasubramaniam. 2022. What do we really know about State of the Art NER? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5983–5993, Marseille, France, June. European Language Resources Association.

Zhou, L., P. A. Moreno-Casares, F. Martínez-Plumed, J. Burden, R. Burnell, L. Cheke, C. Ferri, A. Marcoci, B. Mehrbakhsh, Y. Moros-Daval, S. hÉigeartaigh, D. Rutar, W. Schellaert, K. Voudouris, and J. Hernández-Orallo. 2023. Predictable Artificial Intelligence, October. arXiv:2310.06167 [cs].