# Cross-Lingual Transfer Learning for Complex Word Identification

George-Eduard Zaharia
*Computer Science Department*
*University Politehnica of Bucharest*
Bucharest, Romania
george.zaharia0806@stud.acs.upb.ro

Dumitru-Clementin Cercel
*Computer Science Department*
*University Politehnica of Bucharest*
Bucharest, Romania
dumitru.cercel@upb.ro

Mihai Dascalu
*Computer Science Department*
*University Politehnica of Bucharest*
Bucharest, Romania
mihai.dascalu@upb.ro

***Abstract*—Complex Word Identification (CWI) is a task centered on detecting hard-to-understand words, or groups of words, in texts from different areas of expertise. The purpose of CWI is to highlight problematic structures that non-native speakers would usually find difficult to understand. Our approach uses zero-shot, one-shot, and few-shot learning techniques, alongside state-of-the-art solutions for Natural Language Processing (NLP) tasks (i.e., Transformers). Our aim is to provide evidence that the proposed models can learn the characteristics of complex words in a multilingual environment by relying on the CWI shared task 2018 dataset available for four different languages (i.e., English, German, Spanish, and also French). Our approach surpasses state-of-the-art cross-lingual results in terms of macro F1-score on English (0.774), German (0.782), and Spanish (0.734) languages, for the zero-shot learning scenario. At the same time, our model also outperforms the state-of-the-art monolingual result for German (0.795 macro F1-score).**

***Index Terms*—Complex Word Identification, Transformer, Cross-Lingual Transfer Learning**

## I. Introduction

Texts represent the main source of knowledge for our society. However, they can be written in various manners, thus creating a barrier between the readers and the ideas they intend to convey. Therefore, document comprehension is the main challenge users have to overcome, by understanding the meaning behind troublesome words and becoming familiar with them. Complex Word Identification (CWI) is a task that intends to identify hard-to-understand tokens, highlighting them for further clarification and assisting users to grasping the contents of the document.

***Motivation***. Each culture includes exclusive ideas, available only for the ones who can pass the obstacle of language [1]. However, properly understanding language can prove to be a difficult task. By identifying complex words, users can make consistent steps towards adapting to the culture and accessing the knowledge it has to offer. As an example, entries like "*mayoritariamente*" (eng. "mostly") or "*gobernatura*" (eng. "governance") in the Spanish environment can create understanding problems for non-native Spanish speakers [2], thus requiring users to familiarize themselves with these particular terms.

***Challenges***. The identification task becomes increasingly more difficult, as proper complex word identification is not guaranteed. For example, if we use human identification techniques, language learners may consider a new word to be complex, while others might not share the same opinion by relying on their prior knowledge in that language. Therefore, universal annotation techniques are required, such that a ground truth can be established and the same set of words is considered complex in any context.

***Proposed Approach***. We consider state-of-the-art solutions, namely multilingual Transformer-based approaches, to address the CWI challenge. First, we apply a zero-shot learning approach. This was performed by training Recurrent Neural Networks (RNNs) [3] and Transformer-based [4] models on a source language corpus, followed by validating and testing on a corpus from a target language, different from the source language. A second experiment consists of a one-shot learning approach that considers training on each of the three languages (i.e., English, German, Spanish), but only keeping one entry from the target language, and validating and testing on English, German, Spanish, and French, respectively.

In addition, we performed few-shot learning experiments by validating and testing on a language, and training on the others, but with the addition of a small number of training entries from the target language. The model learns sample structures from the language and, in general, performs better when applied on multiple entries. Furthermore, this training process can help the model adapt to situations in which the number of training inputs is scarce. The dataset provided by the CWI Shared Task 2018 [2] was used to perform all experiments.

This paper is structured as follows. The second section describes related work and its impact on the CWI task. The third section describes the corpus and outlines our method based on multilingual embeddings and Transformer-based models, together with the corresponding experimental setup. The fourth section details the results, alongside a discussion and an error analysis. The fifth section concludes the paper and outlines the main ideas, together with potential extensions.

## II. Related Work

Complex word identification was explored in various other studies and underlying approaches can be split into two main categories: monolingual and cross-lingual.

**Monolingual CWI.** The first category implies the usage of the same language for training, testing, and validation

2375-0197/20/$31.00 ©2020 IEEE
DOI 10.1109/ICTAI50040.2020.00067

processes using a supervised approach. Sheang [5] proposed a solution based on Convolutional Neural Networks [6] trained on both word embeddings and handcrafted features. The author used pretrained GloVe word embeddings [7] for representing words from each of the three languages in the dataset. Furthermore, the author engineered a series of morphological features to obtain additional insights into the structure of the entries, features like the number of vowels, word length, and Tf-Idf. At the same time, the author considered a series of linguistic features, alongside morphological ones, by identifying syntactic dependencies between words. However, the presence of these features together with language-specific word embeddings implies a complex training and evaluation process, performed on each language separately and with different configuration setups.

**Cross-lingual CWI.** Cross-lingual transfer has been successfully used in various NLP tasks, for example: machine translation [8], named entity recognition [9], verb sense disambiguation [10], dependency parsing [11], coreference resolution [12], event detection [13], sentence summarization [14], document retrieval [15], irony detection [16], dialogue systems [17], domain-specific tweet classification [18], as well abusive language identification [19].

In addition, cross-lingual approaches were employed in few works on the CWI task. For example, Finnimore et al. [20] extracted cross-lingual features for each considered language (i.e. English, German, Spanish, and French). They concluded that the best features for cross-lingual approaches are represented by the number of syllables, number of tokens, and number of punctuation marks. However, performing this process can prove to be costly, as it requires re-running the model for each additional language in which the user intends to perform complex word identification.

Another approach for cross-lingual CWI employs traditional classification algorithms, such as K-Nearest Neighbors (kNN), Random Forests (RF), or Support Vector Machines (SVMs) [21]. Alongside these algorithms, the authors introduced different sets of language-independent features, ranging from length and frequency, to syntactic features.

Bingel and Bjerva [22] presented both a multi-task learning architecture and an ensemble voting approach, by using feed-forward neural networks and random-forest classifiers. Gooding and Kochmar [23] proposed a sequence labeling approach for CWI. They used 300-dimensional word embeddings for encoding the input words, and fed this input to a Bidirectional Long Short-Term Memory (BiLSTM) [24] network that considered both word and character-level representations. The authors imposed a probability threshold of 0.5 for classifying a word as complex and applied the same rules for phrase-level classification. The authors used an English dataset based on news articles written with different levels of professionalism. Their approach underlines the effectiveness of sequence labeling models which considerably surpassed prior methods by a margin of up to 3.6% in terms of macro F1-score.

Zampieri et al. [25] developed ensemble classifiers to identify complex words. They used two approaches for classification, namely Plurality Voting [26] and Oracle [27]. Based on multiple subsystems, the authors concluded that the latter approach performed well when integrating the top three methods participating in the SemEval CWI 2016 competition [28].

A different approach to CWI was taken by Thomas et al. [29] who considered simplifying the entire document lexicon, thus making the text more accessible for non-native speakers. The authors introduced different algorithms for reducing the lexicon size, by combining disambiguation and lexical reduction steps.

In contrast to the previous approaches, we developed a system based on state-of-the-art NLP solutions (i.e., Transformers), that can efficiently adapt to a large number of languages, without prior setup or feature engineering. The Transformer multi-lingual models are pretrained on a large number of languages, with various word representations already mapped into the same space. Unlike previous work, our models are universal, can be easily extended to other languages, and can be used for transfer learning.

## III. Method

We consider two main multi-lingual approaches for CWI: a) RNN-based solutions, alongside multilingual word embeddings, and b) multilingual Transformers specialized in token classification. Our aim is to infer cross-lingual features of complex words by training or fine-tuning on a labelled corpus containing different languages, followed by the identification of complex words on a newly encountered language. Preprocessing is minimal and considered only the removal of unknown characters, as well as extra spaces from the dataset.

### A. Corpus

Our analysis uses the dataset provided by the CWI Shared Task 2018 [2], which contains entries in four languages, namely: English, German, Spanish, and French. The English section of the dataset contains articles written at three proficiency levels: professional (news), non-professional (WikiNews), and Wikipedia articles. The German and the Spanish sections contain only one category of entries, taken from Wikipedia pages. Quantitatively, the English section contains 27,299 entries for training and 3,328 for validation. In contrast, the German section offers only 6,151 training elements and 795 for validation. At the same time, the Spanish section provides 13,750 training entries and 1,622 validation entries. We note that there are no training and validation entries for the French language.

As expected, the number of complex words is lower when compared to the number of non-complex words. Table I shows the distribution of complex words in the dataset. While the Spanish and English sections contain a relatively large amount of complex or non-complex words, the vocabulary corresponding to the German section is considerably smaller, with only 17,462 words. The small number of German entries is caused by the general focus on English and Spanish, languages with a greater number of speakers when compared to German[1].

[1]https://www.visualcapitalist.com/100-most-spoken-languages/

Additionally, the test dataset also contains French entries, with a total of 4,507 words.

TABLE I
DISTRIBUTION OF COMPLEX WORDS FOR EACH SECTION OF THE CWI SHARED TASK 2018 DATASET.

| Language | Complex Words | Non-complex Words |
|---|---|---|
| English | 14,100 | 59,944 |
| German | 3,478 | 13,984 |
| Spanish | 9,852 | 28,777 |
| French | 867 | 3,640 |

### B. Multilingual Word Embeddings

Our first experiment consists of using a common embedding for all four languages. We selected pretrained FastText [30] embedding for English, German, Spanish and French. However, these embedding spaces are not aligned one with another. Thus, we mapped them into a merged space by using Facebook MUSE [31], a tool that receives as inputs two embedding files and a target vector space, and maps them into the same space. The mapping process consists of learning a rotation matrix $W$, that intends to align the two distributions by using an adversarial learning technique. The matrix $W$ is then refined by using Procrustes transformations because the initial alignment is rough. The transformation consists of setting frequent words aligned in the previous step to anchor points, followed by minimizing an energy function between the anchor points. Finally, an expansion is performed using the matrix $W$ and a distance metric for the space containing a high density of words, such that the distance between unrelated words is increased.

The tool requires a parallel corpus between the languages. The corpus can be created by selecting the desired ground-truth bilingual dictionaries available on the Facebook MUSE repository[2]. The mapping was performed in two steps, as follows. First, we mapped the English and German vectors by using an English-German parallel corpus. Second, we added the Spanish embeddings, by further using an English-Spanish parallel corpus. The obtained embeddings are then fed into a BiLSTM [24] network, alongside a TimeDistributed layer[3]. The experiments were performed in different scenarios: a) a zero-shot approach that required training on combinations of all the available languages, excepting the target language; b) a one-shot approach that introduces the target language (one entry) into the training corpus; and c) a few-shot approach, introducing 100 target language entries in the training dataset.

### C. Multilingual BERT

Multilingual BERT (mBERT) [32] is a pretrained Transformer architecture trained on over 100 languages, which we selected for multi-lingual token classification. The efficiency of representations generated by the model needs to be maximized because we performed our experiments in a multilingual environment. Fortunately, mBERT offers the possibility of splitting its representations into two categories, language-neutral components and language-specific components, thus sharing certain features between the languages of interest. mBERT was fine-tuned for the CWI task by using the previously mentioned zero-shot and one-shot learning approaches.

### D. XLM-RoBERTa

XLM-RoBERTa [33] is also a multilingual model built with the Masked Language Model objective, that should have an advantage over mBERT because it was pretrained on even more multilingual data (approximately 2.5 TB of raw text data). The model obtains state-of-the-art results for the GLUE benchmark tasks [34], while performing extremely well on Named Entity Recognition and Cross-lingual Natural Language Inference tasks [33].

### E. Other BERT-based Monolingual Models

Alongside mBERT, we decided to experiment with models extensively pretrained on each one of our target languages, alternatives that have shown better performance than the multilingual models in other NLP tasks. Thus, we used new models for the German, Spanish and French languages, namely: German BERT[4], Spanish BERT (BETO) [35], and French BERT (CamemBERT) [36]. Our goal was to increase performance by specifically focusing on a certain language, instead of over 100 languages (as the case of mBERT).

### F. Implementation Details

Six experiments were conducted: a) embeddings aligned with MUSE fed to a BiLSTM network, b) mBERT token classification, c) XLM-RoBERTa token classification, d) German BERT token classification, e) BETO token classification, and f) CamemBERT token classification. Each experiment is also divided into sub-experiments that considered the usage of each language individually, as well as all possible combinations of languages in the training set. The four languages (i.e. English, German, Spanish, and French) were considered, by turn, for validation and testing. The BiLSTM-based solution was trained for 5 epochs, while the others (i.e, the Transformer-based solutions) were trained for 4 epochs. We concluded that this setup offers the best results considering that all our solutions start overfitting after 5 and 4 epochs, respectively. Table II presents the hyperparameters used for training the models during the experiments.

TABLE II
EXPERIMENTAL HYPERPARAMETERS.

| Hyperparameter | MUSE + BiLSTM | Transformer |
|---|---|---|
| Optimizer | RMSprop [37] | AdamW [38] |
| Learning rate | *5e-5* | *2e-5* |
| Weight decay | - | 0.01 |
| Adam epsilon | - | *1e-8* |

[2]https://github.com/facebookresearch/MUSE

[3]https://keras.io/api/layers/recurrent_layers/time_distributed/

[4]https://deepset.ai/german-bert

TABLE III
THE MACRO F1-SCORES OF DIFFERENT MODELS ON BOTH VALIDATION AND TEST DATASETS.

| Model | Train | | | Dev | | | | | Test | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ***EN*** | ***DE*** | ***ES*** | ***EN-W*** | ***EN-WN*** | ***EN-N*** | ***DE*** | ***ES*** | ***EN-W*** | ***EN-WN*** | ***EN-N*** | ***DE*** | ***ES*** | ***FR*** |
| MUSE + BiLSTM | ✓ | | | .606 | .582 | .577 | .622 | .609 | .592 | .587 | .579 | .625 | .640 | .524 |
| | | ✓ | | .487 | .602 | .491 | .479 | .474 | .498 | .500 | .498 | .483 | .513 | .494 |
| | | | ✓ | **.610** | **.611** | .599 | **.638** | .635 | **.603** | .590 | **.592** | .602 | .638 | **.546** |
| | ✓ | ✓ | | .598 | .582 | .571 | .628 | **.618** | .585 | .588 | .577 | .774 | **.641** | .516 |
| | ✓ | | ✓ | .603 | .577 | .569 | .627 | .619 | .598 | .580 | .576 | **.626** | .763 | .513 |
| | | ✓ | ✓ | .590 | .586 | **.609** | .637 | .623 | .589 | **.595** | .579 | .688 | .704 | .519 |
| | ✓ | ✓ | ✓ | .604 | .578 | .570 | .626 | .620 | .587 | .581 | .577 | .774 | .751 | .512 |
| mBERT | ✓ | | | .760 | .790 | .734 | .727 | .756 | .768 | .746 | .721 | .731 | **.734** | .653 |
| | | ✓ | | .728 | .746 | .670 | .806 | .744 | .736 | .696 | .630 | .778 | .697 | **.691** |
| | | | ✓ | **.747** | **.763** | **.703** | **.768** | .733 | **.744** | .702 | **.710** | **.755** | .735 | .671 |
| | ✓ | ✓ | | .750 | .787 | .733 | .784 | **.758** | .766 | .753 | .729 | .766 | .730 | .658 |
| | ✓ | | ✓ | .756 | .788 | .751 | .737 | .730 | .764 | .754 | .721 | .739 | .746 | .649 |
| | | ✓ | ✓ | .736 | .759 | .683 | .783 | .734 | .741 | **.709** | .677 | .746 | .737 | .671 |
| | ✓ | ✓ | ✓ | .755 | .789 | .739 | .782 | .740 | .766 | .752 | .730 | .752 | .735 | .684 |
| XLM-RoBERTa | ✓ | | | .793 | .846 | .780 | .757 | .711 | .808 | .811 | .808 | .770 | **.728** | .647 |
| | | ✓ | | .717 | .697 | .695 | .790 | .710 | .716 | .701 | .670 | .795 | .702 | **.702** |
| | | | ✓ | .749 | **.753** | **.717** | .777 | .730 | .760 | **.720** | .730 | .770 | .756 | .701 |
| | ✓ | ✓ | | .795 | .833 | .808 | .801 | **.720** | .806 | .811 | .808 | .801 | .725 | .674 |
| | ✓ | | ✓ | .795 | .823 | .791 | **.789** | .739 | .785 | .801 | .808 | **.782** | .746 | .688 |
| | | ✓ | ✓ | **.750** | .751 | .711 | .809 | .744 | **.774** | .708 | **.731** | .802 | .737 | .666 |
| | ✓ | ✓ | ✓ | .800 | .817 | .780 | .794 | .748 | .798 | .811 | .807 | .534 | .741 | .688 |
| German BERT | ✓ | | | - | - | - | **.712** | - | - | - | - | **.736** | - | - |
| | | ✓ | | - | - | - | .775 | - | - | - | - | .762 | - | - |
| | | | ✓ | - | - | - | .627 | - | - | - | - | .650 | - | - |
| | ✓ | ✓ | | - | - | - | .771 | - | - | - | - | .770 | - | - |
| | ✓ | | ✓ | - | - | - | .701 | - | - | - | - | .717 | - | - |
| | | ✓ | ✓ | - | - | - | .777 | - | - | - | - | .764 | - | - |
| | ✓ | ✓ | ✓ | - | - | - | .771 | - | - | - | - | .775 | - | - |
| BETO | ✓ | | | - | - | - | - | .603 | - | - | - | - | **.656** | - |
| | | ✓ | | - | - | - | - | .525 | - | - | - | - | .580 | - |
| | | | ✓ | - | - | - | - | .733 | - | - | - | - | .731 | - |
| | ✓ | ✓ | | - | - | - | - | **.652** | - | - | - | - | .649 | - |
| | ✓ | | ✓ | - | - | - | - | .728 | - | - | - | - | .738 | - |
| | | ✓ | ✓ | - | - | - | - | .730 | - | - | - | - | .731 | - |
| | ✓ | ✓ | ✓ | - | - | - | - | .720 | - | - | - | - | .733 | - |
| CamemBERT | ✓ | | | - | - | - | - | - | - | - | - | - | - | .563 |
| | | ✓ | | - | - | - | - | - | - | - | - | - | - | .442 |
| | | | ✓ | - | - | - | - | - | - | - | - | - | - | .604 |
| | ✓ | ✓ | | - | - | - | - | - | - | - | - | - | - | .592 |
| | ✓ | | ✓ | - | - | - | - | - | - | - | - | - | - | .670 |
| | | ✓ | ✓ | - | - | - | - | - | - | - | - | - | - | .669 |
| | ✓ | ✓ | ✓ | - | - | - | - | - | - | - | - | - | - | **.683** |

* We considered: EN-W = English-Wikipedia; EN-WN = English-WikiNews; EN-N = English-News; DE = German; ES = Spanish; FR = French.

## IV. RESULTS

Table III contains the macro F1-scores obtained on the CWI validation and test datasets for each experiment and for each combination of training languages. Table III contains monolingual and zero-shot learning experiments. The best results for the zero-shot approach are marked in bold, while the best results for the monolingual approach are underlined.

### A. Zero-Shot Transfer Evaluation

The best results on both validation and test datasets for the zero-shot learning strategy are obtained using the XLM-RoBERTa model, with a single exception represented by the validation dataset on German. With a considerable margin when compared to its counterparts, XLM-RoBERTa fine-tuned on English and Spanish manages to obtain a macro F1-score of 0.782 on the German test dataset, compared to 0.626 (MUSE+BiLSTM), 0.739 (mBERT), and 0.717 (German BERT). The results are similar for the Spanish and English test datasets (Wikipedia, WikiNews, News) having macro F1-values of 0.702 and 0.774, 0.720, and 0.731, respectively. The increased performance of XLM-RoBERTa can be attributed to the larger corpus it was pretrained on, a clear advantage over other BERT-based solutions. However, if we look at the other BERT-based monolingual models (i.e. German BERT,

TABLE IV
RESULTS ON THE TEST DATASET USING ONE-SHOT AND FEW-SHOT LEARNING.

| Model | Train | | | Macro F1-score (one-shot) | | | | | Macro F1-score (few-shot) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *EN* | *DE* | *ES* | *EN-W* | *EN-WN* | *EN-N* | *DE* | *ES* | *EN-W* | *EN-WN* | *EN-N* | *DE* | *ES* |
| mBERT | ✓ | | | - | - | - | .732 | .723 | - | - | - | .727 | **.738** |
| | | ✓ | | .730 | .684 | .654 | - | .712 | .730 | .688 | .671 | - | .709 |
| | | | ✓ | .741 | **.711** | **.700** | **.743** | - | **.742** | **.691** | **.690** | .740 | - |
| | ✓ | ✓ | | - | - | - | - | **.730** | - | - | - | - | .719 |
| | ✓ | | ✓ | - | - | - | .741 | - | - | - | - | **.768** | - |
| | | ✓ | ✓ | **.751** | .697 | .678 | - | - | .741 | .697 | .663 | - | - |
| XLM-RoBERTa | ✓ | | | - | - | - | .769 | **.732** | - | - | - | .760 | **.730** |
| | | ✓ | | .734 | .688 | .643 | - | .693 | .735 | .691 | .695 | - | .703 |
| | | | ✓ | **.761** | **.731** | **.714** | .779 | - | **.761** | **.733** | **.726** | **.766** | - |
| | ✓ | ✓ | | - | - | - | - | .724 | - | - | - | - | .722 |
| | ✓ | | ✓ | - | - | - | **.783** | - | - | - | - | .765 | - |
| | | ✓ | ✓ | .756 | .723 | .679 | - | - | .755 | .703 | .716 | - | - |
| German BERT | ✓ | | | - | - | - | .699 | - | - | - | - | **.736** | - |
| | | | ✓ | - | - | - | .649 | - | - | - | - | .676 | - |
| | ✓ | | ✓ | - | - | - | **.734** | - | - | - | - | .689 | - |
| BETO | ✓ | | | - | - | - | - | .650 | - | - | - | - | **.686** |
| | | ✓ | | - | - | - | - | .603 | - | - | - | - | .545 |
| | ✓ | ✓ | | - | - | - | - | **.693** | - | - | - | - | .680 |

✓implies the usage of the entire dataset corresponding to that language. Additionally, we randomly selected 1 (for one-shot learning) or 100 (for few-shot learning) training entries from the language corresponding to the result for that line.

BETO, and CamemBERT), we can see that their performance is surpassed by both mBERT and XLM-RoBERTa. These models are pretrained on a main language, and fine-tuning them on different languages can lead to poorer results, as seen in Table III. For example, the difference in performance (macro F1) between XLM-RoBERTa and BETO is of 6.8% on the Spanish validation dataset, a significant discrepancy for a CWI task.

### B. One-Shot Transfer Evaluation

Furthermore, the best values for the one-shot learning approach are marked with bold in Table IV, where we considered only one training entry corresponding to the language of the result. We can observe that, again, the XLM-RoBERTa model offers the best performance. For example, XLM-RoBERTa obtains a macro F1-score of 0.731 on the WikiNews dataset, compared to 0.711 for mBERT. Moreover, the large difference is maintained for the German language as well, with a result of 0.783 versus 0.743. However, the scores for the Spanish language are closer, with a value around 0.730 for both models.

### C. Few-Shot Transfer Evaluation

Next, we included a small number of train entries (i.e., 100) from the same language as the test dataset because we intended to further improve the scores obtained by the Transformer-based solution using the zero-shot learning scenario. Using this approach, the model can infer characteristics of the target language and may perform better when identifying complex words on a wide range of different test entries.

Table IV contains the results obtained in the few-shot learning experiments. Unexpectedly, the models perform slightly worse. This phenomenon can be attributed to the models' incapacity to grasp the main language characteristics, as well as the representations of a complex word, given a small number of training entries.

To conclude, our solution manages to outperform state-of-the-art alternatives on five out of six cross-lingual entries, the only exception being the French language (see Table V). Furthermore, our solution manages to surpass state-of-the-art results for German in the monolingual setup, even though it was created for cross-lingual experiments.

TABLE V
CROSS-LINGUAL AND MONOLINGUAL STATE-OF-THE-ART RESULT COMPARISON WITH OUR PERFORMANCE ON THE TEST DATASET.

| | EN-W | EN-WN | EN-N | DE | ES | FR |
|---|---|---|---|---|---|---|
| Cross-lingual SotA [20] | .652 | .638 | .659 | .734 | .726 | **.758** |
| *Our best solution, zero-shot learning* | **.774** | .720 | **.731** | **.782** | **.734** | .702 |
| *Our best solution, few-shot learning* | .761 | **.733** | .726 | .766 | .730 | - |
| Monolingual SotA [5] | **.811** | **.840** | **.874** | .759 | **.797** | - |
| *Our best monolingual solution* | .808 | .811 | .808 | **.795** | .756 | - |

### D. Error Analysis

Most misclassifications occurred in the English News test dataset, where our models yielded a maximum F1-macro score of 0.733 by using a few-shot learning approach with XLM-RoBERTa. The high number of wrongly categorized tokens can be attributed to the complexity of the dataset, written in a more formal manner, adequate for news articles. This complexity implies the presence of more sophisticated words (e.g., "underwriter") that are not present in the training dataset, thus causing the model to wrongly classify them. In addition, the dataset contains news with series of location names (e.g. "Londonderry") or composed notions (e.g. "better-optimized", "android-running", "java-related") that, once again, are not included in the training set.

At the same time, another aspect that influences the classification performance is represented by the annotators' subjectivity. In certain circumstances, words may not be considered complex (e.g. "with", "connection", "been") in the training set, while they are marked as complex in the test dataset. Similar situations also occur in the English Wikipedia, English WikiNews, German and Spanish datasets, with a series of tokens that either are not present in the training dataset, or have different labels between them.

## V. Conclusions and Future Work

Complex Word Indentification is a challenging task, even when using state-of-the-art Transformer-based solutions. In this work, we introduce an approach that improves the previous results on the cross-lingual and monolingual CWI shared task 2018 by using multilingual and language-specific Transformer models, multilingual word embeddings (non-Transformer), and different fine-tuning techniques. Fine-tuning a model on data from two different languages creates the opportunity of grasping features that empower it to better recognize complex words in certain contexts, even in a different language. In addition, zero-shot, one-shot, and few-shot learning strategies provide good results, surpassing strong baselines [20] and proposing an alternative to help non-native speakers to properly understand the difficult aspects of a certain language.

For future work, we intend to improve our results on the monolingual tasks by integrating additional models, such as XLNet [39] and techniques like adversarial training [40] and multi-task learning [41]. Furthermore, we intend to experiment with other pretraining techniques specific to Transformer models, such that the results for French can benefit from cross-lingual transfer learning.

## Acknowledgments

This work was supported by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

## References

[1] J. Liu and F. G. Fang, "Perceptions, awareness and perceived effects of home culture on intercultural communication: Perspectives of university students in china," *System*, vol. 67, pp. 25–37, 2017.

[2] S. M. Yimam, C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri, "A report on the complex word identification shared task 2018," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 66–78, 2018.

[3] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, Mar 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998–6008, 2017.

[5] K. C. Sheang, "Multilingual complex word identification: Convolutional neural networks with morphological and linguistic features," in *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pp. 83–89, 2019.

[6] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *ArXiv e-prints*, 11 2015.

[7] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[8] Y. Kim, Y. Gao, and H. Ney, "Effective cross-lingual transfer of neural machine translation models without shared vocabularies," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1246–1257, 2019.

[9] Z. Liu, G. I. Winata, P. Xu, and P. Fung, "Coach: A coarse-to-fine approach for cross-domain slot filling," *arXiv preprint arXiv:2004.11727*, 2020.

[10] S. Gella, D. Elliott, and F. Keller, "Cross-lingual visual verb sense disambiguation," 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)*, pp. 1998–2004, 2019.

[11] W. Ahmad, Z. Zhang, X. Ma, E. Hovy, K.-W. Chang, and N. Peng, "On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing," 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)*, pp. 2440–2452, 2019.

[12] G. Urbizu, A. Soraluze, and O. Arregi, "Deep cross-lingual coreference resolution for less-resourced languages: The case of basque," in *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pp. 35–41, 2019.

[13] V. D. Lai, F. Dernoncourt, and T. H. Nguyen, "Extensively matching for few-shot learning event detection," *arXiv preprint arXiv:2006.10093*, 2020.

[14] X. Duan, M. Yin, M. Zhang, B. Chen, and W. Luo, "Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3162–3172, 2019.

[15] R. Zhang, C. Westerfield, S. Shim, G. Bingham, A. R. Fabbri, W. Hu, N. Verma, and D. Radev, "Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3173–3179, 2019.

[16] B. Ghanem, J. Karoui, F. Benamara, P. Rosso, and V. Moriceau, "Irony detection in a multilingual context," in *European Conference on Information Retrieval*, pp. 141–149, Springer, 2020.

[17] S. Schuster, S. Gupta, R. Shah, and M. Lewis, "Cross-lingual transfer learning for multilingual task oriented dialog," 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)*, pp. 3795–3805, 2019.

[18] J. R. Chowdhury, C. Caragea, and D. Caragea, "Cross-lingual disaster-related multi-label tweet classification with manifold mixup," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 292–298, 2020.

[19] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 363–370, 2019.

[20] P. Finnimore, E. Fritzsch, D. King, A. Sneyd, A. U. Rehman, F. Alva-Manchego, and A. Vlachos, "Strong baselines for complex word identification across multiple languages," 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)*, pp. 970–977, 2019.

[21] S. M. Yimam, S. Štajner, M. Riedl, and C. Biemann, "Multilingual and cross-lingual complex word identification," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, (Varna, Bulgaria), pp. 813–822, INCOMA Ltd., Sept. 2017.

[22] J. Bingel and J. Bjerva, "Cross-lingual complex word identification with multitask learning," in *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, (New Orleans, Louisiana), pp. 166–174, Association for Computational Linguistics, June 2018.

[23] S. Gooding and E. Kochmar, "Complex word identification as a sequence labelling task," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1148–1153, 2019.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] M. Zampieri, S. Malmasi, G. Paetzold, and L. Specia, "Complex word identification: Challenges in data annotation and system performance," in *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pp. 59–63, 2017.

[26] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[27] L. I. Kuncheva, J. C. Bezdek, and R. P. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern recognition*, vol. 34, no. 2, pp. 299–314, 2001.

[28] G. Paetzold and L. Specia, "Semeval 2016 task 11: Complex word identification," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560–569, 2016.

[29] S. R. Thomas and S. Anderson, "Wordnet-based lexical simplification of a document.," in *KONVENS*, pp. 80–88, 2012.

[30] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[31] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[32] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, 2019.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[34] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

[35] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," in *Practical ML for Developing Countries Workshop@ ICLR 2020*, 2020.

[36] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," *arXiv preprint arXiv:1911.03894*, 2019.

[37] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *Advances in neural information processing systems*, pp. 5753–5763, 2019.

[40] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu, "Freelb: Enhanced adversarial training for natural language understanding," in *International Conference on Learning Representations*, 2019.

[41] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.