Identifying Semantic Argument Types in Predication and Copredication Contexts: A Zero-Shot Cross-Lingual Approach

Deniz Ekin Yavas[1], **Laura Kallmeyer**[1], **Rainer Osswald**[1],
Elisabetta Jezek[2], **Marta Ricchiardi**[3], **Long Chen**[1]
Heinrich Heine University Düsseldorf[1], University of Pavia[2,3]
{deniz.yavas, laura.kallmeyer, rainer.osswald, chen.long}@hhu.de[1],
jezek@unipv.it[2], marta.ricchiardi01@universitadipavia.it[3]

Abstract

Identifying semantic argument types in predication contexts is not a straightforward task for several reasons, such as inherent polysemy, coercion, and copredication phenomena. In this paper, we train monolingual and multilingual classifiers with a zero-shot cross-lingual approach to identify semantic argument types in predications using pre-trained language models as feature extractors. We train classifiers for different semantic argument types and for both verbal and adjectival predications. Furthermore, we propose a method to detect copredication using these classifiers through identifying the argument semantic type targeted in different predications over the same noun in a sentence. We evaluate the performance of the method on copredication test data with Food•Event nouns for 5 languages.

1 Introduction

This paper is concerned with the question of how to automatically decide which semantic type is targeted in predications over nouns. In our case, the predicate can be a verb or an adjective. This question is particularly interesting in cases where complex type nouns[1] are arguments of predications. But even with nouns that, lexically, have only a single type, the predication can target a different type and thereby trigger a coercion in the noun (Pustejovsky, 1991). Examples are given in (1). In both (1-a) as well as (1-b), the respective predicates target one of the two types of a complex type noun (a *dinner* is inherently both an Event and a Food item). In (1-c), the noun is a simple type noun (*soup* is only of type Food), and its type is targeted in the predication. The predication in (1-d) involves a coercion since it targets a type that is different from the lexical type of the noun. Finally, for complex type nouns, we can have cases where different component types of the same noun are targeted, either by different predicates as in (1-e) or by a single predicate as in (1-f) where *book* is a physical object and an informational content at the same time and the predicate targets both. The first case is an instance of copredication (see below).

(1) a. They chose the *vegetarian dinner*.
→ (target: *Food*)
b. I *organized* a *dinner* for them.
→ (target: *Event*)
c. I *ate* my *soup*.
→ (target: *Food*)
d. I *finished* my *soup*.
→ (target: *Event*, coercion)
e. They *organized* a *vegetarian dinner*.
→ (target: *Event* and *Food*)
f. He *wrote* a lot of *books*.
→ (target: *Phys_Obj•Information*)

Our main goal is to develop classifiers that, given a predicate and an argument noun in their sentential context, decide whether a specific type has been targeted. Furthermore, we exploit the cross-lingual transfer potential of multilingual *pre-trained language models* (LMs) in order to apply this task to different languages without the need of labelled data for all of them.

One interesting application of such classifiers is the detection of instances of copredication with complex type nouns. Copredication is a general term defining a "grammatical construction in which two predicates jointly apply to the same argument" (Asher, 2011, p. 11). We are interested in a specific type of copredication where two predicates that require different semantic types apply to the same noun (Pustejovsky, 1995; Pustejovsky and Jezek, 2008; Asher, 2011). For example, given the occur-

[1] Also "dot object" nouns (Pustejovsky, 1995), "nouns with facets" (Cruse, 1995), "dual aspect nouns" (Asher, 2011) in the literature.

Proceedings of Recent Advances in Natural Language Processing, pages 310–320
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_035

rence of a complex type noun such as *dinner* in a sentence where we have two (or more) predications over that noun, we want to decide whether these predications target different types, as for instance in (1-e). We will apply the classifiers developed in this paper to this task, using the complex type *Food•Event* as a test case.

We start by investigating whether it is possible to train classifiers for both verbal and adjectival predications for this purpose using LMs (BERT (Devlin et al., 2019), mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020a)) as feature extractors. In addition, we investigate whether it is possible to train multilingual classifiers with a zero-shot cross-lingual approach by training the classifiers on one language with the extracted embeddings of multilingual language models and applying them to other languages.

We train monolingual classifiers for Italian using the extracted embeddings of a monolingual BERT model[2] and the multilingual ones using the embeddings of the multilingual models mBERT and XLM-RoBERTa. We start with Italian as our source language due to the availability of annotated data in the T-PAS (Typed Predicate Argument Structures) resource (Jezek et al., 2014). We train classifiers for verbal predications for the semantic types *Human, Information, Event, Artifact,* and *Location* and adjectival predications for the semantic types *Event, Artifact,* and *Information*. The selection of these types is intended to capture the diversity of the semantic type hierarchy.

Finally, we apply the verbal and adjectival classifiers for *Artifact*[3] and *Event* semantic types to the sentences containing *Food•Event* nouns in order to detect certain copredication patterns, as in (1-e), in which a verb and an adjective predicate over the same noun. We evaluate the proposed model on test data for a set of typologically diverse languages; Chinese, English, German, Italian, and Turkish.[4]

2 Related Work

2.1 Selectional Preference and Semantic Type Knowledge of LMs

There is no study to our knowledge that aims at exploiting LMs for a selectional preference task, nor that investigates the transferability of selectional preference knowledge to other languages using multilingual LMs. However, there are studies that investigate the LMs' knowledge about selectional preferences of verbs and semantic types. Their findings suggest that contextual language models encode information about the selectional preferences of verbs (Metheniti et al., 2020; Chersoni et al., 2021; Li et al., 2021; Pedinotti et al., 2021) and the semantic type of the nouns in general (Zhao et al., 2020). Similar to our study, Zhao et al. (2020) and Chersoni et al. (2021) trained classifiers using the extracted representations of the BERT model for their tasks and these classifiers achieved high accuracy scores.

2.2 Zero-Shot Cross-Lingual Transfer using multilingual LMs

Several studies have investigated the performance of multilingual LMs for zero-shot cross-lingual transfer on a variety of tasks, e.g. NER, POS, NLI, QA. They show that these models are effective for this purpose (Pires et al., 2019; Conneau et al., 2020a; Wu and Dredze, 2020; Aghazadeh et al., 2022). It is also shown that these models perform well on multilingual benchmarks such as XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020). Additionally, these papers show that XLM-RoBERTa performs better than mBERT (Conneau et al., 2020a; Hu et al., 2020; Liang et al., 2020; Lauscher et al., 2020).

Recent research has also investigated the effects of language differences in cross-lingual transfer. It has been shown that structural similarity, such as word order or typological similarity affects the transfer (Pires et al., 2019; Conneau et al., 2020b; K et al., 2020; Lauscher et al., 2020; Deshpande et al., 2022). The difference in the language script is also shown to be important, but only when the word order differs as well (Pires et al., 2019; Deshpande et al., 2022).

2.3 Copredication Detection

Jezek and Vieu (2014) adopt a semi-automatic approach for extracting copredications of *Physical_Object•Information* nouns with a verb and an adjective in Italian. First, they manually select a list of predicates for both semantic types: Physical Object and Information. Then, they construct copredication contexts with different predicate combinations and extract examples by searching the corpus for these contexts. As an extension to the

[2]See Appendix A for information about the models, parameters and libraries used for the experiments.

[3]Artifact is the supertype of Food in the T-PAS semantic type hierarchy.

[4]Datasets and code are available at: https://github.com/yavasde/predication-classification

previous study, Vieu et al. (2015) use a latent semantic distributional model in order to select the predicates to avoid the manual process of predicate selection. Compared to these studies, our method is automatic and does not rely on the classification of each predicate, which can be problematic due to their polysemous nature, but relies on the classification of each predication instance. This also allows using this method cross-linguistically without knowledge specific to each language.

3 Method

3.1 Predication Classifiers

We first show that it is possible to train a classifier for the identification of the semantic argument type targeted by a predicate in a specific predication context using the extracted representations of LMs. Furthermore, we aim to investigate whether this knowledge is transferable from one language to another with a zero-shot cross-lingual approach by training classifiers on the source language using the multilingual representations of the multilingual LMs and applying the trained models to the target languages.

For all semantic types and predication types, we use LMs as feature extractors and train monolingual and multilingual classifiers with the SVM algorithm using the extracted embeddings of the models.[5] For monolingual Italian classifiers, we use a BERT model for Italian and for multilingual classifiers, we use the multilingual LMs mBERT and XLM-RoBERTa. We train binary classifiers for each of the semantic types *Artifact, Event, Human, Information* and *Location* for verbal predications and *Artifact, Event,* and *Information* for adjectival predications.[6]

We use the contextualized embeddings of the predicate and the argument in a specific sentence as input for the classifiers. First, we tokenize each sentence with the model tokenizer and give the tokenized sentence as input to the model. Then, we extract the embeddings of the predicate (verb/adj) and the argument (direct object/noun) from the last 4 layers of the model output and average them to create one representation for each item.[7] We use only the last 4 layers because higher layers are more specialized in semantics-related tasks (Liu et al., 2019; Tenney et al., 2019; Zhao et al., 2020). We formalize the task as a relation classification problem where we classify the relation between the predicate and its argument. For this purpose, we concatenate the embeddings of the predicate and the argument and use the final embedding as the input for the classifiers.[8]

3.2 Copredication Detection

In order to detect copredication, the classifiers are applied to the sentences with complex type nouns, where both syntactic types of predications are available for the same noun. First, sentences are parsed using the Stanza library (Qi et al., 2020) in order to identify the predications in sentences. Then, the embeddings of the predicate-argument pairs are extracted from the LM, concatenated and given to the relevant classifiers (verb/adj). Copredication is considered detected if both verbal and adjectival predication classifiers of different semantic types classify the predications as positive.

4 Training Classifiers

4.1 Data

4.1.1 Verbal Predication Classifiers

Training data. We use T-PAS (Typed Predicate Argument Structures; Jezek et al., 2014) as our primary resource. T-PAS provides corpus-derived argument structure patterns for Italian verbs with manually annotated semantic argument types; e.g [Human] mangiare [Food] (*Eng.: [Human] eat [Food]*). Each verb pattern has matching corpus instances extracted from the itWac corpus (Baroni et al., 2009).

In T-PAS, semantic types are organized in a hierarchy. For each semantic type (*Human, Event, Information, Artifact, Location*), we extract sentences whose verbs take direct objects with the target semantic type or a subtype of it.

The training negatives are also selected from T-PAS from the semantic types other than the target semantic type's supertypes, subtypes, or the semantic type itself. The negatives are downsized to make their size equal to the positive samples.

[5] Even though we tested several classification algorithms, we used SVM for the final experiments because it performed best. See Appendix B for the detailed comparison.

[6] We use binary classifiers instead of a multiclass one because there are predicates that can target both semantic types as in example (1-f).

[7] In the cases in which the target words are tokenized into subwords by the model tokenizer, only the first subword is taken into account.

[8] Fine-tuning is the most standard way to use LMs for token or sentence classification but it is not that straightforward to fine-tune the models for relation classification.

Types	Data Size						Model	Languages					
	Training	Test											
		it	de	en	tr	zh		it	de	en	tr	zh	Avg.
Verbal Predication													
Arti.	522	258	248	236	220	182	B	0.95 (0.94)	-	-	-	-	-
							mB	0.92 (0.90)	0.84 (0.75)	0.92 (0.91)	0.75 (0.74)	0.83 (0.87)	0.85 (0.83)
							XR	0.90 (0.92)	0.86 (0.83)	0.93 (0.92)	0.88 (0.84)	0.92 (0.91)	**0.89** (0.88)
Event	643	317	258	268	276	256	B	0.95 (0.95)	-	-	-	-	-
							mB	0.94 (0.93)	0.88 (0.90)	0.94 (0.93)	0.86 (0.88)	0.90 (0.89)	0.90 (0.90)
							XR	0.94 (0.95)	0.89 (0.88)	0.94 (0.93)	0.91 (0.88)	0.91 (0.92)	**0.91** (0.91)
Hum.	292	144	130	128	126	74	B	0.92	-	-	-	-	-
							mB	0.92	0.93	0.98	0.84	0.89	0.91
							XR	0.94	0.98	0.98	0.96	0.94	**0.96**
Info.	176	88	82	86	86	70	B	0.98	-	-	-	-	-
							mB	0.98	0.80	0.98	0.84	0.90	0.90
							XR	0.97	0.98	0.95	0.95	0.97	**0.96**
Loc.	321	159	148	148	142	132	B	0.95	-	-	-	-	-
							mB	0.92	0.91	0.94	0.71	0.93	0.88
							XR	0.95	0.97	0.97	0.89	0.93	**0.94**
Adjectival Predication													
Arti.	252	3680	-	148	-	-	B	0.84 (0.84)	-	-	-	-	-
							mB	0.90 (0.91)	-	0.93 (0.93)	-	-	**0.91** (0.92)
							XR	0.87 (0.85)	-	0.93 (0.92)	-	-	0.90 (0.88)
Event	564	1676	-	148	-	-	B	0.88 (0.89)	-	-	-	-	-
							mB	0.86 (0.88)	-	0.76 (0.77)	-	-	0.81 (0.82)
							XR	0.81 (0.82)	-	0.84 (0.83)	-	-	**0.82** (0.82)
Info.	132	2536	-	78	-	-	B	0.91	-	-	-	-	-
							mB	0.90	-	0.94	-	-	**0.92**
							XR	0.91	-	0.89	-	-	0.90

Table 1: The data size and the test results of each classifier. F1 scores are given. The results of the cross-linguistically best-performing classifiers are given in bold. *T-PAS+CT* results are given in parentheses.

To this end, the sentences are clustered with K-Means algorithm using the Scikit-learn library and an equal number of sentences are selected from each cluster. This undersampling method is chosen to have a balanced representation of the negatives.

The selected sentences for both positives and negatives are parsed with the spacy-udpipe Python library[9] in order to identify and annotate the verb and the direct object in each sentence.[10]

Cross-lingual test data. The test data is selected by splitting the data (test size %33) extracted from T-PAS. The data are then machine translated using DeepL API[11] to the other languages.

It is required that the verbs and objects are correctly identified in the translations. For this purpose, they are translated out-of-context and searched for in the sentences. Additionally, the translations are parsed using the Stanza library and all the verb-object pairs in the sentences are extracted through their dependency labels in order to find the correct pairs. However, sometimes, the pairs are not found automatically, in which case they are manually annotated.

In a final step, all translated sentences are manually checked and corrected by (near-)native speakers of the respective languages following the guideline presented in Appendix C. Sentences that can not be corrected are eliminated. Equal numbers of negatives and positives are selected for each dataset. The resulting data numbers for each language are given in Table 1.

4.1.2 Adjectival Predication Classifiers

Training data. The training data for the adjectival predication classifiers are generated using *Masked Language Modeling* (MLM) with BERT due to the unavailability of annotated data. We generate data for 3 semantic types *Artifact, Event* and *Information* using the verbal predication datasets for these types as the basis. We insert an adjective that is predicted by the model into the sentences in order to modify the direct object. The assumption is that in sentences where the verbal predication over the objects targets a certain type, the adjectives predicted by the model with a high probability score will do so as well.

First, a mask is inserted after the noun, and then the Italian BERT is made to predict a word instead of the mask. Only word predictions over a certain confidence score (0.15) are selected from the model predictions. For the final step, the predicted word is inserted in place of the mask and the resulting sentence is parsed with the spaCy library[12] to check

[9] Available at: https://spacy.io/universe/project/spacy-udpipe

[10] Sentence-level annotations are not provided in T-PAS.

[11] Available at: https://www.deepl.com/

[12] Available at: https://spacy.io/

if the relation between the noun and the predicted word is the desired one (*adjectival modification*). The sentences that meet these conditions are used for the training of the classifiers.[13]

Cross-lingual test data. Since the adjective data is generated, we do not test the performance of the classifiers on this data but on manually constructed data for Italian and English.

The test data for Italian are created by extracting corpus instances from the itWac corpus, identified through a concordance search for the most typical 5-10 lexical items that express each type in corpus instances and their respective most frequent adjective modifiers. The sentences are extracted for 3 semantic types (*Artifact, Event* and *Information*) and the negatives of the test data are selected from the sentences of the other 2 semantic types.

The test data for English are also constructed by extracting corpus examples. First, good representatives of each semantic type noun are selected based on their occurrence in the T-PAS data; these are the nouns that only occur in the target semantic type data and occur more than once. As the next step, we translate the selected nouns to English and extract sentences with these nouns from the ukWac corpus (Baroni et al., 2009) but only consider the ones where the noun is the direct object of a verb and also have a token size between 3 and 20. We parse the sentences using the Stanza library and select the ones where there is an adjective that modifies the noun. Finally, we manually select the sentences with good examples of adjectives. The semantic types are the same as for Italian and the negatives are constructed similarly.

Both the test and training data are balanced in terms of the number of positives and negatives. The data size for adjectival predication classifiers can be seen in Table 1.

4.2 Experiments and Results

We test the monolingual classifiers on Italian test data ('B' for monolingual Italian BERT based classifiers) and the multilingual classifiers on the cross-lingual test data ('XR' for XLM-RoBERTa and 'mB' for mBERT based classifiers). F1 score is used as the metric and cross-lingual performance is evaluated by comparing the average f1 score on cross-lingual test data, see Table 1 for the results. The detailed results with precision and recall scores can be found in the Appendix D. A language-specific evaluation is given in Appendix E.

Overall results. The monolingual classifiers perform very well on the task. Each monolingual verbal predication classifier achieves over 0.92 f1 score and each monolingual adjectival predication classifier achieves over 0.84 f1 score. Similarly, all multilingual verbal predication classifiers achieve over 0.85 average f1 scores for all languages and all multilingual adjectival predication classifiers achieve over 0.81 average f1 scores for English and Italian. Overall, XR-classifiers perform better than mB-classifiers (See Table 1).

Monolingual vs. multilingual. The comparison of the monolingual and multilingual classifiers' performances on Italian test data shows that on average, the monolingual classifiers perform better than the multilingual ones on the source language test data. However, for some semantic types, such as *Human* (verb) and *Artifact* (adj), XR-classifiers perform better than the monolingual classifiers. (See Table 1 for the individual results and Figure 1 for the average for verbs.)

Verbal vs. adjectival predications. Overall, the performance of the verbal predication classifiers is better than the adjectival predication classifiers. Contrary to verbal predication classifiers, mB-classifiers perform better than XR-classifiers for adjectives overall. However, the performance difference is smaller.

5 Copredication Detection

5.1 Classifiers for Complex Type Nouns

Even though T-PAS is not necessarily a resource with simple type nouns, the number of sentences with Food•Event nouns is low in our datasets.[14] Since our task is to detect copredication with complex type nouns, we require classifiers that can disambiguate the meanings of these nouns.

In order to address this, we add, to each classifier's training data, additional data with complex type nouns, in which only one type of the noun is targeted; Food or Event as in (1-a) and (1-b). We add the additional data to the training data of Artifact and Event classifiers for both verbs and adjectives. The original classifiers will be referred to as 'T-PAS' and the latter as 'T-PAS+CT'.

[13] The original adjectives in the sentences are replaced by model-predicted ones, in order to avoid copredication instances in the training data.

[14] There are 2 sentences in the Artifact and 3 in the Event dataset.

Training data with complex type nouns. Additional training is obtained by extracting the sentences of Food•Event nouns with Food or Event predications from corpus. First, we determine the best predicates for each type of predication; best food verbs, event adjectives, etc. For this, we use our datasets. We extract the predicates from each semantic type dataset (Artifact and Event) and select the predicates that occur more than once and that only occur in the target semantic type dataset. In the second step, we select 9 Food•Event nouns (see Appendix F for the selected nouns) and we extract the sentences of these nouns with the selected predicates from the itWac corpus. Finally, we add the complex type sentences both to the positives and negatives of Artifact and Event training data for verbs and adjectives with the amount of 20%.

Training results. The performance of the T-PAS+CT classifiers on the test data can be seen in Table 1. Their performance is close to the T-PAS classifiers overall, with some slight differences for some semantic types and languages.

5.2 Evaluation

We apply both semantic type classifiers (Artifact and Event) to classify the verbal and adjectival predications in the sentences of the test data. We investigate how often copredication is detected both in the positives and negatives of the test data. However, we do not consider the correct classification of individual predications in this evaluation method. We use an additional evaluation method to investigate how often the predications are identified correctly.

We test both T-PAS and T-PAS+CT classifiers on the cross-lingual copredication test data comprising 5 languages; Chinese, English, German, Italian, and Turkish. We use monolingual classifiers for Italian and XR-classifiers for other languages since they performed better on single predication classification overall.

Additionally, we investigate the effects of the complex type nouns on copredication detection. We do that by comparing the performance of the method on two types of negatives: negatives with simple type nouns and complex type nouns.

5.2.1 Evaluation Data

The test data is manually created for Italian and machine translated into Chinese, English, German, and Turkish. The translations are manually corrected by (near-)native speakers of the respective languages. A similar correction procedure is applied to the test data, following the data correction guidelines in Appendix C.

The test data contains 30 positive and 24 negative examples of copredication with different semantic types (for more details, see Appendix G). In the positives, verbs and adjectives target different types of Food•Event nouns, whereas in the negatives, both predicates target the same type of Food•Event nouns (either Event or Food). An example of positives is given in (1-e), where the verb 'organize' targets the Event type and the adjective 'vegetarian' targets the Food type. As an example of the negatives, in (2-a), both the verb 'eat' and the adjective 'cold' target the Food type.

We prepare additional data for negatives with simple type nouns. We do this by substituting the Food•Event nouns in the negatives with a Food or Event simple type noun as in (2-b) (see Appendix G for more details).

(2) a. It's depressing to *eat* a *cold lunch*.
b. It's depressing to *eat* a *cold soup*.

5.2.2 Results

We evaluate the results using three metrics; *sensitivity* (recall), to measure the ability to detect the positives and *specificity*, to measure the ability to detect the negatives, and finally, the geometric mean of sensitivity and specificity, for the overall performance. The results can be seen in Table 2.

Lang.	Classifier	Scores		
		Sens.	Spec.	*g*
it	T-PAS	0.66	0.35 (0.79)	0.48
	T-PAS+CT	0.46	0.62 (0.87)	**0.53**
de	T-PAS	0.66	0.25 (0.83)	0.40
	T-PAS+CT	0.53	0.58 (0.91)	**0.55**
en	T-PAS	0.83	0.29 (0.79)	0.49
	T-PAS+CT	0.70	0.66 (0.83)	**0.67**
tr	T-PAS	0.76	0.25 (0.75)	0.43
	T-PAS+CT	0.53	0.45 (0.87)	**0.48**
zh	T-PAS	0.82	0.59 (0.83)	**0.69**
	T-PAS+CT	0.68	0.65 (0.83)	0.66
Random Baseline		*0.25*	*0.25*	*0.25*

Table 2: Performance on the cross-lingual copredication test data. *g* stands for the geometric mean of specificity and sensitivity. The results of the classifiers with the best overall performance are given in bold. The specificity scores in the parenthesis refer to the specificity over simple type nouns.

Overall. T-PAS classifiers achieve higher sensitivity scores compared to specificity scores for all languages. Even though the sensitivity scores achieve

0.80, specificity scores are around the random baseline for most of the languages. The difference between both scores is lower for Chinese and the specificity score is also good. With T-PAS+CT classifiers, there is an increase in specificity scores but also a drop in sensitivity scores for all languages. The scores for sensitivity and specificity are closer to each other. Overall, TPAS+CT classifiers perform better in terms of their overall performance for all languages except for Chinese.

Simple type nouns. The results of specificity scores on different types of negatives show that the low specificity score is much higher in the negatives with complex type nouns compared to the negatives with simple type nouns. The specificity scores increase with T-PAS+CT classifiers also for the second type of negatives however the difference between the two types of classifiers is much lower. For example, the increase for Italian is from 0.79 to 0.87 compared to 0.35 to 0.62.

6 Discussion

The findings of the previous studies suggest that LMs encode information about the selectional preferences of verbs (Metheniti et al., 2020; Chersoni et al., 2021; Li et al., 2021; Pedinotti et al., 2021) and semantic types of nouns (Zhao et al., 2020). Our study shows that it is possible to exploit this knowledge of LMs to train classifiers for the identification of the semantic types targeted by both verbs and adjectives.

From a cross-lingual point of view, our results show that it is possible to use the embeddings of the multilingual LMs to train classifiers in order to transfer knowledge from one language to another. Our results are in line with the previous studies in terms of the performance of individual models. XLM-RoBERTa yields better performance compared to other multilingual LMs (Conneau et al., 2020a; Hu et al., 2020; Liang et al., 2020; Lauscher et al., 2020) and its performance is comparable to monolingual models (Conneau et al., 2020a). Even though we have limited test data for adjectival predication classifiers, we expect the transfer to work similarly for both types of predications (verbal and adjectival) and the results for English show this is the case.

In the copredication detection task, our results show that classifiers that are trained only with data with simple type nouns are not able to disambiguate the meanings of complex type nouns. This is evident in the tendency of false positives (low specificity) with T-PAS classifiers. Even when both predications target the same semantic type in a sentence, i.e. in negatives, copredication is detected. This is because both semantic type classifiers tend to classify the predications as positive when a complex type noun is involved. However, this tendency is absent with simple type nouns, which is also evident in the specificity scores. We think that this tendency is due to the nature of complex type nouns and how they are represented by LMs, which is a topic we intend to investigate in the future. The false positive tendency is overcome by adding more data with complex type nouns and this improves the overall performance which shows that copredication detection is possible with the proposed model.

Cross-linguistically, the performance on copredication detection shows a similar pattern for all languages and for both monolingual and multilingual classifiers. In the future, we plan to use this method for building a cross-lingual collection of corpus-based copredication instances that includes also other complex types and copredication constructions.

7 Conclusion

In this study, we focused on training classifiers for the identification of the semantic argument types targeted by the predicates in a specific predication context using the extracted embeddings of LMs. We trained both monolingual and multilingual classifiers for different semantic types and for both verbal and adjectival predications. The training results for individual classifiers show that it is possible to train classifiers for this purpose using LMs and to train multilingual classifiers with zero-shot cross-lingual transfer using multilingual LMs . Furthermore, we proposed a method to detect copredications using these classifiers and evaluated the method's performance on cross-lingual copredication test data. Our results show that copredication detection is a more complicated task. However, the method achieves reasonable scores for all languages and good scores for English.

Acknowledgments

This study is a part of the project "Coercion and Copredication as Flexible Frame Composition" funded by DFG (Deutsche Forschungsgemeinschaft). We would like to thank Younes Samih for his invaluable insights and feedback.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Nicholas Asher. 2011. *Lexical Meaning in Context. A Web of Words*. Cambridge University Press, Cambridge.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

D. Alan Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. *Computational lexical semantics*, pages 33–49.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. Proceedings of Machine Learning Research.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS; a resource of typed predicate argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895, Reykjavik, Iceland. European Language Resources Association (ELRA).

Elisabetta Jezek and Laure Vieu. 2014. Distributional analysis of copredication: towards distinguishing systematic polysemy from coercion. In *1st Italian Conference on Computational Linguistics (CLiC-it 2014)*, volume 1, pages 219–223, Pisa, Italy.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations (ICLR'20)*, Addis Ababa, Ethiopia.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021. How is BERT surprised? layerwise detection of linguistic anomalies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4215–4228, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-*

guage Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Eleni Metheniti, Tim Van de Cruys, and Nabil Hathout. 2020. How relevant are selectional preferences for transformer-based language models? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1266–1278, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the cat drink the coffee? challenging transformers with generalized event knowledge. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 1–11, Online. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky and Elisabetta Jezek. 2008. *Semantic coercion in language: Beyond distributional analysis*, volume 20 of *Italian Journal of Linguistics / Rivista di Linguistica :*. De Gruyter Mouton, Berlin, Boston. 2010.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Laure Vieu, Elisabetta Jezek, and Tim Van de Cruys. 2015. Quantitative methods for identifying systematic polysemy classes. In *6th Conference on Quantitative Investigations in Theoretical Linguistics (QITL 2015)*, pages 1–5.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.

A Models, Parameters and Libraries Used for the Experiments

The embeddings are extracted using the Transformers library (Wolf et al., 2020). The classifiers are trained using the Scikit-learn library (Pedregosa et al., 2011). The Scikit-learn library is also used for the clustering of the negative dataset.

For the SVM algorithm, the radial basis function kernel is used with a C value of 100 and a gamma value of 0.001. K-Means is used as the clustering algorithm for the negative dataset, and the number of clusters (k) is determined as 10.

We use the BERT model *dbmdz/bert-base-italian-base-cased* as feature extractor for monolingual Italian classifiers and multilingual LMs mBERT *bert-base-multilingual-cased* and XLM-RoBERTa *xlm-roberta-base* for multilingual classifiers. All models are available at `https://huggingface.co/`.

B Comparison of Classification Algorithms

The performance of several classification algorithms (*Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine*) is compared for

Model		Languages													
Verbal Predication		it		de		en		tr		zh		Adjectival Predication			
												it		en	
		p	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*
	B	0.94	0.95	-	-	-	-	-	-	-	-	0.94	0.76	-	-
Arti.	mB	0.93	0.92	0.90	0.79	0.92	0.93	0.91	0.64	0.92	0.76	0.91	0.88	0.92	0.94
	XR	0.90	0.91	0.88	0.83	0.94	0.93	0.93	0.84	0.92	0.92	0.93	0.82	0.91	0.95
	B	0.94	0.97	-	-	-	-	-	-	-	-	0.88	0.88	-	-
Event	mB	0.94	0.91	0.90	0.86	0.97	0.91	0.91	0.81	0.92	0.88	0.86	0.86	0.85	0.68
	XR	0.94	0.88	0.82	0.97	0.92	0.95	0.88	0.94	0.84	0.99	0.78	0.84	0.90	0.79
	B		0.93	0.92	-	-	-	-	-	-	-	-	-	-	-
Hum.	mB	0.92	0.92	0.96	0.90	0.98	0.98	0.97	0.74	0.96	0.83	-	-	-	-
	XR	0.93	0.96	0.98	0.98	0.98	0.98	0.96	0.95	0.97	0.91	-	-	-	-
	B	1	0.97	-	-	-	-	-	-	-	-	0.86	0.96	-	-
Info.	mB	1	0.97	0.93	0.70	1	0.97	0.96	0.74	1	0.82	0.86	0.95	0.97	0.91
	XR	0.97	0.97	0.97	1	0.97	0.93	0.97	0.93	0.97	0.97	0.90	0.92	1	0.80
	B	0.95	0.95	-	-	-	-	-	-	-	-	-	-	-	-
Loc.	mB	0.91	0.92	1	0.85	0.92	0.97	0.97	0.56	0.93	0.92	-	-	-	-
	XR	0.97	0.94	1	0.94	0.98	0.95	0.96	0.83	0.96	0.90	-	-	-	-

Table 3: The test results of each classifier. Precision, Recall scores are given.

	Artifact	Event	Human	Info.	Loc.
Log. Reg.	0.94	**0.95**	**0.92**	**0.98**	**0.95**
Naive Bayes	0.87	0.92	**0.92**	0.97	0.92
Rand. Forest	0.88	0.91	0.90	0.97	0.94
SVM	**0.95**	**0.95**	**0.92**	**0.98**	**0.95**

Table 4: Performance of different classification algorithms on Italian verbal predication test data. Best performing classifiers for each semantic type are given in bold.

the monolingual verbal predication classification task. See Table 4 for the f1 scores of the classifiers trained with different algorithms. Overall, SVM is the best performing one.

C Data Correction Guideline

Please, follow these points for the manual correction of the translated test data:

- If the verb and the object are not identified correctly, they should be annotated manually.
- The sentences should be corrected if they sound unnatural or the predicate does not target the desired semantic type.
- For the correction, the sentences can be changed or the verb and the noun can be changed.
- The noun should be the object of the verb. If the verb takes a prepositional phrase instead, it should be changed with another verb.
- If any of the target words is a *multi-word expression*, the headword should be considered as the target word.
- If the sentence is passivized in translation, it should be turned into an active one.

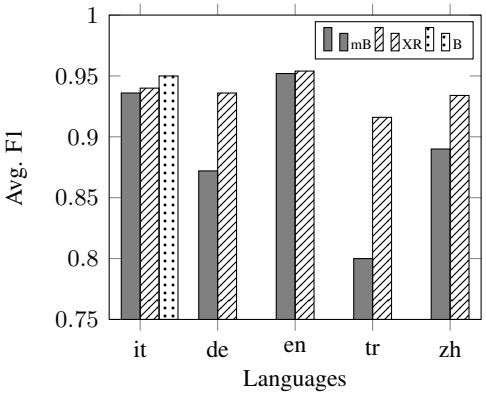

Figure 1: The average f1 score of different model-based classifiers on each language for verbal predication.

D Detailed Test Results

See Table 3 for the precision and recall scores of each classifier.

E Language-Specific Evaluation

The performance of the multilingual classifiers for verbal predication changes depending on the target language (See Figure 1). Similar to the studies that investigate the effects of structural differences of languages on cross-lingual transfer (Pires et al., 2019; K et al., 2020; Conneau et al., 2020b; Lauscher et al., 2020; Deshpande et al., 2022), our results show that the performance of the XR-classifiers on the typologically more distant languages Turkish and Chinese is worse. Similarly, the mB-classifiers perform worse on Turkish. We don't think the quality of the translations is the reason since native speakers manually checked the translations of these languages. However, even the worst performance is still good with over 0.8 f1 score.

The classifiers perform best on English test data, which is not the source language. One possible reason is that the Italian test data was not manually corrected, in contrast to the target languages. For this reason, the test data for the source language may contain more noise due to, e.g. parsing errors.

Even though XLM-RoBERTa improves the results for all languages, we see that the improvement changes depending on the language. One possible explanation is the larger size of training data for the XLM-RoBERTa model for these languages compared to mBERT.

F Food•Event Nouns

pranzo ('lunch'), *cena* ('dinner'), *colazione* ('breakfast'), *merenda* ('snack'), *aperitivo* ('aperitif'), *buffet* ('buffet'), *picnic* ('picnic'), *pasto* ('meal'), *spuntino* ('snack')

G Data Information for Copredication Test Data

The test data contains 30 positive and 24 negative examples of copredication with different semantic types targeting a Food•Event noun. There are both *Food verb-Event adj* and *Event verb-Food adj* combinations in the positives. Similarly, there are both *Food verb-Food adj* and *Event verb-Event adj* combinations in the negatives. The distributions of the types can be seen in Table 4. The cross-lingual copredication test data contains the same number of sentences and distribution for all languages, except for Chinese, which lacks one sentence for *Food verb-Event adj* and one sentence for *Event verb-Event adj*.

Positives	**Negatives**
Total: 30	Total: 24
food-event: 15	*food-food:* 15
event-food: 15	*event-event:* 9

Table 5: Data size and type distribution of copredication test data. The first semantic type refers to the verbal predication and the second one to the adjectival predication, e.g. *food-event*: Food verb-Event adj.

In addition to these data, another type of negative instances is created in order to test the effects of the complex type nouns in copredication detection. This data contains negative instances of copredication with simple type nouns, in which both a verb and an adjective targeting the same semantic type (also the same as the noun's semantic type) predicate over the noun. This type of negative instances are produced by substituting the Food•Event nouns in the negatives with a Food or Event simple type noun. However, in some cases, the sentences are also changed in order to make them more natural. In 24 sentences, 9 sentences are exactly the same except for the noun. However, 10 sentences are changed to some extent, leaving the predicates the same, and 5 sentences are changed completely.