Neural Metaphor Detection for Slovene

Matej Klemen **Marko Robnik-Šikonja**
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, Ljubljana, Slovenia
{matej.klemen, marko.robnik}@fri.uni-lj.si

Abstract

Metaphors are linguistic expressions using comparison with another concept to potentially improve the language expressivity. Due to relevant downstream applications, metaphor detection is an active topic of research. Most of the research is focused on English, while other languages are less covered. In our work, we focus on Slovene, presenting the first word-level metaphor detection experiments. We apply multiple transformer-based large language models on four versions of two publicly available Slovene corpora: KOMET and G-KOMET. We perform monolingual, multilingual, and cross-lingual experiments, using the VU Amsterdam metaphor corpus as an additional source of metaphor knowledge. We evaluate the models quantitatively using word-level F_1 score and find that (1) the most consistently well-performed model is the trilingual CroSloEngual BERT model, (2) the addition of English data in multilingual experiments does not improve the performance significantly, and (3) the cross-lingual models achieve significantly worse results than their monolingual and multilingual counterparts.

1 Introduction

A metaphor is an expression that uses a comparison with another concept for rhetorical effect. For example, instead of saying *"his words offended me"* we might say *"his words cut deeper than a knife"*, comparing the effect of offensive words to the physical pain of a knife cut. Metaphors are ubiquitous in language and add color to conversations. The ability to detect and form metaphors can be applied to creative writing, such as news headline generation or rephrasal (Stowe et al., 2021), and enables the analysis of public discourse and evolution of language (Prabhakaran et al., 2021; Zwitter Vitez et al., 2022; Kutuzov et al., 2018).

Although humans are able to detect and understand metaphors with relative ease, metaphor detection has proven to be a challenging problem for computational models (Strapparava, 2018). Most existing work on metaphor detection has dealt with broadly-spoken languages such as English. The earlier approaches relied on handcrafted features such as abstractness and concreteness features in combination with machine learning models (Turney et al., 2011). Recent work is shifting towards the use of large language models that model the word context (Choi et al., 2021).

An example of a language with less researched metaphors is Slovene, where little work has been done on metaphor detection, although annotated datasets exist (Antloga, 2020; Antloga and Donaj, 2022). The only computational approach to metaphor detection for Slovene was done by Zwitter Vitez et al. (2022). Authors detect metaphors at the sentence-level by first training a model for idiom detection and then tuning it for metaphor detection on the KOMET corpus (Antloga, 2020). The idea behind their work is that although metaphors and idioms are different concepts, both are forms of figurative language that heavily rely on a word's context.

In our work, we approach the metaphor detection at a fine-grained (i.e. word) level and present the results of initial experiments applying state-of-the-art language models to Slovene metaphors, leveraging multiple CLARIN resources in the process. Although metaphors are multi-word units, we model their

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Matej Klemen and Marko Robnik-Šikonja 2023. Neural Metaphor Detection for Slovene. *Selected papers from the CLARIN Annual Conference 2022. Ed. by Tomaž Erjavec and Maria Eskevich.* Linköping Electronic Conference Proceedings 198, pp. 77–89. DOI: https://doi.org/10.3384/ecp198008

detection at a word level as the metaphors in the datasets are annotated at this level, following the MIPVU annotation scheme (Steen et al., 2010). In contrast to detection at the sentence level, this approach enables a clearer insight into what part of the text is metaphorical. On the other hand, such an approach runs the risk of the model only learning to detect the "easier" subset of words in a multi-word expression as the model only considers dependence between words implicitly, an aspect we discuss in our evaluation. We perform experiments on the KOMET metaphor dataset and the previously unexplored spoken language dataset G-KOMET. We test metaphor detection in a monolingual, multilingual, and cross-lingual setting, using the VU Amsterdam metaphor corpus (VUAMC) (Krennmayr and Steen, 2017) as an additional source of knowledge:

- In the monolingual experiments, we test the performance of large language models on word-level metaphor detection setting for Slovene.
- In the multilingual experiments, we test if the inclusion of English language can improve the performance on Slovene, for example by acting as a regularization mechanism.
- In the cross-lingual experiments, we test if the models are able to learn any transferable language-agnostic ideas behind metaphors on the English data and test it on the Slovene data.

We make the code to rerun our experiments publicly available[1].

The remainder of our paper is structured as follows. In Section 2, we describe the relevant existing work on metaphor detection. In Section 3, we describe the data used in our experiments. Section 4 presents our approach for metaphor detection, while the results are contained in Section 5. We present conclusions and discuss possible future directions in Section 6.

2 Related Work

The work on metaphor detection has progressed from initial manually engineered feature-based approaches in combination with machine learning algorithms towards increasingly more automated (end-to-end) feature learning in combination with deep neural networks.

In the initial work on metaphor detection, authors have tested various features such as concreteness of words, named entity features, and part of speech tag features (Tsvetkov et al., 2013; Beigman Klebanov et al., 2014) in combination with machine learning algorithms such as logistic regression (Cox, 1958). With the introduction of word embedding algorithms, authors have discovered that the embeddings are able to replace or simplify the process of feature construction. For example, Do Dinh and Gurevych (2016) use word2vec word embeddings (Mikolov et al., 2013) in their metaphor detection system and demonstrate their feasibility on English data. In addition to word embeddings, they also use part of speech tag embeddings, replacing the manual linguistic features with automatically learned ones. Although powerful, in contrast to the inherently contextually dependent nature of metaphors, earlier proposed word embeddings such as word2vec cannot produce a contextual word representation. To amend this, researchers have incorporated the embeddings in combination with models that are capable of capturing an increasing amount of context, such as long short term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and convolutional neural networks (CNNs) (LeCun et al., 2010). For example, Pramanick et al. (2018) propose an approach using a bidirectional LSTM and conditional random fields (Lafferty et al., 2001) to improve the English metaphor detection performance. Lately, authors have started replacing LSTMs and CNNs with transformer models (Vaswani et al., 2017) due to their wide success on other tasks in natural language processing. Despite the powerful representation capability of transformer-based models, latest metaphor detection models continue to benefit from the inclusion of external information such as concreteness (Alnafesah et al., 2020) or the customization of architecture to account for the specifics of the metaphor detection task. For example, Choi et al. (2021) include custom components in their approach that model the process used in the annotation of metaphor datasets. They show that their system performs better than an out-of-the-box transformer model.

[1] https://github.com/matejklemen/metaphor-recognition

Most of the work is focused on processing English metaphors, and the topic has been the focus of multiple shared tasks (Leong et al., 2018; Leong et al., 2020; Saakyan et al., 2022). Some other broadly spoken languages such as Chinese and Russian have also received significant attention and customized architectures (Song et al., 2021; Lu and Wang, 2017; Badryzlova et al., 2022), while work on other languages is relatively scarce in comparison. Examples of metaphor detection systems can be found for diverse languages, such as Spanish (Sanchez-Bayona and Agerri, 2022), Greek (Florou et al., 2018), and Uyghur (Qimeng et al., 2021), although they are typically focused on a narrower domain (e.g., literary) or present initial studies on the feasibility of metaphor detection.

With our work, we expand the research on non-English metaphor detection. The only previous approach for Slovene metaphor detection is by Zwitter Vitez et al. (2022), who detect metaphors at the sentence level using the KOMET dataset. In contrast, we approach metaphor detection at a more fine-grained (word) level and test the state-of-the-art transformer models in a monolingual, multilingual and cross-lingual setting.

3 Metaphor Datasets

In our experiments, we use three datasets for evaluating the metaphor detection performance: KOMET (Antloga, 2020), G-KOMET (Antloga and Donaj, 2022), and VUAMC (Krennmayr and Steen, 2017). In addition to these original datasets, we create modified versions of KOMET and G-KOMET containing "semantically interesting metaphors", i.e. metaphors containing at least one noun or verb (marked as *NV* in Table 1).

The datasets are annotated using the MIPVU annotation scheme (Steen et al., 2010). Using the scheme, a word is annotated as "metaphor-related" if its contextual meaning differs from its basic meaning. In Table 1, we provide basic statistics of the three datasets.

	KOMET		G-KOMET		VUAMC
	full	NV	full	NV	
# documents	62		287		117
# sentences	13 963		5695		16 202
# words	259 881		52 955		238 509
# met.	(5.2%) 13 574	(1.2%) 3100	(1.1%) 560	(0.7%) 357	(9.5%) 22 620
# MRWi	13 191	2893	527	324	22 254
# MRWd	364	205	33	33	341
# MRWimp	19	2	0	0	25

Table 1: Statistics of the used datasets: number of documents, sentences, words, and metaphors, as well as types of metaphors (indirect - MRWi, direct - MRWd, implied - MRWimp). The columns marked NV show statistics for the processed version of datasets where only metaphors containing a noun or verb are kept as metaphors.

KOMET is a Slovene metaphor corpus containing 13 963 sentences. It contains journalistic, fiction and web texts extracted from the Slovene youth literature corpus MAKS (Verdonik et al., 2020). The corpus was annotated by one annotator. Approximately 5.2% of the words are marked as metaphors in the original version and approximately 1.2% in the NV version.

G-KOMET is a Slovene metaphor corpus containing 5695 sentences. In contrast to KOMET, it contains transcripts of spoken language extracted from the GOS corpus (Verdonik et al., 2013). The corpus was annotated by one annotator. Approximately 1.1% of the words are marked as metaphors in the original version and approximately 0.7% in the NV version.

VUAMC is an English metaphor corpus containing 16 202 sentences. It contains academic, news, conversational, and fiction texts from the BNC-Baby corpus (BNC Consortium, 2007). The corpus was annotated by four annotators.

In our experiments, we are interested in detecting the following three metaphor types annotated in all used corpora.

- Indirect metaphor (MRWi): a metaphor where a comparison is indirectly stated. For example, in "Tim se je zlajal nanj" (*"Tim barked at him"*), Tim's yelling is compared to the barking of a dog and Tim is indirectly compared to a dog.
- Direct metaphor (MRWd): a metaphor where a comparison is directly stated. For example, in "Tim je osel" (*"Tim is a donkey"*), Tim is being compared to a donkey to outline his stubbornness.
- Implied metaphor (MRWimp): a lexical unit that is not necessarily metaphorical by itself but refers to a previous metaphorically used word. For example, in *"Naturally, to embark on such a step is not necessarily to succeed immediately in realising it"*, "step" is related to a metaphor and "it" refers to "step", so it is considered an implicit metaphor (Steen et al., 2010).

We use KOMET and G-KOMET for training and evaluation in mono-, multi-, and cross-lingual settings, while we use VUAMC for training in the multi- and cross-lingual setting. We do not evaluate the performance on English as there exist many works on the topic (Leong et al., 2020).

4 Metaphor Detection

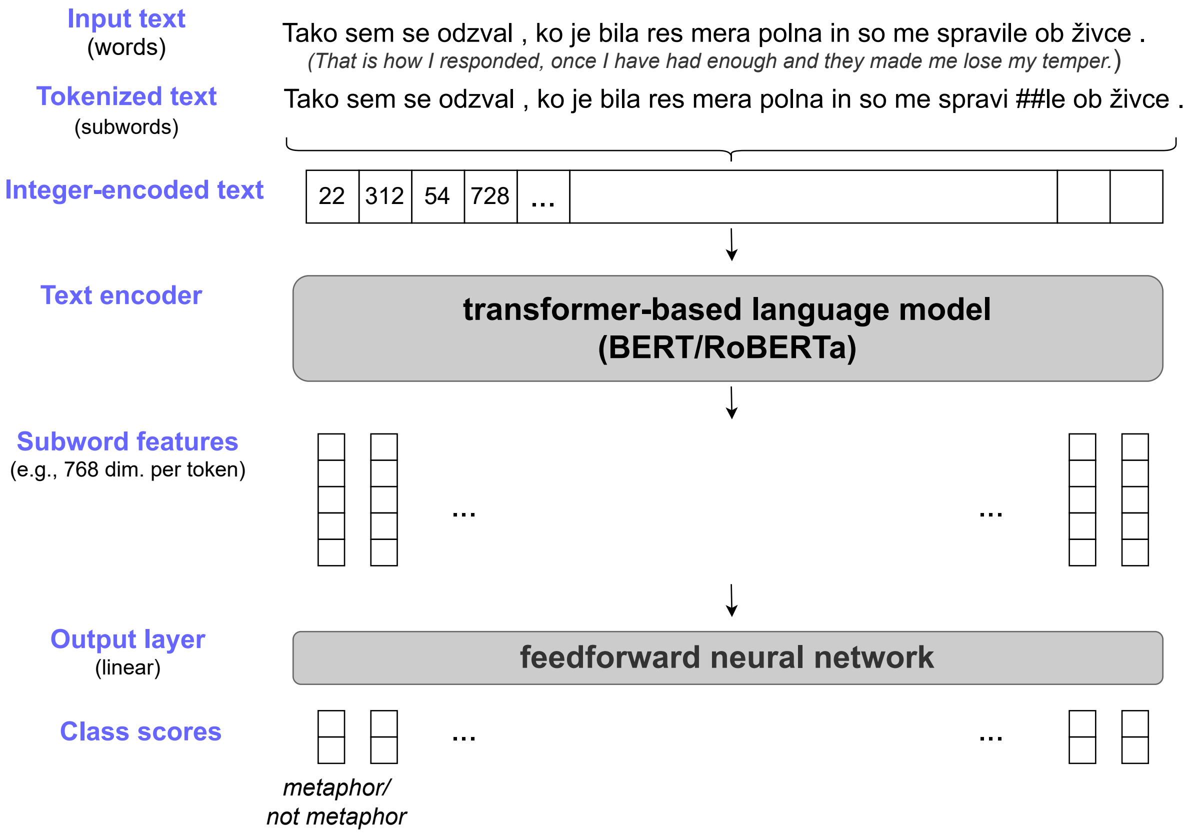

Input text
(words)
Tako sem se odzval , ko je bila res mera polna in so me spravile ob živce .
(That is how I responded, once I have had enough and they made me lose my temper.)
Tokenized text
(subwords)
Tako sem se odzval , ko je bila res mera polna in so me spravi ##le ob živce .
Integer-encoded text
22
312
54
728
...
Text encoder
transformer-based language model
(BERT/RoBERTa)
Subword features
(e.g., 768 dim. per token)
...
...
Output layer
(linear)
feedforward neural network
Class scores
...
...
metaphor/
not metaphor

Figure 1: Schematic representation of the metaphor detection neural networks used in our experiments. Each integer-encoded subword of the input text is passed through the transformer-based language model to obtain a representation in the form of an embedding vector. Then, the embeddings are passed through the final classification layer to obtain a score of the subwords being a metaphor or non-metaphor.

We use the described datasets in a binary token classification setting, described next. The token classification procedure follows the best practice in existing literature (Devlin et al., 2019), to which we make

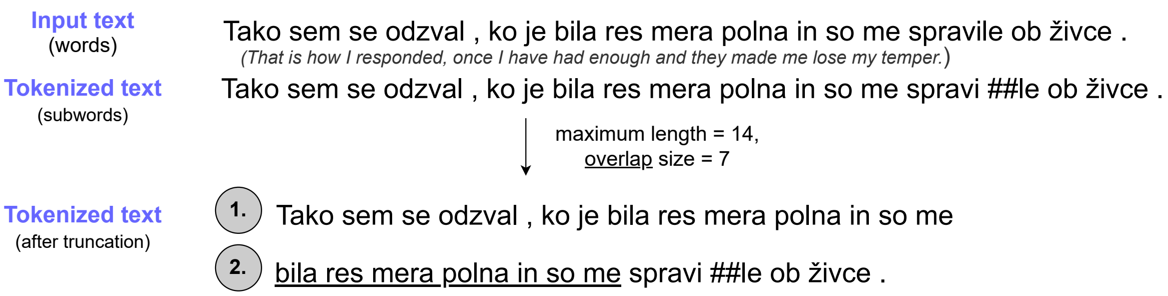

Input text
(words)
Tako sem se odzval , ko je bila res mera polna in so me spravile ob živce .
(That is how I responded, once I have had enough and they made me lose my temper.)
Tokenized text
(subwords)
Tako sem se odzval , ko je bila res mera polna in so me spravi ##le ob živce .
maximum length = 14,
overlap size = 7
Tokenized text
(after truncation)
1. Tako sem se odzval , ko je bila res mera polna in so me
2. bila res mera polna in so me spravi ##le ob živce .

Figure 2: A toy example of the truncation logic in our modeling process. The input text gets tokenized into subwords and then broken up into two inputs as it is initially longer than the maximum model length (14 in this example). The second example has a partial overlap with the first example (marked with underline) of half the size of the maximum input length.

minor modifications to account for the nature of the metaphor detection task. In Figure 1, we show an example of detecting metaphors in a sentence.

The input text is first tokenized and encoded using the model-specific tokenizer, which converts the input into integer-encoded subwords. The subwords may be identical to the corresponding words or they can be smaller constituents of the words, depending on encoding of the input word in the model-specific vocabulary. The encoded text is then passed through the transformer-based model, which produces an internal representation in the form of a fixed dimensional (e.g., 768) vector for each subword. Intuitively, this representation captures the context in which a word appears. Throughout the training procedure, the representation is optimized to capture features that are useful for determining if a subword is a metaphor or not. In the final step, the token representations are passed through the feedforward neural network, which produces a score for each possible class, in our case metaphor and non-metaphor. By applying the softmax function on the scores, they can be interpreted as the probabilities of a subword being a metaphor or not. If the probability is higher than T, the subwords is classified as a metaphor. We determine the probability threshold T on the validation set based on the F_1 score.

Transformer models have a limited maximum input length, i.e. the maximum number of subwords they accept as an input. The limit reflects primarily the computational constraints, more specifically the required amount of GPU memory and the maximal batch sizes used during model training. Inputs longer than the maximum length need to be handled separately. We break long input texts into multiple partially-overlapping input texts of the maximum allowed length. The size of the overlap is a hyperparameter and presents a trade-off: setting it too low drops potentially important context while setting it too high increases the amount of computation required for classification. We set it to half of the determined maximum length of a model as a sensible middle ground. In addition, we set the maximum input length high enough so that the examples do not need to be broken up frequently. The prediction for overlapping words is only taken into account once, i.e. when the overlapping segment is first observed. We show a toy example in Figure 2: the input “Tako sem se odzval, ko je bila res mera polna in so me spravile ob živce.” gets broken up into two inputs, with the second input partially overlapping the first one.

Another decision is how to classify words that the tokenizer splits into multiple subwords. For example, in Figure 1 the word “spravile“ gets split into “spravi‘ and “le”. In the prediction phase, each subword is classified separately, so the predictions need to be aggregated to determine if a word is a metaphor. By default, we aggregate the predictions by setting the class of a word to the class of its first subword but further aggregation experiments are done in Section 5.4.

5 Evaluation

In this section, we analyze the performance of metaphor detection models on the described datasets. We describe the three evaluation settings in Section 5.1. Then we analyze the results quantitatively and qualitatively in Sections 5.2 and 5.3. Last, we present an additional experiment analyzing the effect of subword score aggregation on the metaphor detection performance in Section 5.4.

5.1 Experimental Settings

In our experiments, we use four transformer models: monolingual SloBERTA (Ulčar and Robnik-Šikonja, 2021), trilingual CroSloEngual BERT (CSE BERT) (Ulčar and Robnik-Šikonja, 2020), and multilingual bert-base-multilingual-cased (mBERT$_{\text{BASE}}$) (Devlin et al., 2019) and XLM-RoBERTa$_{\text{BASE}}$ (XLM-R$_{\text{BASE}}$) (Conneau et al., 2020). SloBERTa was trained on Slovene, CSE BERT on Slovene, Croatian, and English, and mBERT$_{\text{BASE}}$ and XLM-R$_{\text{BASE}}$ were trained on 104 and 100 languages, respectively. In terms of the number of parameters, SloBERTa and CSE BERT are comparable at 110 million parameters, mBERT$_{\text{BASE}}$ is larger with 172 million parameters, and XLM-R$_{\text{BASE}}$ is the largest with 270 million parameters. As the representation of the tokens (token features), we take the hidden state of the last transformer layer. In our preliminary experiments, we have also experimented with using a learned combination of all the hidden states, but found no significant difference on the metaphor detection performance. We first test all models in a monolingual setting, i.e. training them on the Slovene training set and evaluating them on the Slovene test set. We test the multilingual models also in a multi-lingual and cross-lingual settings, i.e. training them on the combined Slovene and English training set or just the English training set, and evaluating them on the Slovene test set. As the models use different tokenization which could lead to incomparable subword-level metrics, we use the word-level F_1 score of the positive class (i.e. *metaphor*) for evaluation.

We train the models for 10 (on KOMET and VUAMC) or 20 epochs[2] (on G-KOMET) and select the best model based on the validation set F_1 score. We use the learning rate $2e-5$ and set the batch size to the maximum number that is possible on a 11GB GPU. We set the maximum input length of the models to the 99th percentile of the input lengths in the training set. This is different for every tokenizer-dataset pair but is typically between 80 and 100 subwords.

As we noticed a high variation in the results in our preliminary experiments, we perform the evaluation using 5-fold cross validation. The examples are split into folds on the document-level to reduce the possibility of an information leak. The folds are determined in a way that the proportion of metaphors is approximately the same in each fold. This simplification is due to the focus of the paper being the feasibility of metaphor detection on Slovene and not the ability of models to handle distribution shift. In each of the five evaluation runs, we set aside 10% of the training documents as the validation set. We split VUAMC in the same distribution-preserving fashion in the ratio 80%:20% training:test documents, although we do not evaluate models on English.

When comparing scores of different variants of a model, we use the Wilcoxon signed-rank test (Wilcoxon et al., 1970). We test the null hypothesis that the mean scores are the same using the significance level $\alpha = 0.01$.

5.2 Metaphor Word-Level Detection Results

In Table 2, we show the mean F_1 scores and standard deviations of the models. In one setting (SloBERTa on the G-KOMET$_{\text{NV}}$ dataset), the model did not converge, so we mark its performance as *N/A* and exclude it from further analysis. In addition to the models mentioned in the previous section, we include a naïve baseline which predicts the metaphor class with probability 0.5 to check if the models have learned anything beyond random guessing. We see that the models in all cases surpass this baseline.

In all three settings (monolingual, multilingual, and cross-lingual), the F_1 scores on the original (i.e. full sized) datasets are higher than on the NV counterparts. For example, the highest mean F_1 score on KOMET$_{\text{full}}$ is 0.607, while it is only 0.401 on KOMET$_{\text{NV}}$. This implies that a noticeable portion of the score on the full version comes from correctly detecting metaphors that are not nouns or verbs, e.g., adpositions such as *"na"* (*"on"* in English). These make up a significant amount of all the annotated metaphors, but are less interesting for practical applications such as rephrasing sentences with metaphors. The differences between full and NV variants on KOMET are higher than on G-KOMET as

[2] In preliminary experiments, we have found that the validation F_1 score kept increasing after 10 epochs, so we increased it.

Model	KOMET$_{\text{full}}$	KOMET$_{\text{NV}}$	G-KOMET$_{\text{full}}$	G-KOMET$_{\text{NV}}$
0.5/0.5 baseline	0.095 (0.008)	0.024 (0.003)	0.022 (0.003)	0.013 (0.001)
(monolingual)				
CSE BERT	0.606 (0.069)	0.361 (0.031)	0.261 (0.021)	0.243 (0.040)
SloBERTa	0.596 (0.071)	**0.401** (0.040)	0.243 (0.021)	N/A
XLM-R$_{\text{BASE}}$	0.591 (0.073)	0.348 (0.027)	0.245 (0.033)	0.205 (0.049)
mBERT$_{\text{BASE}}$	0.575 (0.058)	0.304 (0.028)	0.216 (0.033)	0.173 (0.028)
(multilingual $_{EN\,+\,SL}$)				
CSE BERT	**0.607** (0.068)	0.389 (0.031)	**0.313** (0.039)	0.276 (0.038)
XLM-R$_{\text{BASE}}$	0.599 (0.077)	0.354 (0.028)	0.282 (0.024)	**0.283** (0.045)
mBERT$_{\text{BASE}}$	0.573 (0.065)	0.320 (0.020)	0.223 (0.016)	0.237 (0.039)
(cross-lingual $_{EN \Rightarrow SL}$)				
CSE BERT	0.351 (0.035)	0.178 (0.036)	0.124 (0.041)	0.073 (0.021)
XLM-R$_{\text{BASE}}$	0.408 (0.046)	0.134 (0.021)	0.110 (0.010)	0.070 (0.010)
mBERT$_{\text{BASE}}$	0.374 (0.041)	0.112 (0.012)	0.089 (0.009)	0.053 (0.008)

Table 2: Mean word-level F_1 scores of metaphor detection models measured using 5-fold cross validation. The corresponding standard deviations are shown in parentheses. *N/A* indicates that the model did not converge, i.e. the validation metric did not improve in the training process. We mark the highest mean F_1 score in bold.

the proportion of metaphors is lower in the latter, so the proportion of metaphors in the NV version is not reduced as significantly as in KOMET. We further analyze the performance of models concerning part of speech tags of metaphors in Section 5.3.

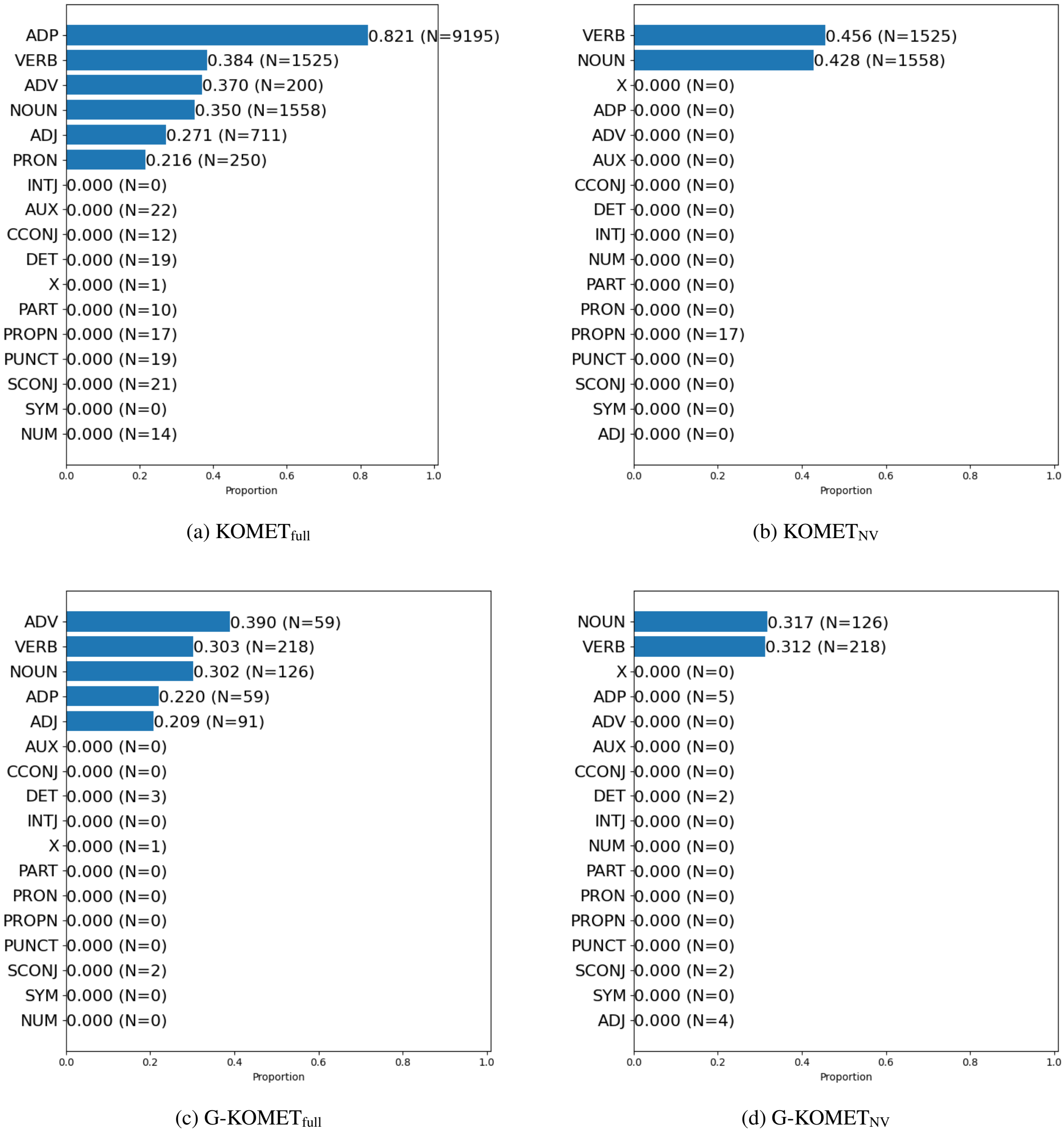
In the monolingual experiments, we see that the models achieve comparable results with minor differences. In general, the best performing model across the four datasets is CSE BERT, achieving the best mean F_1 score on three (KOMET$_{\text{full}}$, G-KOMET$_{\text{full}}$, and G-KOMET$_{\text{NV}}$) and the second best mean F_1 score on one (KOMET$_{\text{NV}}$) dataset. However, the differences in performance between the models are not statistically significant.

In the multilingual experiments, the inclusion of English training data does not have a significant effect on the model performance. Comparing the multilingual models to their monolingual counterparts shows that there are potential differences in the performance on the G-KOMET dataset, although their statisical significance cannot be confirmed due to the high deviation in model performance. The improvement is possibly a consequence of the heavily imbalanced dataset and a smaller amount of training examples: G-KOMET is five times smaller than KOMET and contains only approximately 1% of metaphors.

In the cross-lingual experiments, we see a significant drop in performance across all the datasets. The results indicate that there is a limited amount of metaphor knowledge that can be transferred from the used English source to the Slovene target datasets. This is especially clear on G-KOMET, where an additional obstacle is the specialized domain, i.e. spoken language. Due to this, the worst performing cross-lingual models closely approach the random baseline performance, although the improvement in performance over the random baseline is still statistically significant.

5.3 POS-Tag Analysis of the Detected Metaphors

To observe what the models successfully detect, we analyze the model predictions based on universal part of speech (UPOS) tags of metaphors. We obtain them using the Trankit (Nguyen et al., 2021) library, using the Slovenian-SSJ large model. In Figure 3, we show the proportion of correctly detected metaphors for each UPOS tag for the CSE BERT model on the four datasets in the monolingual setting. We select this model as it consistently performs well, but we have noticed that the proportions are similar for other models in the monolingual and multilingual setting. On KOMET$_{\text{full}}$, we observe that the model most accurately detects the adpositions (82.1% of them), which present the majority of the annotated metaphors. On the other hand, only 35% of nouns and 38.4% of verbs are correctly detected. The proportion of correctly detected noun and verb metaphors increases in the more narrowly focused KOMET$_{\text{NV}}$ dataset. A similar conclusion can be drawn for G-KOMET, shown on Figure 3c and 3d, although adpositional metaphors do not present the majority in G-KOMET$_{\text{full}}$; instead, nouns and verbs present the majority. Therefore, the increase in the detection performance for nouns and verbs is significantly less noticeable in the G-KOMET$_{\text{NV}}$ dataset.



(a) KOMET$_{\text{full}}$

(b) KOMET$_{\text{NV}}$

(c) G-KOMET$_{\text{full}}$

(d) G-KOMET$_{\text{NV}}$

Figure 3: Proportion of metaphors correctly detected by CSE BERT in a monolingual setting, grouped by their UPOS tag.

We perform the same analysis for the best performing model in the cross-lingual setting on the KOMET dataset. The results are shown on Figure 4. Interestingly, although it performs significantly worse in terms of F_1 score, this is primarily a consequence of poorly detecting the adpositional metaphors. On the other hand, the proportion of correctly detected nouns and verbs is equal or better than on KOMET$_{\text{NV}}$.

While it does improve the ability of detecting semantically interesting metaphors, ignoring non-noun and non-verbal metaphors is not the ideal solution as these may still be used in some linguistic analyses. More importantly, they can act as constituents of a metaphor composed of multiple words. For example, in the metaphor "spravile ob živce" (*"to make someone lose their temper"*), the word "ob" is essential for the phrase to be considered a metaphor. We note that the annotation of multi-word metaphors in existing datasets is imperfect as they are commonly annotated as multiple single-word metaphors: e.g., "spravile", "ob", and "živce" are labelled as three metaphors.

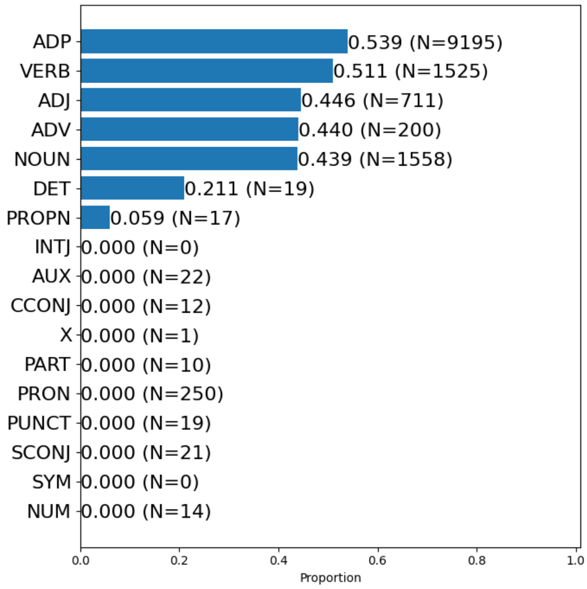


Figure 4: Proportion of metaphors correctly detected by XLM-R$_{\text{BASE}}$ on KOMET$_{\text{full}}$ in a cross-lingual setting, grouped by their UPOS tag.

5.4 Analysis of Subword Score Aggregation

In this additional experiment we test the effect of the strategy used to aggregate subword predictions. As the model tokenizers may break up the input word into multiple subwords, the predictions need to be aggregated to the word level. In our implementation, we use by default the prediction for the first subword as the prediction of the word - we refer to this strategy as *first*. However, as the subword predictions are made independently, meaning the prediction $\hat{y}_i$ is independent of $\hat{y}_{i-1}$, it is possible that subwords belonging to the same word obtain inconsistent predictions. Therefore, we test two additional aggregation strategies to check if using a different strategy leads to a significant difference in performance:

- *majority*: the prediction of a word is determined as the majority prediction of its subwords;
- *any*: the prediction of a word is metaphor if any of the subwords is a metaphor, and non-metaphor otherwise.

To avoid overfitting the test set, we compare the validation set F_1 scores. We test this only for the XLM-R$_{\text{BASE}}$ model as its tokenizer is designed to handle 100 languages, so the words are far more likely to get divided into subwords than in the monolingual SloBERTa or trilingual CSE BERT model. From the results shown in Table 3, we can observe that the difference in performance across the three strategies is minimal and statistically insignificant. The results indicate that the model is likely to track the dependence between predictions in its hidden layers. As the results on XLM-R$_{\text{BASE}}$ show no difference, we skip comparisons for other models as the corresponding tokenizers are equally or less likely to split words into subwords, so the results are unlikely to be different.

Model	Strategy	KOMET$_{\text{full}}$	KOMET$_{\text{NV}}$	G-KOMET$_{\text{full}}$	G-KOMET$_{\text{NV}}$
XLM-R$_{\text{BASE}}$	first	0.662 (0.035)	0.419 (0.024)	0.270 (0.021)	0.226 (0.045)
	majority	0.663 (0.036)	0.418 (0.021)	0.266 (0.009)	0.219 (0.049)
	any	0.664 (0.037)	0.420 (0.032)	0.272 (0.026)	0.227 (0.056)

Table 3: Comparison of subword prediction aggregation strategies. The prediction for a token is either the prediction for its first subword (*first*), the majority prediction of all its subwords (*majority*), or determined as the metaphor if at least one of the subwords is a metaphor, and non-metaphor otherwise (*any*). The scores are mean word-level F_1 scores measured using 5-fold cross validation. The corresponding standard deviations are shown in parentheses.

6 Conclusion

We presented the results of the first word-level metaphor detection attempt on Slovene data, analyzing the performance of the models in a monolingual, multilingual, and cross-lingual setting. Our approach considers metaphor detection as a standard token classification task with minor modifications, such as prediction aggregation and threshold optimization, to account for the specifics of the task. The results show that the models have plenty of room for improvement and perform best at detecting semantically less interesting metaphors, such as adpositions. The inclusion of English data in multilingual experiments has a minor and insignificant effect. The performance drops significantly in the cross-lingual experiments, indicating that there is a limited amount of knowledge that is transferrable from English to the Slovene datasets.

An issue in current metaphor datasets is the disjoint nature of metaphor annotations, i.e. multi-word metaphors are commonly annotated as multiple single-word metaphors which potentially limits options to improve their modeling. In future work, we plan to tackle this issue by proposing an automatic grouping mechanism which will allow modeling metaphor detection as a span extraction task instead of a token classification task.

An additional direction for future work is the introduction of new datasets as resources large enough to train neural networks are only available for a limited set of languages. The introduction of new datasets will enable detection in new domains and languages as well as potentially enable cross-lingual transfer.

Acknowledgements

The research was supported by the Slovene Research Agency through research core funding no. P6-0411, project J6-2581 (CANDAS - Computer-assisted multilingual news discourse analysis with contextual embeddings), and the young researcher grant.

References

[Alnafesah et al.2020] Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. Augmenting neural metaphor detection with concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 204–210, July.

[Antloga and Donaj2022] Špela Antloga and Gregor Donaj. 2022. Corpus of metaphorical expressions in spoken Slovene language G-KOMET 1.0. Slovenian language resource repository CLARIN.SI.

[Antloga2020] Špela Antloga. 2020. Korpus metafor KOMET 1.0. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 167–170.

[Badryzlova et al.2022] Yulia Badryzlova, Olga Lyashevskaya, and Anastasia Nikiforova. 2022. Automated metaphor identification in Russian and its implications for metaphor studies. In *Distributed Computing and Artificial Intelligence, Volume 2: Special Sessions 18th International Conference*, pages 86–96.

[Beigman Klebanov et al.2014] Beata Beigman Klebanov, Ben Leong, Michael Heilman, and Michael Flor. 2014. Different texts, same metaphors: Unigrams and beyond. In *Proceedings of the Second Workshop on Metaphor in NLP*, pages 11–17, June.

[BNC Consortium2007] BNC Consortium. 2007. British national corpus, XML edition. Oxford Text Archive.

[Choi et al.2021] Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773.

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

[Cox1958] D. R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242.

[Devlin et al.2019] 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.

[Do Dinh and Gurevych2016] Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, June.

[Florou et al.2018] Eirini Florou, Konstantinos Perifanos, and Dionysis Goutsos. 2018. Neural embeddings for metaphor detection in a corpus of greek texts. In *2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4.

[Hochreiter and Schmidhuber1997] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780, nov.

[Krennmayr and Steen2017] Tina Krennmayr and Gerard J. Steen, 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer.

[Kutuzov et al.2018] Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

[Lafferty et al.2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[LeCun et al.2010] Yann LeCun, Koray Kavukcuoglu, and Clement Farabet. 2010. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256.

[Leong et al.2018] Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, June.

[Leong et al.2020] Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29.

[Lu and Wang2017] Xiaofei Lu and Ben Pin-Yun Wang. 2017. Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation*, 51(3):663–694.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

[Nguyen et al.2021] Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

[Prabhakaran et al.2021] Vinodkumar Prabhakaran, Marek Rei, and Ekaterina Shutova. 2021. How metaphors impact political discourse: A large-scale topic-agnostic study using neural metaphor detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):503–512.

[Pramanick et al.2018] Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An LSTM-CRF based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing*, pages 67–75, June.

[Qimeng et al.2021] Yang Qimeng, Yu Long, Tian Shengwei, and Song Jinmiao. 2021. Uyghur metaphor detection via considering emotional consistency. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 895–905.

[Saakyan et al.2022] Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. A report on the FigLang 2022 shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, December.

[Sanchez-Bayona and Agerri2022] Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, December.

[Song et al.2021] Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2021. A knowledge graph embedding approach for metaphor processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:406–420.

[Steen et al.2010] Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

[Stowe et al.2021] Kevin Stowe, Nils Beck, and Iryna Gurevych. 2021. Exploring metaphoric paraphrase generation. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 323–336, November.

[Strapparava2018] Carlo Strapparava. 2018. Metaphor: A Computational Perspective. *Computational Linguistics*, 44(1):191–192, 03.

[Tsvetkov et al.2013] Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, June.

[Turney et al.2011] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690.

[Ulčar and Robnik-Šikonja2021] Matej Ulčar and Marko Robnik-Šikonja. 2021. Slovenian RoBERTa contextual embeddings model: SloBERTa 2.0. Slovenian language resource repository CLARIN.SI.

[Ulčar and Robnik-Šikonja2020] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: Less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020*, page 104–111.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

[Verdonik et al.2013] Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47(4):1031–1048.

[Verdonik et al.2020] Darinka Verdonik, Sandi Majninger, Kaja Dobrovoljc, Špela Antloga, Aleksandra Zögling Markuš, Ines Voršič, Melita Zemljak Jontes, Melita Koletnik, Alenka Valh Lopert, Polonca Šek, Iztok Kosem, Majhenič Simona, and Ferme Marko. 2020. Korpus mladinske književnosti MAKS.

[Wilcoxon et al.1970] Frank Wilcoxon, SK Katti, and Roberta A Wilcox. 1970. Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259.

[Zwitter Vitez et al.2022] Ana Zwitter Vitez, Mojca Brglez, Marko Robnik Šikonja, Tadej Škvorc, Andreja Vezovnik, and Senja Pollak. 2022. Extracting and analysing metaphors in migration media discourse: towards a metaphor annotation scheme. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2430–2439.