

Multilingual Metaphor Detection on the LCC Datasets

COSI 114A: Fundamentals of NLP

Anna Brežgis

annabrezgis@brandeis.edu

brezgis.com

Abstract

I classify source/target word pairs in the LCC Metaphor Datasets as metaphorical or literal, in English, Spanish, Russian and Farsi. A sentence-level model that cannot see which pair is being judged reaches test F1 0.59 on English. Building the representation around the pair, with handcrafted features and then frozen and fine-tuned encoders, raises this step by step. A multilingual mDeBERTa model fine-tuned on all four languages at once, with the pair marked in the text, reaches F1 0.83 (English), 0.74 (Spanish), 0.86 (Russian) and 0.68 (Farsi).

1 Introduction

Metaphor is one of the most fascinating phenomena in natural language. A sentence like “the foundation of our government is strong” makes an underappreciated but compelling rhetorical move, discussing abstract concepts in terms of concrete ones (Lakoff and Johnson, 1980). From a computational perspective, this raises an interesting challenge. Can we model these relations without flattening the inference?

In this project, I built a binary classifier to distinguish metaphorical from literal word pairs using the LCC datasets. Each example is a source word and a target word inside a sentence, in English, Spanish, Russian or Farsi. I started from a sentence-level baseline, then worked through handcrafted features of the pair and frozen encoder representations, and ended with a multilingual mDeBERTa model (He et al., 2023) fine-tuned on all four languages at once.

2 Data

I used the Language Computer Corporation (LCC) Metaphor Datasets (Mohler et al., 2016), specifically the small English bundle as the benchmark and the large English, Spanish, Russian and Farsi bundles for training and for the other languages.

Bundle	Pairs	Metaphor	Test pairs
en_small	8,724	35%	866
en_large	52,118	39%	(en_small)
es_large	42,165	26%	4,284
ru_large	22,978	38%	2,199
fa_large	26,685	33%	2,652

Table 1. Data after filtering and binarization. All English models are tested on the en_small test set, and the other bundles are split 80/10/10 by document.

The English data was sourced from the ClueWeb09 corpus and the Debate Politics online forum. Each instance in the dataset is a source/target word pair inside a sentence, as in “the **War** on **Poverty**”, and has scores indicating how metaphorical annotators judged that pair to be. Scores of -1 indicate invalid or unclear cases. For valid instances, higher scores (up to 3) indicate stronger metaphoricality. I use only the human scores, although the large bundles also contain system scores.

I filtered out all of the pairs whose mean score is below 0, leaving 8,724 valid pairs in the small English bundle. I then binarized the labels. Pairs with mean score ≥ 2 were labeled as metaphorical (1) and pairs with mean score < 2 were labeled as literal (0).

I split the small English bundle 80/10/10 into training, development, and test sets (7,016 / 842 / 866 pairs), grouped by document so that no document appears in more than one split. The large English bundle contains the small one, so it inherits the same split by document, and any test sentence repeated in another document is kept out of training. The Spanish, Russian and Farsi bundles are each split 80/10/10 by document. The class imbalance (roughly 2:1 literal to metaphorical in English) is notable and affects model performance, so the pair-level models below use class weighting.

Some pairs were scored by two annotators, which gives a sense of how hard the task is for

people. Scoring one annotator against the other gives F1 0.66 in English, 0.60 in Spanish, 0.82 in Russian and 0.39 in Farsi, with Cohen’s κ of 0.51, 0.48, 0.69 and 0.12. The Farsi labels are close to noise.

2.1 Examples from the data

Literal examples (source $\rightarrow$ target)

- “Taxes on ‘the rich’ are **going up**...” (*going up* $\rightarrow$ *Taxes*)
- “When Kennedy **cut taxes**...” (*cut* $\rightarrow$ *taxes*)

Metaphorical examples

- “Real enjoyment is at the **heart** of **faith**.” (*heart* $\rightarrow$ *faith*)
- “Thank Lyndon Johnson, his Great Society, and the **War** on **Poverty**.” (*War* $\rightarrow$ *Poverty*)

Looking at the data, it becomes clear that one source was a political debate website. A noticeable amount of the data references some kind of political or governmental language.

These examples illustrate the challenge. Metaphorical pairs are not obviously different in structure or vocabulary from literal ones. The distinction often depends on subtle semantic relationships between words. A single sentence can also contain several pairs with opposite labels, so a model has to know which pair it is judging.

3 Cross-Validation Results

I experimented with four main approaches, each scored with 5-fold cross-validation over the English training and development sets, grouped by document. Everything learned from data (vocabularies, scalars, SVD) was fit inside each fold, and the test sets stayed locked until the end.

1. Sentence baseline. Bag-of-words and length features of the whole sentence.
2. Handcrafted pair features. Lemmas, part of speech, character n-grams and a ± 3 -word window around the source and target, WordNet supersenses, concreteness norms (Brysbaert et al., 2014), and language-model signals such as surprisal.
3. Frozen encoders. Vectors for the source and target spans from frozen encoders, plus the source word encoded on its own, a contrast in

Features (logistic regression)	CV F1
Sentence BoW + length	.538
Frequency / concreteness	.502 / .525
Dependency neighbourhood / ± 3 window	.553 / .562
LM signals	.577
WordNet supersenses + polysemy	.578
Char n-grams of source/target	.677
Lexical (lemmas, POS, concept, chain)	.692
All handcrafted	.713
+ sentence BoW	.695
Frozen mDeBERTa spans	.772
with RBF SVM	.782

Table 2. Cross-validated F1 on en_small training and development data. Frozen encoder rows use the source, target and isolated source vectors.

the spirit of the Metaphor Identification Procedure (Steen et al., 2010) as used in MeIBERT (Choi et al., 2021).

4. Fine-tuned mDeBERTa. The pair is marked in the text as « source » and [[target]], and the classifier reads [CLS] together with the mean of the marked source and target tokens.

Table 2 shows that which word is the source carries most of the signal, and that sentence context hurts. Among frozen encoders, middle layers beat the last layer (BERT 0.751 at layer 5 versus 0.723 at layer 12), and mDeBERTa-v3-base at layer 7 was the best, ahead of BERT and RoBERTa (0.75), XLM-R base and large (0.744 / 0.755) and mBERT (0.73). Nonlinear classifiers gained little over logistic regression, about 0.01 for an RBF SVM on the encoder vectors. Training on the large English bundle (about 47k pairs per fold instead of about 6.3k) helped much more, raising all handcrafted features from 0.713 to 0.779 and an MLP on mDeBERTa vectors from 0.770 to 0.804.

For the fine-tuned model I used mdeberta-v3-base with AdamW (learning rate 2×10^{-5} , batch size 32), a class-weighted loss and 3 epochs, choosing the epoch by development F1. It was trained on the training splits of all four large bundles, with three seeds, and also on English alone for comparison.

4 Test Set Results

A single script scored every finalist on the test sets exactly once, with 95% bootstrap confidence intervals.¹ Table 3 shows the English results, and

¹The one exception is an earlier set of sentence-level models, which were scored on the English test set before it was

Model	Train	F1	95% CI
Sentence BoW + length, balanced	small	.588	.544–.631
All handcrafted, LR	small	.717	.674–.756
Frozen mDeBERTa spans, LR	small	.772	.734–.809
Frozen mDeBERTa spans, LR	large	.809	.775–.840
Frozen mDeBERTa spans, MLP	large	.812	.778–.844
Fine-tuned, English only	large	.833	.800–.862
Fine-tuned, joint (3 seeds)	large×4	.827	≈.79–.86
<i>Annotator vs. annotator</i>		.66	

Table 3. English test F1 (866 pairs). “Small” and “large” are the English training bundles.

	Precision	Recall	F ₁	Acc.	N
English	.776 ± .007	.885 ± .012	.827 ± .002	.870 ± .001	866
Spanish	.694 ± .016	.786 ± .014	.737 ± .003	.855 ± .004	4,284
Russian	.833 ± .002	.883 ± .008	.857 ± .003	.889 ± .002	2,199
Farsi	.618 ± .003	.760 ± .010	.682 ± .003	.765 ± .001	2,652

Table 4. Test results of the joint fine-tuned model, as mean ± standard deviation (population) over three seeds. Metaphor is the positive class.

Table 4 shows the joint fine-tuned model in every language.

The best English model is the English-only fine-tuned one (0.833), and the joint model is close behind (0.827) while also being the best in Spanish, Russian and Farsi (Table 6).

5 Discussion

The results reveal interesting patterns about what information matters for metaphor detection. A sentence-level bag-of-words model achieves a reasonable performance by capturing lexical associations. For example, common words like “cut” or “break” may appear more often in metaphorical contexts when applied to abstract concepts (“cut ties with them”, “break a promise”). The same holds at the pair level, where lexical features of the source and target are the strongest handcrafted group (CV F1 0.692). The bag-of-words baseline, however, reaches only test F1 0.588, because it represents the whole sentence rather than the specific semantic arithmetic between words that characterizes metaphor. This is exactly why representing the pair helped. Handcrafted pair features alone take test F1 from 0.588 to 0.717, and encoder vectors of the source and target spans take it to 0.772 with the same amount of data.

Theory-driven signals behave as the literature predicts. Very concrete source words are metaphors locked. They are not reported here.

	Pred Literal	Pred Metaphor
True Literal	488 (56.4%)	73 (8.4%)
True Metaphor	39 (4.5%)	266 (30.7%)

Table 5. Confusion matrix for the joint fine-tuned model on the English test set (seed 0).

Model	en	es	ru	fa
Frozen + LR, in-language	.801	.708	.833	.670
Frozen + LR, English only	.801	.663	.746	.649
Frozen + LR, joint	.797	.701	.815	.676
Fine-tuned, English only	.833	.669	.733	.642
Fine-tuned, joint (3 seeds)	.827	.737	.857	.682

Table 6. Test F1 by training regime. “English only” rows are zero-shot on Spanish, Russian and Farsi.

50% of the time versus 21% for very abstract ones, and metaphorical sources are much less predictable from context (mean masked-LM log-probability -7.2 versus -5.2). They are small, though, and mostly overlap with knowing the lemma. Overall, the representation and the amount of training data mattered far more than the choice of classifier.

The fine-tuned model reaches 87.0% accuracy on English, against about 65% for always predicting literal. Its errors are not evenly split. False positives (literal pairs predicted as metaphorical) are 8.4% of the test set and false negatives are 4.5%, so the model leans toward calling pairs metaphorical, and recall is higher than precision in every language.

Joint training across languages helps only when the encoder is fine-tuned. With frozen features, a model trained on English alone keeps 90–97% of the in-language F1 on the other languages, but joint training does not improve on in-language training (it is lower in Spanish and Russian and only 0.006 higher in Farsi). With fine-tuning, the joint model is best in every non-English language, perhaps because a trainable encoder can absorb differences between the four annotation teams.

Per-language scores track annotator agreement. Russian has the highest agreement (κ 0.69) and the best scores. Every language beats one annotator scored against another, because the model learns the average annotator rather than any one person’s judgments. With a Farsi κ of only 0.12, though, the Farsi labels limit what any Farsi number means. Finally, with 866 English test pairs, the confidence interval of frozen mDeBERTa with an MLP (0.778–0.844) overlaps that of the fine-tuned models, so the two are not clearly separable on English. All results come from a single split, and differences

under about 0.02 should not be over-read.

6 Conclusion

I built a binary metaphor classifier for source/target pairs in four languages, comparing a sentence-level baseline, handcrafted pair features, frozen encoders and a fine-tuned multilingual encoder. The jointly fine-tuned mDeBERTa model achieved the best overall results, with test F1 0.83 in English, 0.74 in Spanish, 0.86 in Russian and 0.68 in Farsi, compared with 0.59 for the sentence-level baseline in English.

The biggest lesson is that the model has to see the pair it is judging. After that, more data and a better encoder mattered far more than the classifier on top.

With more time, the natural next step is evaluating on the VU Amsterdam Metaphor Corpus (VUA), annotated with MIPVU (Steen et al., 2010), which is the standard word-level benchmark, so the model can be compared with established state-of-the-art systems. LCC’s pair-level framing is not directly comparable to VUA’s token-level tagging, so this would require adapting the task format as well.

Code Availability

The code and instructions to reproduce every number are at <https://github.com/brezgis/metaphor-classifier-cosi114a>. The LCC data is licensed CC BY-NC-SA 4.0 and is not re-distributed there.

AI Use Statement

The extension and re-run of this project were developed using Claude Code with Claude Opus 5.5, and I reviewed the code.

References

- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.
- 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, Online. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations (ICLR)*.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.
- Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. [Introducing the LCC metaphor datasets](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4221–4227, Portorož, Slovenia. European Language Resources Association (ELRA).
- Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins, Amsterdam.