Lemmatization of Polish Multi-word Expressions

Aleksander Smywiński–Pohl
Computer Science Institute
AGH University of Krakow, Poland
apohllo@agh.edu.pl

Magdalena Król
Computer Science Institute
AGH University of Krakow, Poland
magdakrol@agh.edu.pl

Paweł Lewkowicz
Computer Science Institute
AGH University of Krakow, Poland
pawel.lewkowicz@agh.edu.pl

Zbigniew Kaleta
Computer Science Institute
AGH University of Krakow, Poland
zkaleta@agh.edu.pl

Abstract

This paper explores the lemmatization of multi-word expressions (MWEs) and proper names in Polish – tasks complicated by linguistic irregularities and historical factors. Instead of using rule-based methods, we apply a machine learning approach with fine-tuned `plT5` and `mT5` models.

We trained and validated the models on enhanced gold-standard data from the 2019 PolEval task and evaluated the impact of additional fine-tuning on a silver-standard dataset derived from Wikipedia. Two setups were tested: one without context, and one using left-side context of the target MWE.

Our best model achieved 86.23% AccCS (Accuracy Case-Sensitive), 89.43% AccCI (Accuracy Case-Insensitive), and a combined score of 88.79%, setting a new state-of-the-art for Polish MWE and named entity lemmatization, as confirmed by the PolEval maintainers. We also evaluated optimization and quantization techniques to reduce model size and inference time with modest quality loss.

1 Introduction

Polish has a complex inflectional system and this naturally impacts lemmatization. For example, the Polish noun *język* (language) appears as *język**ami*** in: *Posługując się różnymi* język**ami***, ludzie mogą komunikować emocje i koncepcje wyrażane w różnych kulturach i przez różne pokolenia* (eng. *With languages, people are able to communicate complex ideas and emotions across cultures and generations*). Though present in English to a limited extent (e.g., plural nouns, verb tenses), such variation is more pronounced in languages like Polish. Lemmatization – a key NLP task – maps inflected forms to base forms (lemmas), enabling consistent indexing in applications like full-text search (Schütze et al., 2008), topic modeling or keyword extraction and allowing to correctly present extracted phrases in the user interface (UI). Without lemmatization lexical matching will fail: a search for *język* will not retrieve *językami*, two inflected forms will decrease topic integrity and the last but not the least, the user will see awkward extracted keywords in the UI.

However, these algorithms and systems typically treat tokens individually and struggle with multi-word expressions (MWEs; also multi-word phrases or multi-word units), which often convey non-compositional meanings. This is especially problematic in tasks like information extraction (IE), where recognizing semantically identical phrases (e.g., inflected variants of a named entity) across documents is crucial. MWEs – such as names of people, places, and organizations – must be treated as single units. Their lemmas cannot simply be derived by lemmatizing each word separately. For instance, the proper name: *Ministerstwo Nauki i Szkolnictwa Wyższego* (eng. *Ministry of Science and Higher Education*) has a structured form: only the first word inflects, while others maintain fixed morphosyntactic features. Moreover, each component, though common in isolation, is capitalized in the final lemma, due to its function as a proper name. Naive, token-by-token lemmatization would yield ***m**inisterstwo **n**auk**a** i **s**zkolnictw**o** **w**y**soki***.

This illustrates the complexity of MWE lemmatization in inflectional languages which has to take into account numerous rules and relationships existing between the words.

This paper explores the use of neural models for lemmatizing MWEs in Polish, focusing on data-driven rather than rule-based or prompt-engineered approaches. We argue that, given the abundance of training data, compact models can rival larger systems in both accuracy and effi-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22160–22169
November 4-9, 2025

ciency (Kocoń et al., 2023). Our study targets MWEs including proper names (people, locations, institutions) and common expressions functioning as linguistic units.

We address the following research questions:

RQ1. **Performance of Text-to-Text Models:** How well do text-to-text models handle Polish MWE lemmatization?

RQ2. **Impact of Context:** How does adding context affect model performance?

RQ3. **Transfer Learning Benefits:** Can silver-labeled data enhance lemmatization quality?

RQ4. **Interpreting Place Names:** How effectively can models lemmatize and interpret Polish place and proper names?

RQ5. **Efficiency Optimizations:** How do techniques like quantization influence performance and accuracy?

The paper is structured as follows: Section 2 reviews related work on MWE lemmatization, especially in Polish. Section 3 explains the rules and challenges of MWEs' inflection. Section 4 outlines the datasets and models used. Section 5 presents experiments, results, and answers to the research questions. Section 6 concludes the paper.

2 Related work

In the related work section we first discuss recent advances in the domain of multi-word lemmatization, focusing on papers that take into account inflectional languages and then we concentrate on works which discuss Polish MWE lemmatization as the primary task.

2.1 Recent Multi-word Lemmatization Approaches

Shortest Edit Scripts (Myers, 1986), based on Levenshtein distance (Levenshtein, 1966), use basic string operations – insertions, deletions, and substitutions – to transform inflected forms into lemmas. These scripts can be hand-crafted or learned from form-lemma pairs, and selecting the correct one is essential. Toporkov and Agerri (2024) evaluate various script types and extraction methods.

Some approaches use hand-made or automatically derived rules. For example, Jongejan and Dalianis (Jongejan and Dalianis, 2009) improve lemmatization with rules that modify not only suffixes, but also prefixes and infixes. Their patterns, such as *ge*a*d → ***en (where asterisk is a wildcard), are tested across 10 languages with varying inflection levels, from English to Slovene.

A similar method is proposed for MWEs by Małyszko et al. (2018), where rules map morphosyntactic tags on both sides to guide lemmatization. Token forms are determined using Hunspell. The approach achieves 75.7% accuracy overall, and 82.1% when excluding misextracted phrases or those with unknown words.

A comparison of several then-state-of-the-art lemmatization methods, including decision trees and SVMs, on Croatian and Serbian datasets is presented in (Agić et al., 2013). Akhmetov et al. (2020) apply a random forest classifier to lemmatize words in 25 languages from six language families. Forms are vectorized using character-level tf-idf, and lemmas are encoded via character ordinals in a multiclass setting. The average accuracy is 72%, highly varying by language, from 38% for Farsi to 96% for Turkish, and for Slavic languages from 48% (Czech) to 92% (Slovak).

Stanković et al. (2016) present a system for extracting and lemmatizing MWEs from domain-specific corpora using a general-domain MWE dictionary. They identify 14 MWE classes covering 98% of cases, used to build detection transducers and lemmatization rules. Ambiguities are resolved using additional corpus occurrences. The lemmatization accuracy exceeds 95%.

Schmitt and Constant (2019) propose an encoder-decoder model using character, word, POS, and left-context embeddings. They assume a one-to-one mapping between input and output tokens[1]. Tested on Italian, French, Portuguese, and Polish, the model performs well except for Polish[2]. The accuracy ranges from 82.7–98% (French), 92.2% (Italian), 95.1% (Portuguese), 90.6% (Brazilian Portuguese), and 58.6–88.9% (Polish). Further context-sensitive neural approaches include the Universal Lemmatizer of Kanerva et al. (2020), a character-level seq2seq model that conditions generation on morphosyntactic context (UPOS, XPOS, FEATS) instead of a raw surface context window. Evaluated on 52 languages and 76 UD treebanks, it outperforms UD-

[1]This assumption is valid for Polish MWEs (Małyszko et al., 2018), but not necessarily for languages like English or French.

[2]The authors' conclusion.

Pipe Future on 62/76 treebanks with an average relative error reduction of 19%. The authors also show that autoencoder pretraining and transducer-based augmentation particularly help low-resource languages.

In the SIGMORPHON 2019 shared task (Task 2), Straka et al. (2019) modify UDPipe 2.0 by adding BERT contextual embeddings, regularizing with individual morphological categories, and (for some languages) merging corpora. Their system ranked first in lemmatization with 95.78 accuracy averaged over 107 corpora in 66 languages, and second in morphological analysis (93.19). Ablations indicate consistent gains from BERT and feature-level regularization, with additional benefit from careful ensembling and corpus merging.

2.2 Lemmatization of Multi-word Entitites in Polish

Marcińczuk (2017) introduce PoLem, a rule-based tool for lemmatizing Polish multi-word noun phrases and named entities[3]. It assumes unambiguous morphological tags and known entity types. The system uses 27 handcrafted transformation rules, each defined by a *head* (tag constraints) and a *body* (tag modifications), with final lemmas generated by Morfeusz (Woliński, 2006). Datasets from the KPWr corpus include ~4,000 noun phrases and over 21,000 named entities. Additional resources include Morfeusz SGJP (39k+ geographic names) and NELexicon2 (110k+ Wikipedia-based names). Reported accuracy is 97.99% for common nouns and 88.45% for named entities (86.17% with case sensitivity).

Pałka and Nowakowski (2023) describe the AMU system for the 4th SlavNER Shared Task (Yangarber et al., 2023), which addressed named entity recognition (NER), normalization, and cross-lingual linking for Polish, Czech, and Russian. The system used BERT (Devlin et al., 2018), `T5` (Raffel et al., 2020), and `plT5` models, trained on news datasets provided by organizers. For Polish lemmatization, additional datasets were used: SEJF (Czerepowicka and Savary, 2018), SEJFEK (Savary et al., 2012) (88k and 146k MWEs, respectively), and PolEval 2019 (Ogrodniczuk and Łukasz Kobyliński, 2019). SEJF and SEJFEK were machine-translated to Czech and Russian due to lack of native resources. Fine-tuned `plT5` models were used for Polish, and `mT5` (Xue et al., 2021) for other languages. Adding external datasets improved Polish performance, and PolEval data boosted multilingual results, though machine-translated lexicons degraded quality. While the Tilde system (Vīksna et al., 2023) based on XLM-RoBERTa (Conneau et al., 2020) scored highest overall, AMU outperformed it in normalization, with an F1 of 82.4% for Polish vs. Tilde's 53.9%. The authors' `T5`-based models are publicly available on HuggingFace.

[3]Named entities may also be single tokens.

3 Theoretical background

Multi-word expressions (MWEs) typically consist of a head (main segment) and one or more subordinates. The head's morphological category governs the inflection of the entire phrase. Subordinate segments relate to the head through either agreement or case governance. In agreement, a subordinate adjusts to match the head in case, number, and gender, e.g., *pann**a** młod**a***, *pann**y** młod**ej***, *pann**ie** młod**ej***... (eng. *bride*, lit. *young lady*).

In case governance, the subordinate retains a fixed form (usually genitive) regardless of the head, as in Ministerstw**o**/Ministerstw**a**/-Ministerstw**em** Sprawiedliwoś**ci** (eng. *Ministry of Justice*)

Sometimes, a subordinate agrees not with the head, but with another subordinate that is governed by the head. In *Ministerstwo Nauki i Szkolnictwa **Wyższego*** (eng. *Ministry of Science and Higher Education*) *Wyższego* agrees with *Szkolnictwo*, which itself is governed by the head.

Lemmatizing MWEs typically involves:

1. identifying the MWE boundaries,
2. finding the head,
3. analyzing relationships with subordinates,
4. assigning morphosyntactic tags,
5. retrieving base forms.

Our focus is on MWEs primarily made of nouns. The lemma is usually in the singular nominative form, though some plural phrases are exceptions, e.g., *prac**e** budowlan**e*** (eng. construction works) is preferred over theoretically correct singular *prac**a** budowlan**a***.

Adjectives are generally lemmatized in masculine singular form. However, in artistic titles, feminine forms may be correct if the work content calls

for it. E.g., *Rozważn**a** i romantyczn**a*** (eng. *Sense and Sensibility*) – although dictionaries list masculine forms for each of the segments (*rozważn**y***, *romantyczn**y***), the phrase functions as a recognized compound title and should be lemmatized accordingly.

4 Approach

While lemmatization can be approached with rule-based methods, we chose a machine learning solution due to the vast variety of MWEs and numerous exceptions, such as plural lemmas or non-masculine adjectives. Capitalization also poses challenges that are difficult to handle with rules alone.

Machine learning, however, depends on annotated datasets. Fortunately, corpora of inflected and lemmatized MWEs are available for both training and evaluation. We opted for transformer-based models pre-trained on large text corpora, expecting them to learn common patterns and reflect the linguistic biases of frequently used MWEs.

We used two datasets:

- a *silver-standard* dataset built from Wikipedia,
- a *gold-standard* dataset from the 2019 PolEval MWE lemmatization task.

The gold dataset, being task-specific, served as the main resource for evaluation, while the silver dataset was used for initial fine-tuning. This setup also allowed us to explore the benefits of transfer learning (RQ3).

4.1 Dataset 1 – Silver-standard Dataset Based on Wikipedia Links

Supervised and semi-supervised machine learning methods are commonly used in NLP, requiring data annotated with correct answers. The most reliable source is *gold-standard* corpora – carefully curated by human experts, often evaluated through inter-annotator agreement. However, a more scalable alternative is the use of *silver-standard* corpora. These are produced automatically, either by combining outputs of multiple systems (similar to ensemble methods) or by repurposing data from tasks with related structure.

In our work, we employed a silver-standard dataset derived from a semi-structured Wikipedia corpus, originally prepared for PolEval 2019 Task 3 (entity linking) (Smywiński-Pohl, 2019).[4] This dataset takes advantage of Wikipedia's internal links, where the visible link text may differ from the title of the linked page. An example of such a link is shown in figure 1.

Wikipedia links occur in two formats. If the visible caption matches the target page title, the format is `[[PageTitle]]`. If it differs, e.g., due to inflection, abbreviation, or stylistic adjustment, the format `[[PageTitle|LinkText]]` is used. For example, in *Warszawa jest stolicą Polski* (eng. *Warsaw is the capital of Poland*) *Polski* links to *Polska* (eng. *Poland*), forming `[[Polska|Polski]]`. Similarly: *Polska dołączyła do NATO w 1999 roku* (eng. *Poland joined NATO in 1999*) produces the link `[[Pakt Północnoatlantycki|NATO]]`.

The file `links.tsv`, included with PolEval Task 3 data, lists these pairs, separating `PageTitle` and `LinkText` by tabs. Since its original purpose was entity linking, many entries were unsuitable for MWE lemmatization, for example, abbreviations like `NATO` or incomplete event names.

We applied several filtering steps:

1. **Noise removal:** Entries with non-alphabetic characters (e.g., `:$1`, `004-2005)Vancouver Canucks`[5]) and disambiguation notes (e.g., *Coulomb (crater)*) were removed.
2. **Single-word exclusion:** Titles composed of a single token were excluded, as they are not MWEs.
3. **Token mismatch:** Pairs with differing token counts were removed. These often arise from abbreviation (*NATO*), shortened event names (e.g., *Mistrzostwa Świata w Piłce Nożnej 2014* (eng. *FIFA World Cup 2014*) → *Mistrzostwa Świata w Piłce Nożnej*), or omission of people's middle names.
4. **Heuristic filtering:** To eliminate synonyms (e.g., *struktura matematyczna* vs. *struktura algebraiczna* (eng. *mathematical structure*, *algebraic structure*)), we required that the first two letters of each corresponding token

[4] This dataset is publicly available at `https://2019.poleval.pl/index.php/tasks/task3`. No licensing information is given.

[5] This is likely a result of an error in the Wikipedia scraping process.

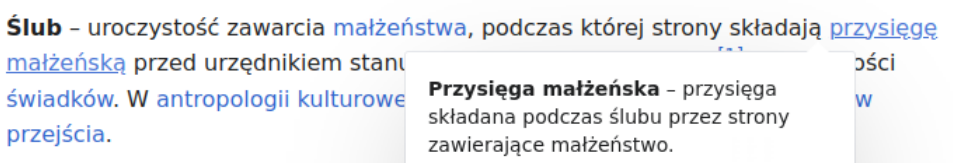

Figure 1: Example of a Wikipedia link in the Polish language. The link `[[przysięga małżeńska|przysięgę małżeńską]]` connects the inflected form *przysięgę małżeńską* (eng. *marriage vows*) with the base form *przysięga małżeńska*, which is a title of the linked page. (Screenshot from the Polish Wikipedia page `https://pl.wikipedia.org/wiki/Ślub`).

match. This works well in Polish, where inflection rarely alters word beginnings (exceptions include irregular cases like *rok* → *lata* (singular and plural of *year*)).

5. **POS tag consistency:** Pairs with mismatched parts of speech (e.g., adjective vs. noun) were excluded, even if etymologically related. For example, adjective *angielskie* (eng. *English*) should map to *angielski*, not the noun *Anglia* (eng. *England*). We used KRNNT (Wróbel, 2017) for POS tagging to enforce this.

After filtering, many valid entries still had identical inflected and base forms. This may occur when:

- the link form is in nominative (subject or part of enumeration),
- the form is accusative, which sometimes matches nominative,
- the page title is not subject to inflection (e.g. *Tadź Mahal* 'Taj Mahal').

To avoid biasing the model toward copying input, we limited such identity cases to 1/6 of the final dataset, ensuring that the majority (5/6) required transformation.

The original dataset had nearly 5.9 million entries. After filtering, about 1 million high-quality pairs remained, offering a reliable training source for silver-standard MWE lemmatization.

4.2 Dataset 2 – Gold-standard Dataset from PolEval 2019 Task 2

The gold-standard dataset was created for the 2019 PolEval MWE lemmatization task (Marcińczuk and Bernaś, 2019)[6]. It includes over 24 thousand phrase pairs from nearly 1,7 thousand documents sourced from the KPWr corpus (Broda et al., 2012). Each document is in XML format, with MWEs annotated as *phrase* elements, sometimes also marking subphrases (e.g., surnames within full names).

A corresponding TSV file provides the document ID, phrase ID, inflected form, and lemmatized form.

To improve quality and consistency, we applied the following filters:

- phrases containing digits,
- phrases where any word's first two letters (case-insensitive) differ between the inflected and lemmatized form,
- abbreviations in the format `<capital letter><dot><space><capital letter>...`.

This yielded version 0.5 of the dataset, which we split into training and validation sets in a 4:1 ratio (departing from the original split to retain more control and more training data). Version 1.0 extended the dataset by adding left context from source documents.

To better test generalization, version 1.1 removed from the validation set all expressions whose lemmas also appeared in the training set. While effective, this drastically reduced validation size. To mitigate this, version 1.2 grouped expressions by lowercased lemmas, then split these groups between training and validation. This ensured no lemma appeared in both sets.

Version 1.3 was based on manual analysis of model errors. We performed 2-fold cross-validation by splitting the training set in half and training two models. Discrepancies between model output and reference data were manually reviewed. This revealed: 79 inflection errors in the gold data vs. 51 in the model output, 14 vs. 6 capitalization errors, 13 vs. 8 cases of plural common

[6]This dataset is publicly available at `https://2019.poleval.pl/index.php/tasks/task2`. No licensing information is given.

noun lemmas, and 12 double spaces (only in gold data). Interestingly, the models produced fewer total errors (65) than the reference data (118), with lemmatization errors being the most common (67% in the gold data, 78% in model outputs). Other error types occurred at similar rates in both. This suggests that 2-fold cross-validation is an effective way to detect annotation errors in gold-standard data.

4.3 Test dataset

During experiments, we tracked model performance using validation sets from both the silver- and gold-standard datasets. For final evaluation, we followed the original PolEval setup: generating lemmatized forms for expressions in the test set (99 documents from the Polish Spatial Texts, containing 1997 phrases (Oleksy et al., 2018)). As reference answers are not publicly available, we submitted results to the task maintainer (Michał Marcińczuk) for scoring. Due to the time required, we report test results only for models that performed best on the validation sets. These results serve as our primary benchmark.

4.4 Language models

We followed a fine-tuning approach using the full transformer architecture from the `T5` model family to perform MWE lemmatization in Polish. Fine-tuned models are known to outperform larger general-purpose models on specific tasks (Raffel et al., 2020; Kocoń et al., 2023).

Raffel et al. (2020) introduced the Text-to-Text Transfer Transformer (`T5`), which frames all NLP tasks as text-to-text problems. Its architecture includes both encoder and decoder components, unlike `BERT` or `GPT`. The model was pre-trained on the Colossal Clean Crawled Corpus (C4), a large English-only dataset collected and cleaned from Common Crawl. `T5` achieved state-of-the-art results on many benchmarks, with variants up to 11 billion parameters.

The multilingual `mT5` model (Xue et al., 2021) uses the same architecture, but is trained on mC4, a Common Crawl-derived dataset spanning 101 languages over nearly six years. It includes massive variants (*xl* and *xxl*), which we excluded due to resource constraints.

The `plT5` model (Chrabrowa et al., 2022) builds on `mT5` and is further pre-trained on six Polish corpora. At publication, it achieved SOTA results on Polish tasks. Its largest version contains 1.2 billion parameters. As `T5` was trained only on English, we excluded it, but tested `mT5` (small, base, large) and `plT5` (base, large).[7]

5 Experiments

In all our experiments we have used the well-known Transformers library (version 4.44.2) to train all models. For training, we used the training script provided by the library authors.[8] We have used the same metrics as in PolEval task 4, which are case-sensitive accuracy (AccCS) and case-insensitive accuracy (AccCI). For all of the models, we report the best values achieved by the model on the validation set described in section 4.2. For some of the models, we also report the values achieved on the test dataset.[9] We report only the arguments that differ from the default values[10].

5.1 RQ1: Best Base Model for Polish MWE Lemmatization

We began by identifying the best-performing `T5`-family model for lemmatization (excluding `mT5-xl` and `mT5-xxl` due to hardware constraints). All models were trained on dataset 2 (v0.5) for 30 epochs using identical hyperparameters. We tuned the learning rate (tested: 1e-3 to 1e-5) and gradient accumulation (1 to 64). The best result was achieved with a learning rate of 5e-5 and gradient accumulation of 8.

The highest validation accuracy (AccCS) of 93.96% was obtained with `plt5-large`. The `mT5` models achieved 72.91% (small), 85.41% (base), and 89.58% (large). Given a 4.38 pp. lead over the next best model, `plt5-large` was selected for further experiments. The performance gap between base and large variants also justified using the larger model, aligning with findings from other fine-tuning studies.

Due to the strong performance, we submitted `plt5-large` to the task maintainer for evaluation on the test set. The model achieved 86.23% AccCS, 89.43% AccCI, and 88.79% Score (weighted

[7] `mT5` model is available at `https://huggingface.co/google/mt5-large` under Apache-2.0 license and `plT5` at `https://huggingface.co/allegro/plt5-base` under CC BY 4.0 license. The `mt5` models have 300 M, 580 M and 1.2 B parameters, respectively, and `plT5` models have 280 M and 820 M parameters, respectively.

[8] Available at `https://github.com/huggingface/transformers/blob/main/examples/pytorch/translation/run_translation.py`

[9] See section 4.3 for more information.

[10] The defaults can be found in `transformers/-training_args.py`

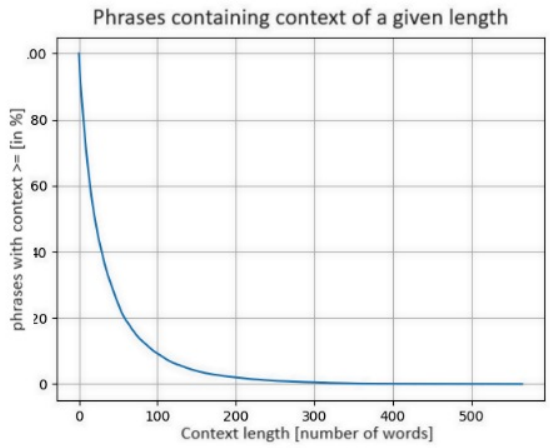

Figure 2: Context length in version `1.0` of the dataset 2.

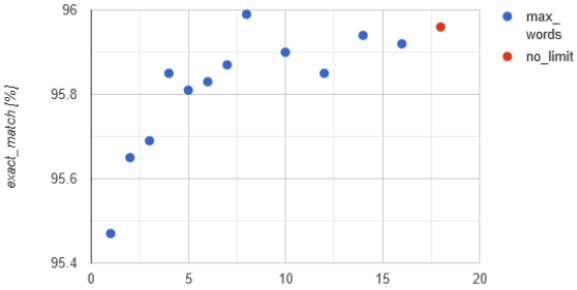

Figure 3: Results (AccCS) obtained with different context lengths during inference for dataset 2 version 1.0.

average with 0.2 and 0.8 weights), setting a new SOTA for the task. However, the drop from validation to test performance indicated possible issues with training or validation procedures.

5.2 RQ2: The Impact of the Context on the Lemmatization Quality

We conducted a series of experiments to assess how left-side context affects lemmatization quality. Context lengths varied widely, with a median of ~20 tokens and some exceeding 500. The distribution is shown in Figure 2.

To distinguish context from the target phrase during training, we inserted a special token `<c>` between them. Initially, we omitted this token from the target, which led to worse results than using no context at all. Including `<c>` in both input and target resolved this issue and improved performance.

We ran experiments on dataset 2 versions 1.0, 1.2, and 1.3 using full context. Results (Table 1) show that context improves lemmatization, with average gains of 0.57 pp. AccCI and 0.61 pp. AccCS. The performance dip in v1.2 is expected, as it uses a harder validation set with overlapping lemmas removed.

To study the effect of context length, we trained a model with full context and evaluated it using truncated contexts of various lengths. Results for v1.0 (Figure 3) show that even minimal context helps: no context yields 92.31% AccCI[11], while a single token improves it to 95.47%. Contexts of 4+ tokens consistently exceed 95.80%. Gains diminish beyond 9 tokens, suggesting that a context length of 4–9 is optimal.

5.3 RQ3: The Impact of Transfer Learning

To address the third research question, we evaluated how dataset size and transfer learning (TL) affect model performance. Results are shown in Table 2.

Models labeled "No TL" were fine-tuned directly on dataset 2 (PolEval), starting from `plT5-large`. In contrast, "TL" models were first fine-tuned on dataset 1 (see Section 4.1) and then on dataset 2. The context setting relates to the dataset 2, as dataset 1 does not contain context.

Transfer learning consistently improved performance. The largest gains occurred without context, with AccCS increasing by 0.83–1.09 pp., and AccCI by 0.92–0.93 pp. With context, improvements were smaller but still meaningful: 0.27–0.52 pp. for AccCI. While these gains may appear modest, they are significant given the already high baseline scores.

5.4 RQ4: Proper Name Lemmatization

Lemmatizing proper names can be particularly challenging due to regional linguistic variation, historical influences, and the need to avoid confusion with common nouns. To evaluate model performance on such cases, we used data from the *Państwowy Rejestr Nazw Geograficznych*[12] (National Registry of Geographical Names), which lists Polish geographical names in nominative and genitive forms.

We sampled 590 names containing at least one noun or adjective and evaluated the best models (trained with context and transfer learning on dataset v.1.2 and v.1.3). They achieved lemmatization accuracy of 87.62% and 84.06%, respectively.

[11] ommited on the chart for readability

[12] `https://dane.gov.pl/pl/dataset/780,panstwowy-rejestr-nazw-geograficznych-prng`

	AccCS		AccCI	
Dataset version	**No context**	**Context**	**No context**	**Context**
1.0	95.40	**95.99** (+0.59)	95.96	96.57 (+0.61)
1.2	92.51	93.43 (+0.92)	94.02	94.73 (+0.71)
1.3	94.76	95.33 (+0.57)	96.89	**97.30** (+0.41)

Table 1: Comparison of the context impact on the model's quality with different dataset versions.

		AccCI		AccCS	
Dataset version	**Context**	**No TL**	**TL**	**No TL**	**TL**
v 1.2	no	92.51	93.43 (+0.92)	93.43	94.52 (+1.09)
v 1.2	yes	94.02	94.54 (+0.52)	94.73	95.13 (+0.40)
v 1.3	no	94.76	95.69 (+0.93)	95.33	96.16 (+0.83)
v 1.3	yes	96.89	**97.16** (+0.27)	97.30	**97.58** (+0.28)

Table 2: The impact of transfer learning on the model's quality.

As this is the first known application of automatic lemmatization for proper names in Polish on this dataset, we cannot compare against existing results. However, manual error analysis was conducted for the better model (v.1.2). Out of 73 errors 11 were due to missing vowel alternations, e.g., *Malinowy Doł* instead of *Malinowy Dół*, *Wielki Dęb* vs. *Wielki Dąb*. 2 errors involved incorrect number: singular instead of plural, e.g., *Łąka pod Mazurem* instead of *Łąki pod Mazurem*. 5 errors resulted from incorrect gender assignment (a neuter noun misinterpreted as masculine or feminine), e.g., *Jarowyszcz Niżny* instead of *Jarowyszcze Niżne*. And finally 2 cases failed due to incorrect adaptation of the foreign origin name to Polish morphology, e.g., *Szyroka Droga* instead of *Szyroka Droha*.

5.5 RQ5: Quantization and Optimization

Quantization reduces model size and speeds up inference by converting parameters from 32-bit floats to lower-precision formats (e.g., 8-bit integers), usually with only minor quality loss. While model size can shrink 4× and inference time 2× and more[13].

We tested two quantization methods for the `plT5-large` model: ONNX Runtime with AVX-512 instructions and PyTorch quantization to `qint8`.

[13]Theoretically it should also be 4×, but some slowdown may occur due to required conversions, when hardware lacks native low-precision support

The original model occupies ~3 GB RAM. ONNX reduced this to ~1.1 GB, PyTorch to ~0.9 GB. Inference was also faster. For batch size 25, inference times were: ONNX – 9.5 s, PyTorch – 11.4 s, baseline – 20.5 s; for batch size 50: 10.3 s, 10.4 s, and 17.2 s respectively.

Quality loss varied: ONNX yielded 87.94% (4.56 pp. below baseline at 92.50%), while PyTorch dropped to 74.36% (–18.14 pp.) and couldn't be saved to disk – quantization would be required at each load.

We also tested ONNX graph optimization at O3 level, both standalone and combined with quantization. Standalone optimization had no effect on quality, slightly increased size, and improved speed (down to 4.9 s and 0.9 s for batch sizes 25 and 50). Combined with quantization, it slightly improved performance (–1.0 s for batch size 25, +1.2 s for 50) but reduced quality by 0.88 pp. relative to ONNX quantized, or 5.44 pp. below baseline.

Finally, BetterTransformers slightly improved speed but had no impact on size or quality. However, this model was also unsavable to disk.

6 Conclusions

We evaluated various models and fine-tuning techniques to identify the most effective solution for Polish MWE lemmatization. As in prior work, `plT5-large` proved best among models under 2B parameters, outperforming multilingual `T5-large` and smaller variants.

Adding context consistently improved performance. Notably, even a short context of 4 words yielded nearly the same accuracy (95.85%) as full context (95.93%), showing diminishing returns with longer input.

Transfer learning also led to better results. Pretraining on a silver Wikipedia-derived dataset followed by fine-tuning on gold data improved accuracy by 0.27–1 percentage point compared to training solely on the gold data. Despite the small gain, this mattered as the model approached near-perfect validation performance.

The models handled proper names well, and quantization further improved efficiency with modest performance loss. The most significant gain, however, came from enhancing the gold dataset itself. By reviewing cross-validation errors, we refined the training and validation sets, leading to state-of-the-art PolEval scores: 89.85% without context and 91.93% with context.

Still, a notable gap remains between validation and test set results. We suspect this stems from errors in the official gold dataset that were absent in our improved subsets – a hypothesis we plan to explore further.

7 Limitations

We have developed our solution for the Polish language, using Wikipedia as the primary source of data and tested our solution on a PolEval 2019 Task 2 test set, which consists of data from travel blogs, through Corpus of Polish Spatial Texts. Since the inflection (and thus lemmatization) rules are the same in all kinds of language (formal, colloquial, specialist, etc.), the model should generalize correctly. However, this assumption has not been tested. Also, we have not tested this solution on other inflectional languages, Slavic or otherwise. The system is based on pre-trained models (`mT5` and `plT5`), so any bias present in them, resulting from the choice of a training corpus, may affect the performance of our system.

References

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and morphosyntactic tagging of croatian and serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57.

Iskander Akhmetov, Alexandr Pak, Irina Ualiyeva, and Alexander Gelbukh. 2020. Highly language-independent word lemmatization using a machine-learning classifier. *Computación y Sistemas*, 24(3):1353–1364.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. Kpwr: Towards a free corpus of polish. In *Proceedings of LREC*, volume 12, pages 3218–3222.

Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch, Mikołaj Koszowski, Robert Mroczkowski, and Piotr Rybak. 2022. Evaluation of transfer learning for Polish with a text-to-text model. *arXiv preprint arXiv:2205.08808*.

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, Online. Association for Computational Linguistics.

Monika Czerepowicka and Agata Savary. 2018. Sejf – a grammatical lexicon of Polish multiword expressions. In *Human Language Technology. Challenges for Computer Science and Linguistics: 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015, Revised Selected Papers 8*, pages 59–73. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bart Jongejan and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2020. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27:1–30.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, and 1 others. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet Physics-Doklady*.

Jacek Małyszko, Witold Abramowicz, Agata Filipowska, and Tomasz Wagner. 2018. Lemmatization of multi-word entity names for Polish language using rules automatically generated based on

the corpus analysis. In *Human Language Technology. Challenges for Computer Science and Linguistics: 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015, Revised Selected Papers 8*, pages 74–84. Springer.

Michał Marcińczuk. 2017. Lemmatization of multiword common noun phrases and named entities in polish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2017)*, pages 483–491.

Michał Marcińczuk and Tomasz Bernaś. 2019. Results of the poleval 2019 task 2: Lemmatization of proper names and multi-word phrases. In *Proceedings of the PolEval 2019 Workshop*, pages 15–21, Warsaw, Poland. Institute of Computer Science, Polish Academy of Sciences.

Eugene W Myers. 1986. An o(nd) difference algorithm and its variations. *Algorithmica*, 1(1):251–266.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2019. *Proceedings of the PolEval 2019 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

Marcin Oleksy, Jan Wieczorek, Tomasz Bernaś, and Michał Marcińczuk. 2018. Polish spatial texts (PST) 1.0. CLARIN-PL digital repository.

Gabriela Pałka and Artur Nowakowski. 2023. Exploring the use of foundation models for named entity recognition and lemmatization tasks in slavic languages. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 165–175.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Agata Savary, Bartosz Zaborowski, Aleksandra Krawczyk-Wieczorek, and Filip Makowiecki. 2012. Sejfek – a lexicon and a shallow grammar of Polish economic multi-word units. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 195–214.

Marine Schmitt and Mathieu Constant. 2019. Neural lemmatization of multiword expressions. In *Proceedings of the joint workshop on multiword expressions and wordnet (MWE-WN 2019)*, pages 142–148. Association for Computational Linguistics.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Aleksander Smywiński-Pohl. 2019. Results of the poleval 2019 shared task 3: Entity linking. In *Proceedings of the PolEval 2019 Workshop*, pages 23–36. Institute of Computer Science, Polish Academy of Sciences.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *LREC*, pages 507–514.

Milan Straka, Jana Strakova, and Jan Hajič. 2019. Udpipe at sigmorphon 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. pages 95–103.

Olia Toporkov and Rodrigo Agerri. 2024. Evaluating shortest edit script methods for contextual lemmatization. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, pages 6451–6463. European Language Resources Association (ELRA).

Rinalds Vīksna, Inguna Skadiņa, Daiga Deksne, and Roberts Rozis. 2023. Large language models for multilingual slavic named entity linking. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 172–178.

Marcin Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In *Intelligent Information Processing and Web Mining*, pages 511–520. Springer.

Krzysztof Wróbel. 2017. Krnnt: Polish recurrent neural network tagger. In *Proceedings of the 8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 386–391. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Roman Yangarber, Jakub Piskorski, Anna Dmitrieva, Michał Marcińczuk, Pavel Přibáň, Piotr Rybak, and Josef Steinberger. 2023. Slav-NER: the 4th cross-lingual challenge on recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 179–189, Dubrovnik, Croatia. Association for Computational Linguistics.