In search of efficient, parsing-free encodings of word structure:
efficacy comparison among n-grams, skippy n-grams and extended skippy n-grams against on noun classification tasks

Kow Kuroda
Medical School, Kyorin University

Abstract

This study explores efficient, parsing-free methods for encoding word structure by comparing regular n-grams, skippy n-grams, and extended skippy n-grams in the context of inflectional classification tasks for noun gender, plurality, and case. The classification was tested on the nouns of four languages: Czech, French, German, and Irish. While the outcomes were mixed and complex, the findings suggest that extended skippy n-grams (with or without boundary marking) outperform skippy n-grams, and skippy n-grams perform better than regular n-grams in terms of classification efficiency. This study provides evidence that (extended) skippy n-grams offer a more effective approach for encoding word structure.

1 Introduction

All words, or more precisely, *surface word forms*, possess internal structures. This is true even in languages where the concept of a word is difficult to define, as certain languages may not exhibit the same clear-cut distinctions between words. However, the existence of internal structures in words is irrefutable. Words often exhibit interesting properties that sentences do not have, crucially because they can be *classified*. For example, nouns in several languages display *declensions*, and adjectives often follow suit. Similarly, verbs exhibit *conjugations*, which would not be possible if words lacked internal structure.

A central question arises: How can the internal structures of words be encoded? While it is widely accepted that sentences and phrases can be *parsed*, few assert that words can be parsed in the same way. This discrepancy arises because the structure of a word is not as easily broken down into clear categories such as Noun (N), Adjective (A), Preposition (P), and Verb (V). But does this mean that words do not have internal structures? No, that is not the case. Words in many languages reveal complex internal patterns, even though these patterns may not conform to traditional parsing categories.

The challenge lies in encoding these internal structures, as there are currently no widely accepted parsing models for word structure[1)]. This research addresses this challenge by exploring parsing-free methods to encode word structure, focusing specifically on the use of skippy n-grams. Skippy n-grams, first introduced in prior work [3], are extended in this study to assess their efficiency for encoding the internal structure of words.

2 Methodology

2.1 Task

The task at hand is a word (form) classification problem, where the classifier predicts the inflectional class of a given word. Specifically, the classification involves three attributes: gender, plurality, and case. For example, in French, the noun *maison* (meaning “house” in English) is a **feminine singular** noun, while *maisons* (meaning “houses” in English) is the **plural** form of *maison* and remains **feminine**. The classifier must predict the correct gender and plurality for each form.

For the purposes of this study, decision tree (DT), random forest (RF), and neural network (NN) classifiers were used. These classifiers were optimized to the extent possible for each dataset.[2)] It is important to note that the objective of this research is not to identify the best classifi-

1) Arguably, Morfessor https://github.com/aalto-speech/morfessor [1] is one of them, but it has two problems. First, it is based on a statistical model that require an ample to train with for better performance. Second, it is a **segmentation** tool unable to handle overlaps prevalent in morphology. Prevalence is overlapping in Japanese morphology was reported in [2].

2) There seems to be no room for detailed explanation on this paper. Refer the Jupyter Notebook scripts available at: https://github.com/kow-k/ngram-based-noun-classification

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

cation method but to evaluate the most effective encoding. Each encoding was assessed based on the best performance achieved using any of the classifiers (DT, RF, or NN).

2.2 Data

A sample of 2,000 random sentences was taken from tagged corpora of Czech, French, German, and Irish, available through the Sketch Engine[6] using seed lemmas in Table 2. The data obtained was tagged with relevant inflectional attributes such as **gender**, **plurality**, and **case**. Table 1 summarizes the target attributes. These sentences were manually parsed to extract the relevant tags.

It is important to note that the construction of the training data may contain imperfections for two primary reasons. First, the tags used for annotation are not error-free and may be inaccurate. Second, certain information may be missing from the data, particularly for case, due to syncretism[7], where the same form may correspond to multiple values.

2.3 Encodings under assessment

The study compares three types of encodings: (a) regular (consecutive) n-grams (abbreviated as n-`grams`), (b) skippy n-grams (abbreviated as sk$n$`grams`), and (c) extended skippy n-grams (abbreviated as `xsk`n`grams`)[8], for various values of n (2, 3, 4). These encodings were tested across the following tasks: i) Gender, plurality, and case classification for Czech nouns ii) Gender and plurality classification for French nouns iii) Gender, plurality, and case classification for German nouns iv) Gender, plurality, and case classification for Irish nouns

Table 3 provides examples of how different types of n-grams are formed for the word "fig," based on the degree of "skippiness" (i.e., the gaps between the consecutive characters). Skippy n-grams allow for gaps between the positions of the characters, represented by character "_".

Cleary n-grams with larger n are inefficient. This limitation can be attenuated by adding inclusiveness. Note that the comparison below is the one among inclusive versions.

The inclusion of gaps in skippy n-grams makes them more flexible than regular n-grams, but also less efficient. This inefficiency can be mitigated by adding inclusiveness. This means including $(n-1)$-grams along with n-grams. Inclusive n-grams include the $(n-1)$-grams for each n. Table 3 demonstrates this for the word "fig" with inclusive n-grams.

During the exploratory stages of the experiment, it was discovered that explicitly marking word boundaries improved performance in several cases. Thus, this option was included for testing. Examples of relevant cases are shown in Table 5.

2.4 Other training parameters

For training and validation, three different data sizes were used: 1.2k, 2k, and 3k samples. For cross-validation, 10% of the data was held out as test data.

An upper limit was set on the length of words in the training data. This parameter, called the `max_doc_size`, had values of either 9 or 11 characters.

To ensure computational and cognitive efficiency, a limit was imposed on how far a gap could extend within skippy n-grams. The `max_gap_size` parameter was chosen relative to the maximum document size (i.e., `max_doc_size`). It was defined by a `max_gap_ratio`, which had values of .33, .67, or 1.00, corresponding to `max_gap_val` values of 3, 6, or 9 characters when the `max_doc_size` was set to 9.

The final parameter in the training setup concerned the inclusion of **supplementary attributes** in the encoding. If supplementary attributes were not used, words were encoded solely by n-grams. However, when supplementary attributes were included, they were added to the n-gram-based encodings. This modification often led to better performance, though it was not always effective.

3 Results

Experiments were conducted across various combinations of training data sizes (1.2k, 2k, and 3k), maximum document sizes (9 and 11 characters), and maximum gap ratios (.33, .67, and 1.00). Due to space limitations, a full report of all results is not feasible. We focus here on one specific analysis, where training was performed with a 1.2k sample, a maximum document size of 9, a maximum gap ratio of .67 (which corresponds to a maximum gap value of 6), and the use of supplementary attributes in training.

6) Sources are Project Gutengerg corpora of the four languages available at https://www.sketchengine.eu

7) Syncretism is (the term for) a situation in which different functions are expressed by the same form. Case system is notoriously susceptible to syncretism. Both in Czech and German, for example, many nouns have the same form for Accusative and Nominative.

8) This was not defined in [3]. It is worth a mention that extended skippy n-gram was designed to get skippy n-gram to incorporate the effect of boundary marking after the effect was accidentally found.

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Table 1 Target attribute values

Attribute	German	French	Irish	Czech
gender	Fem, Masc, Neu	Fem, Masc, Comm[3]	Fem, Masc	Fem, Masc{0,1}[4], Neut
plurality	Sg, Pl	Sg, Pl, Inv[5]	Sg, Pl, Inv	Sg, Pl
case	Nom, Acc, Gen, Dat	n.a.	Nom, Gen	Nom, Acc, Gen, Dat, Instr, Loc

Table 2 lemmas used for data construction

English	German	French	Irish	Czech
book	Buch	livre	leabhar	kniha
cat	Kat	chat	cat	kočka
dog	Hund	chien	madra	pes
man	Mann	homme	fear	muž
sea	Meer	mer	farraige	moře
water	Wasser	eau	uisce	voda

Table 3 *n*-gram encodings for "fig"

n	regular	skippy	extended skippy
1	f, i, g	f, i, g	f_, _i_, _g
2	fi, ig	fi, f_g, ig	fi_, f_g, _ig
3	fig	fig	fig

The results are presented in two sections: the first compares performance across languages, and the second compares performance across attributes within each language.

3.1 Language-wise comparison

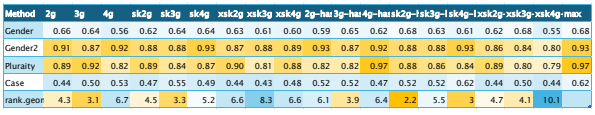


Figure 1 Czech all attributes under mgr 0.67 on 1.2k sample

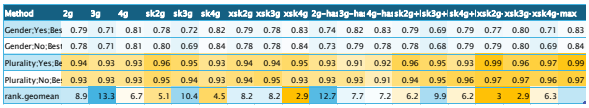


Figure 2 French all attributes under mgr 0.67 on 1.2k sample

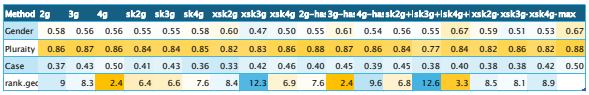


Figure 3 German all attributes under mgr 0.67 on 1.2k sample

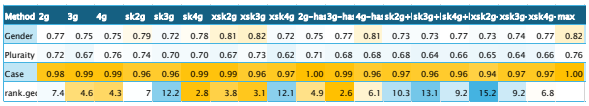


Figure 4 Irish all attributes under mgr 0.67 on 1.2k sample

Figures 1 to 4 display the language-wise accuracy distributions for Czech, German, and Irish. The accuracy distributions for each language are organized by the classification task (gender, plurality, and case) and by the encoding methods used. From the results, we observe the following: Czech: The best-performing encodings are the skippy 2-gram, skippy 4-gram, and regular 3-gram, in that order. These results highlight the effectiveness of skippy *n*-grams for this language. French: The best performers are extended skippy 4-grams and extended skippy 3-gram with hashing. The inclusion of hash-based encodings appears to boost performance significantly. German: The most

Table 4 inclusive *n*-gram encodings for "fig"

n	regular	skippy	extended skippy
1	f, i, g	f, i, g	f_, _i_, _g
2	fi, ig, f, i, g	fi, f_g, ig, f, i, g	fi_, f_g, _ig, f_, _i_, _g
3	fig, fi, ig, f, i, g	fig, fi, f_g, ig, f, i, g	fig, fi_, . . . , _i_, _g

Table 5 Non-inclusive hashed *n*-gram encodings for "fig"

n	regular	skippy	extended skippy
1	#, f, i, g, #	#, f, i, g, #	#_, _f_, _i_, _g_, _#
2	#f, fi, ig, g#	#f, #_i, . . . , _g#	#f_, #_i_, . . . , _ig_, _g#
3	#fi, fig, ig#	#fi, #f_g, . . . , ig#	#fi, #f_g, . . . , _fig_, _ig#

effective encodings are the 3-gram with hashing and the skippy 4-gram with hashing. These results demonstrate the advantage of using skippy *n*-grams in combination with hash-based encodings. Irish: The best-performing encodings include the regular 4-gram and the 3-gram with hashing. Extended skippy 3-grams also perform well, but overall, regular *n*-grams seem to be more effective for Irish.

3.2 Attribute-wise comparison

This section provides a detailed analysis of the results for plurality, gender, and case classification tasks across the languages studied. The accuracy distributions for each attribute (gender, plurality, and case) are presented in figures 9 to 15, which show how each encoding method performed across different attributes.

3.2.1 Plurality classification

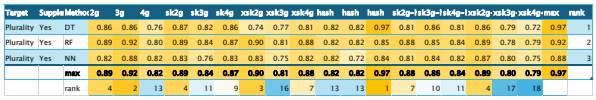


Figure 5 Czech plurality under mgr 0.67 on 1.2k

Figure 6 French plurality under mgr 0.67 on 1.2k

Figure 7 German plurality under mgr 0.67 on 1.2k

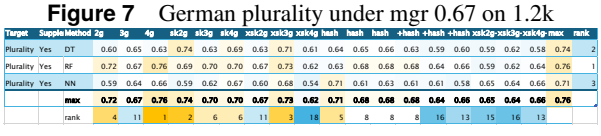


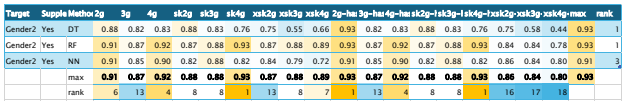
Figure 8 Irish plurality under mgr 0.67 on 1.2k sample

Figures 5 to 8 show the accuracy distributions for plurality classification of nouns in Czech, French, German, and Irish. The best-performing methods for plurality clas-

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

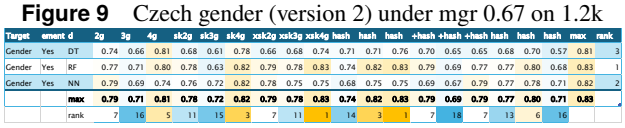
sification include: Czech: The highest-performing method is skippy 2-gram, followed by skippy 4-gram and regular 3-gram. French: Extended skippy 4-grams with hashing yield the best results, followed by regular 3-grams and skippy 3-grams. German: The most effective methods are 3-grams with hashing and skippy 4-grams with hashing. Irish: Regular 4-grams and 3-grams with hashing are the top performers, followed by extended skippy 3-grams.

3.2.2 Gender classification



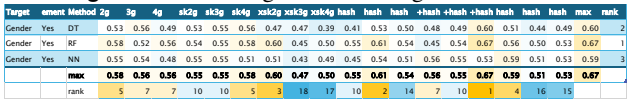
Target	Supple	Method	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	2g–ha	3g–ha	4g–ha	sk2g–l	sk3g–l	sk4g–l	xsk2g-	xsk3g-	xsk4g-	max	rank
Gender2	Yes	DT	0.88	0.82	0.83	0.88	0.83	0.76	0.75	0.55	0.66	0.93	0.82	0.83	0.88	0.83	0.76	0.75	0.58	0.44	0.93	1
Gender2	Yes	RF	0.91	0.87	0.92	0.87	0.88	0.93	0.87	0.88	0.89	0.93	0.87	0.92	0.87	0.88	0.93	0.84	0.84	0.78	0.93	1
Gender2	Yes	NN	0.91	0.85	0.90	0.82	0.88	0.82	0.84	0.79	0.72	0.91	0.85	0.90	0.82	0.88	0.82	0.86	0.84	0.80	0.91	3
		max	**0.91**	**0.87**	**0.92**	**0.88**	**0.88**	**0.93**	**0.87**	**0.88**	**0.89**	**0.93**	**0.87**	**0.92**	**0.88**	**0.88**	**0.93**	**0.86**	**0.84**	**0.80**	**0.93**	
		rank	6	13	4	8	8	1	13	8	7	1	13	4	8	8	1	16	17	18		

Figure 9 Czech gender (version 2) under mgr 0.67 on 1.2k



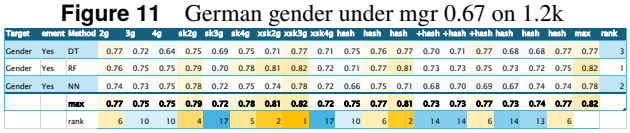
Target	ement	d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	hash	hash	hash	max	rank
Gender	Yes	DT	0.74	0.66	0.81	0.68	0.61	0.78	0.66	0.68	0.74	0.71	0.71	0.76	0.70	0.65	0.65	0.68	0.70	0.57	0.81	3
Gender	Yes	RF	0.77	0.71	0.80	0.78	0.63	0.82	0.79	0.78	0.83	0.74	0.82	0.83	0.79	0.69	0.77	0.77	0.80	0.68	0.83	1
Gender	Yes	NN	0.79	0.69	0.74	0.76	0.72	0.82	0.78	0.75	0.75	0.68	0.75	0.75	0.69	0.67	0.79	0.77	0.78	0.71	0.82	2
		max	**0.79**	**0.71**	**0.81**	**0.78**	**0.72**	**0.82**	**0.79**	**0.78**	**0.83**	**0.74**	**0.82**	**0.83**	**0.79**	**0.69**	**0.79**	**0.77**	**0.80**	**0.71**	**0.83**	
		rank	7	16	5	11	15	3	7	11	1	14	3	1	7	18	7	13	6	16		

Figure 10 French gender under mgr 0.67 on 1.2k



Target	ement	Method	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	hash	hash	hash	max	rank
Gender	Yes	DT	0.53	0.56	0.49	0.53	0.55	0.56	0.47	0.47	0.39	0.41	0.53	0.50	0.48	0.49	0.60	0.51	0.44	0.49	0.60	2
Gender	Yes	RF	0.58	0.52	0.56	0.54	0.55	0.58	0.60	0.45	0.50	0.55	0.61	0.54	0.45	0.54	0.67	0.56	0.50	0.53	0.67	1
Gender	Yes	NN	0.55	0.54	0.48	0.55	0.55	0.51	0.51	0.43	0.49	0.45	0.54	0.51	0.56	0.55	0.53	0.59	0.51	0.53	0.59	3
		max	**0.58**	**0.56**	**0.56**	**0.55**	**0.55**	**0.58**	**0.60**	**0.47**	**0.50**	**0.55**	**0.61**	**0.54**	**0.56**	**0.55**	**0.67**	**0.59**	**0.51**	**0.53**	**0.67**	
		rank	5	7	7	10	10	5	3	18	17	10	2	14	7	10	1	4	16	15		

Figure 11 German gender under mgr 0.67 on 1.2k



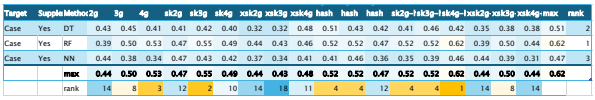
Target	ement	Method	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	hash	hash	hash	max	rank
Gender	Yes	DT	0.77	0.72	0.64	0.75	0.69	0.75	0.71	0.77	0.71	0.75	0.76	0.77	0.70	0.71	0.77	0.68	0.68	0.77	0.77	3
Gender	Yes	RF	0.76	0.75	0.75	0.79	0.70	0.78	0.81	0.82	0.72	0.71	0.77	0.81	0.73	0.73	0.75	0.73	0.72	0.75	0.82	1
Gender	Yes	NN	0.74	0.73	0.75	0.78	0.72	0.75	0.74	0.78	0.72	0.66	0.75	0.71	0.68	0.70	0.69	0.67	0.74	0.74	0.78	2
		max	**0.77**	**0.75**	**0.75**	**0.79**	**0.72**	**0.78**	**0.81**	**0.82**	**0.72**	**0.75**	**0.77**	**0.81**	**0.73**	**0.73**	**0.77**	**0.73**	**0.74**	**0.77**	**0.82**	
		rank	6	10	10	4	17	5	2	1	17	10	6	2	14	14	6	14	13	6		

Figure 12 Irish gender under mgr 0.67 on 1.2k

Figures 9 to 12 display the accuracy distributions for gender classification in Czech, French, German, and Irish. The best-performing methods for gender classification include: Czech: The extended skippy 4-gram with hashing, followed by the regular 3-gram, showed the best results. French: The best-performing method is extended skippy 3-gram with hashing, followed by regular 4-grams. German: Skippy 4-grams and extended skippy 3-grams with hashing produced the best results, with the 3-gram with hashing also performing well. Irish: Extended skippy 3-grams with hashing and skippy 4-grams produced the best performance for gender classification.

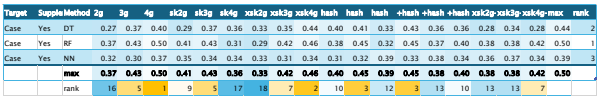
3.2.3 Case classification

Figures 13 to 15 present the accuracy distributions for case classification in Czech, German, and Irish. The results indicate the following: Czech: Skippy 2-grams with hashing and regular 4-grams are the top performers for case classification. German: The best results are obtained with 3-grams with hashing and extended skippy 4-grams.



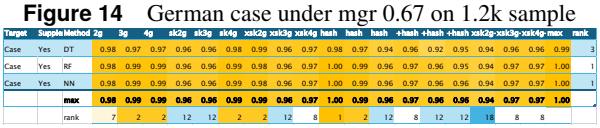
Target	Supple	Method	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	sk2g–tsk3g–tsk4g–l	xsk2g-	xsk3g-	xsk4g-	max	rank
Case	Yes	DT	0.43	0.45	0.41	0.41	0.42	0.40	0.32	0.32	0.48	0.51	0.43	0.42	0.41	0.46	0.42	0.35	0.38	0.38	0.51	2
Case	Yes	RF	0.39	0.50	0.53	0.47	0.55	0.49	0.44	0.43	0.46	0.52	0.52	0.47	0.52	0.52	0.62	0.39	0.50	0.44	0.62	1
Case	Yes	NN	0.44	0.38	0.34	0.47	0.43	0.42	0.37	0.34	0.41	0.41	0.46	0.36	0.35	0.39	0.46	0.44	0.39	0.31	0.47	3
		max	**0.44**	**0.50**	**0.53**	**0.47**	**0.55**	**0.49**	**0.44**	**0.43**	**0.48**	**0.52**	**0.52**	**0.47**	**0.52**	**0.52**	**0.62**	**0.44**	**0.50**	**0.44**	**0.62**	
		rank	14	8	3	12	2	10	14	18	11	4	4	12	4	4	1	14	8	14		

Figure 13 Czech case under mgr 0.67 on 1.2k sample



Target	Supple	Method	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-	xsk3g-	xsk4g-	max	rank
Case	Yes	DT	0.27	0.37	0.40	0.29	0.37	0.36	0.33	0.35	0.44	0.40	0.41	0.33	0.43	0.36	0.36	0.28	0.34	0.28	0.44	2
Case	Yes	RF	0.37	0.43	0.50	0.41	0.43	0.31	0.29	0.42	0.46	0.38	0.45	0.32	0.45	0.37	0.40	0.38	0.38	0.42	0.50	1
Case	Yes	NN	0.32	0.30	0.37	0.35	0.34	0.34	0.33	0.31	0.34	0.31	0.32	0.39	0.33	0.38	0.34	0.36	0.37	0.34	0.39	3
		max	**0.37**	**0.43**	**0.50**	**0.41**	**0.43**	**0.36**	**0.33**	**0.42**	**0.46**	**0.40**	**0.45**	**0.39**	**0.45**	**0.38**	**0.40**	**0.38**	**0.38**	**0.42**	**0.50**	
		rank	16	5	1	9	5	17	18	7	2	10	3	12	3	13	10	13	13	7		

Figure 14 German case under mgr 0.67 on 1.2k sample



Target	Supple	Method	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-	xsk3g-	xsk4g-	max	rank
Case	Yes	DT	0.98	0.97	0.97	0.96	0.96	0.98	0.99	0.96	0.97	0.98	0.97	0.94	0.96	0.92	0.95	0.94	0.96	0.96	0.99	3
Case	Yes	RF	0.98	0.99	0.99	0.96	0.96	0.99	0.98	0.96	0.97	1.00	0.99	0.96	0.97	0.96	0.95	0.94	0.97	0.97	1.00	1
Case	Yes	NN	0.98	0.99	0.99	0.96	0.96	0.99	0.98	0.96	0.97	1.00	0.99	0.96	0.97	0.96	0.96	0.94	0.97	0.97	1.00	1
		max	**0.98**	**0.99**	**0.99**	**0.96**	**0.96**	**0.99**	**0.99**	**0.96**	**0.97**	**1.00**	**0.99**	**0.96**	**0.97**	**0.96**	**0.96**	**0.96**	**0.97**	**0.97**	**1.00**	
		rank	7	2	2	12	12	2	2	12	8	1	2	12	8	12	12	18	8	8		

Figure 15 Irish case under mgr 0.67 on 1.2k sample

Irish: Extended skippy n-grams and 3-grams with hashing produced the highest accuracy for case classification.

In general, the best performing methods for case classification are similar to those for gender and plurality classification. Skippy n-grams, especially those extended or combined with hashing, show superior performance across different tasks and languages.

3.3 Discussion

The results obtained from these experiments provide several insights into the effectiveness of various n-gram encoding methods for word classification tasks. One key observation is that, in certain cases, ((extended) skippy) n-grams perform worse as n increases, while in other cases, they show improved results with larger n-values. This inconsistency suggests that while skippy n-grams offer flexibility in encoding word structure, the relationship between n-gram size and performance is not always straightforward and depends on the specific language and task.

4 Conclusion

This study investigated the efficacy of different n-gram-based encoding methods for word structure classification tasks, focusing on gender, plurality, and case attributes across multiple languages (Czech, French, German, and Irish). It is sugggested that (extended) skippy n-grams, when used with boundary markers and hash-based encoding, offer an effective and efficient method for word structure classification. However, further exploration is required to understand the nuances of n-gram size, encoding type, and language-specific factors in order to fine-tune these methods for optimal performance.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

Acknowledgements

To run decision tree and random forest analyses, relevant modules in `scikit-learn` (https://scikit-learn.org/) were used. To run neural network analysis, `keras` (https://keras.io/) was used. For other data analysis and visualizations, Anaconda 3 (https://www.anaconda.com) version 24.11.x was used, running Jupyter Notebook 7.0.x on Python 3.11.

References

[1] Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. Morfessor 2.0: Toolkit for statistical morphological segmentation. In **Proceedings of Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 21–24, 2014.

[2] 黒田航, 相良かおる, 東条佳奈, 麻子軒, 西嶋佑太郎, 山崎誠. LDA を使った専門用語の教師なしクラスタリング. 言語処理学会 30 回年次大会発表論文集, pp. 2858–63, 2024.

[3] Kow Kuroda. Finding structure in spelling and pronunciation using latent dirichlet allocation. In **Proceedings of the 30th Annual Meeting of the Natural Language Processing Association**, 2024.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

A Appendix: Attribute-wise analysis of mgv: 1.00 results

This appendix gives the results of another analysis with parameters `max_gap_ratio` = 1.00 on 1.2k sample.

A.1 Plurality classification

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Plurality	Yes	DT	0.89	0.78	0.82	0.76	0.74	0.87	0.79	0.78	0.84	0.91	0.86	0.88	0.73	0.84	0.83	0.82	0.81	0.80	0.91	2
Plurality	Yes	RF	0.88	0.82	0.86	0.82	0.82	0.88	0.84	0.81	0.84	0.88	0.88	0.89	0.83	0.85	0.86	0.83	0.89	0.85	0.89	3
Plurality	Yes	NN	0.81	0.75	0.81	0.72	0.74	0.82	0.88	0.78	0.92	0.84	0.78	0.80	0.82	0.82	0.85	0.82	0.82	0.74	0.92	1
		max	**0.89**	**0.82**	**0.86**	**0.82**	**0.82**	**0.88**	**0.88**	**0.81**	**0.92**	**0.91**	**0.88**	**0.89**	**0.83**	**0.85**	**0.86**	**0.83**	**0.89**	**0.85**	**0.92**	1:xsk4
		rank	3	15	9	15	15	6	6	18	1	2	6	3	13	11	9	13	3	11		

Figure 16 Czech plurality under mgr 1.00 on 1.2k

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Plurality	Yes	DT	0.93	0.85	0.86	0.85	0.89	0.89	0.95	0.82	0.90	0.88	0.82	0.91	0.93	0.90	0.87	0.86	0.82	0.89	0.95	3
Plurality	Yes	RF	0.94	0.92	0.92	0.90	0.91	0.90	0.97	0.93	0.96	0.89	0.87	0.96	0.96	0.94	0.89	0.93	0.93	0.97	0.97	1
Plurality	Yes	NN	0.91	0.88	0.91	0.93	0.95	0.93	0.97	0.95	0.96	0.89	0.88	0.88	0.96	0.97	0.91	0.93	0.97	0.97	0.97	1
		max	**0.94**	**0.92**	**0.92**	**0.93**	**0.95**	**0.93**	**0.97**	**0.95**	**0.96**	**0.89**	**0.88**	**0.96**	**0.96**	**0.97**	**0.91**	**0.93**	**0.97**	**0.97**	**0.97**	1:xsk2
		rank	10	14	14	11	8	11	1	8	5	17	18	5	5	1	16	11	1	1		

Figure 17 French plurality under mgr 1.00 on 1.2k

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Plurality	Yes	DT	0.84	0.82	0.78	0.74	0.75	0.81	0.75	0.75	0.81	0.82	0.83	0.84	0.83	0.77	0.82	0.78	0.79	0.75	0.84	3
Plurality	Yes	RF	0.82	0.87	0.83	0.83	0.80	0.83	0.78	0.79	0.85	0.79	0.85	0.88	0.87	0.86	0.81	0.82	0.79	0.82	0.88	2
Plurality	Yes	NN	0.84	0.80	0.78	0.78	0.78	0.80	0.77	0.82	0.78	0.76	0.77	0.74	0.89	0.82	0.74	0.82	0.81	0.84	0.89	1
		max	**0.84**	**0.87**	**0.83**	**0.83**	**0.80**	**0.83**	**0.78**	**0.82**	**0.85**	**0.82**	**0.85**	**0.88**	**0.89**	**0.86**	**0.82**	**0.82**	**0.81**	**0.84**	**0.89**	
		rank	7	3	9	9	17	9	18	12	5	12	5	2	1	4	12	12	16	7		

Figure 18 German plurality under mgr 1.00 on 1.2k

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Plurality	Yes	DT	0.69	0.70	0.65	0.67	0.57	0.64	0.58	0.59	0.70	0.67	0.74	0.68	0.63	0.69	0.70	0.62	0.58	0.70	0.74	2
Plurality	Yes	RF	0.65	0.69	0.66	0.68	0.60	0.72	0.63	0.65	0.66	0.72	0.75	0.71	0.69	0.72	0.74	0.64	0.63	0.68	0.75	1
Plurality	Yes	NN	0.59	0.59	0.61	0.65	0.60	0.65	0.70	0.54	0.61	0.70	0.68	0.71	0.68	0.68	0.73	0.56	0.59	0.61	0.73	3
		max	**0.69**	**0.70**	**0.66**	**0.68**	**0.60**	**0.72**	**0.70**	**0.65**	**0.70**	**0.72**	**0.75**	**0.71**	**0.69**	**0.72**	**0.74**	**0.64**	**0.63**	**0.70**	**0.75**	
		rank	11	7	14	13	18	3	7	15	7	3	1	6	11	3	2	16	17	7		

Figure 19 Irish plurality under mgr 1.00 on 1.2k sample

Figures 16–19 give the accuracy distributions for plurality of nouns in Czech, French, German and Irish. Performace of extended skippy *n*-grams, with or without hash, seem to be improved.

A.2 Gender classification

Target	Suppl ement	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	2g-hash	3g-hash	4g-hash	sk2g-hash	sk3g-hash	sk4g-hash	xsk2g+hash	xsk3g+hash	xsk4g+hash	max	rank
Gender	Yes	DT	0.58	0.58	0.51	0.57	0.49	0.56	0.57	0.57	0.40	0.57	0.59	0.53	0.47	0.44	0.50	0.52	0.44	0.46	0.59	3
Gender	Yes	RF	0.65	0.66	0.68	0.60	0.54	0.66	0.60	0.68	0.62	0.62	0.59	0.62	0.59	0.62	0.63	0.65	0.55	0.65	0.68	2
Gender	Yes	NN	0.57	0.55	0.62	0.54	0.58	0.59	0.61	0.63	0.67	0.56	0.52	0.69	0.59	0.67	0.54	0.57	0.53	0.62	0.69	1
		max	**0.65**	**0.66**	**0.68**	**0.60**	**0.58**	**0.66**	**0.61**	**0.68**	**0.67**	**0.62**	**0.59**	**0.69**	**0.59**	**0.67**	**0.63**	**0.65**	**0.55**	**0.65**	**0.69**	1:4g+h
		rank	8	6	2	14	17	6	13	2	4	12	15	1	15	4	11	8	18	8		

Figure 20 Czech gender under mgr 1.00 on 1.2k

Target	Suppl ement	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	2g-hash	3g-hash	4g-hash	sk2g+hash	sk3g+hash	sk4g+hash	xsk2g-xsk3g hash	xsk3g-xsk4g hash	max	rank
Gender	Yes	DT	0.71	0.68	0.68	0.72	0.69	0.68	0.64	0.69	0.67	0.63	0.72	0.61	0.68	0.68	0.67	0.65	0.72	0.57	0.72	3
Gender	Yes	RF	0.77	0.80	0.69	0.81	0.80	0.79	0.74	0.75	0.75	0.73	0.72	0.71	0.80	0.74	0.77	0.76	0.76	0.75	0.81	1
Gender	Yes	NN	0.72	0.79	0.69	0.79	0.78	0.73	0.75	0.76	0.73	0.72	0.68	0.68	0.79	0.74	0.77	0.79	0.72	0.69	0.79	2
		max	**0.77**	**0.80**	**0.69**	**0.81**	**0.80**	**0.79**	**0.75**	**0.76**	**0.75**	**0.73**	**0.72**	**0.71**	**0.80**	**0.74**	**0.77**	**0.79**	**0.76**	**0.75**	**0.81**	1:xsk2g
		rank	7	2	18	1	2	5	11	9	11	15	16	17	2	14	7	5	9	11		

Figure 21 French gender under mgr 1.00 on 1.2k

Target	Suppl ement	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	2g-hash	3g-hash	4g-hash	sk2g+hash	sk3g+hash	sk4g+hash	xsk2g-xsk3g hash	xsk3g-xsk4g hash	max	rank
Gender	Yes	DT	0.52	0.47	0.47	0.48	0.57	0.53	0.56	0.48	0.45	0.47	0.52	0.50	0.47	0.51	0.52	0.54	0.59	0.53	0.59	2
Gender	Yes	RF	0.57	0.54	0.57	0.61	0.57	0.57	0.61	0.54	0.55	0.56	0.54	0.61	0.59	0.50	0.62	0.56	0.61	0.58	0.62	1
Gender	Yes	NN	0.52	0.44	0.56	0.57	0.54	0.53	0.45	0.55	0.55	0.46	0.49	0.55	0.56	0.52	0.54	0.50	0.52	0.50	0.57	3
		max	**0.57**	**0.54**	**0.57**	**0.61**	**0.57**	**0.57**	**0.61**	**0.55**	**0.55**	**0.56**	**0.54**	**0.61**	**0.59**	**0.52**	**0.62**	**0.56**	**0.61**	**0.58**	**0.62**	1:sk4g
		rank	8	16	8	2	8	8	2	14	14	12	16	2	6	18	1	12	2	7		

Figure 22 German gender under mgr 1.00 on 1.2k

Target	ement	d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	hash	hash	hash	max	rank
Gender	Yes	DT	0.82	0.67	0.68	0.76	0.66	0.70	0.68	0.58	0.59	0.79	0.75	0.75	0.70	0.72	0.72	0.69	0.56	0.68	0.82	1
Gender	Yes	RF	0.76	0.77	0.74	0.78	0.71	0.81	0.78	0.68	0.70	0.75	0.73	0.75	0.78	0.74	0.79	0.75	0.68	0.70	0.81	2
Gender	Yes	NN	0.75	0.70	0.68	0.78	0.71	0.80	0.76	0.68	0.69	0.69	0.75	0.77	0.70	0.75	0.74	0.68	0.61	0.70	0.80	3
		max	**0.82**	**0.77**	**0.74**	**0.78**	**0.71**	**0.81**	**0.78**	**0.68**	**0.70**	**0.79**	**0.75**	**0.77**	**0.78**	**0.75**	**0.79**	**0.75**	**0.68**	**0.70**	**0.82**	
		rank	1	8	13	5	14	2	5	17	15	3	10	8	5	10	3	10	17	15		

Figure 23 Irish gender under mgr 1.00 on 1.2k

Figures 20 to 23 provide the accuracy distributions for gender classification of nouns in Czech, French, German and Irish. Performace of extended skippy *n*-grams, with or without hash, seem to be improved.

A.3 Case classification

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	sk2g-hash	sk3g-hash	sk4g-hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Case	Yes	DT	0.56	0.51	0.48	0.47	0.30	0.41	0.32	0.36	0.39	0.42	0.45	0.53	0.47	0.40	0.36	0.36	0.31	0.34	0.56	2
Case	Yes	RF	0.40	0.53	0.56	0.52	0.50	0.52	0.42	0.59	0.49	0.50	0.51	0.50	0.49	0.54	0.53	0.55	0.38	0.53	0.59	1
Case	Yes	NN	0.36	0.37	0.41	0.41	0.32	0.40	0.39	0.42	0.37	0.37	0.40	0.45	0.42	0.33	0.43	0.36	0.27	0.43	0.45	3
		max	**0.56**	**0.53**	**0.56**	**0.52**	**0.50**	**0.52**	**0.42**	**0.59**	**0.49**	**0.50**	**0.51**	**0.53**	**0.49**	**0.54**	**0.53**	**0.55**	**0.38**	**0.53**	**0.59**	1:xsk3
		rank	2	6	2	10	13	10	17	1	15	13	12	6	15	5	6	4	18	6		

Figure 24 Czech case under mgr 1.00 on 1.2k sample

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Case	Yes	DT	0.33	0.35	0.32	0.38	0.34	0.29	0.37	0.25	0.33	0.35	0.41	0.28	0.39	0.42	0.33	0.40	0.37	0.33	0.42	3
Case	Yes	RF	0.37	0.35	0.38	0.43	0.36	0.41	0.42	0.31	0.33	0.30	0.43	0.32	0.35	0.42	0.41	0.47	0.46	0.38	0.47	1
Case	Yes	NN	0.28	0.33	0.23	0.36	0.35	0.24	0.47	0.35	0.32	0.26	0.43	0.31	0.34	0.39	0.38	0.42	0.32	0.24	0.47	1
		max	**0.37**	**0.35**	**0.38**	**0.43**	**0.36**	**0.41**	**0.47**	**0.35**	**0.33**	**0.35**	**0.43**	**0.32**	**0.39**	**0.42**	**0.41**	**0.47**	**0.46**	**0.38**	**0.47**	
		rank	12	14	10	4	13	7	1	14	17	14	4	18	9	6	7	1	3	10		

Figure 25 German case under mgr 1.00 on 1.2k sample

Target	Supple ment	Metho d	2g	3g	4g	sk2g	sk3g	sk4g	xsk2g	xsk3g	xsk4g	hash	hash	hash	+hash	+hash	+hash	xsk2g-xsk3g	xsk3g-xsk4g	max	rank
Case	Yes	DT	0.92	0.97	0.97	0.98	0.97	0.96	0.94	0.97	0.95	0.95	0.93	0.96	0.95	0.95	0.96	0.96	0.93	0.97	0.98	3
Case	Yes	RF	0.91	0.98	0.98	0.99	0.97	0.98	0.95	0.96	0.95	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.97	0.99	1
Case	Yes	NN	0.91	0.98	0.98	0.99	0.97	0.98	0.95	0.96	0.95	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.97	0.97	0.99	1
		max	**0.92**	**0.98**	**0.98**	**0.99**	**0.97**	**0.98**	**0.95**	**0.97**	**0.95**	**0.96**	**0.96**	**0.96**	**0.96**	**0.95**	**0.96**	**0.96**	**0.97**	**0.97**	**0.99**	
		rank	18	2	2	1	5	2	15	5	15	9	9	9	9	15	9	9	5	5		

Figure 26 Irish case under mgr 1.00 on 1.2k sample

Figures 24–26 give the accuracy distributions for plurality classification of nouns in Czech, German and Irish. Like the two cases above, performace of extended skippy *n*-grams, with or without hash, seems to be improved but not quite remarkably.

A.4 Discussion

In addition to increased performance of extended skippy *n*-grams mentioned above, overall performance is improved with the full `max_gap_ratio` = 1.00. The potential problem with larger mgv is that it takes more computational resources, and it works only under reasonably shorter words. With larger `max_doc_sizes`, adverse effects are quite likely.

This work is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).