UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE
FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

Natural language processing methods

AUTHOR

Author: Mgr. Marko Penzeš
Supervisor: prof. RNDr. Michal Munk, PhD.

INTRODUCTION

Natural language processing is a challenging topic in computer science due to diversity and heterogeneity of human language. This is because computers have problem to interpret information conveyed through natural languages.

OBJECTIVE

The goal of our work was to use natural language processing methods to compare the quality of machine and human translation with regard to lexical diversity and density, as well as a deeper understanding and analysis of issues in this area.

METHODOLOGY

In this study we focus on English to Slovak translations and we have decided to specifically examine lexical diversity in human translations in comparison to machine translation. Lower lexical diversity in machine translation would suggest a less varied and "creative" output with a smaller set of translation equivalents than suggested by human translator. We used several techniques to analyze the translations, including tagging, stemming and contextual ratio analysis. To perform these analyses, we used Python programming language and various libraries including Pandas, NLTK, LexicalRichness and Stanza. To measure lexical diversity we used metrics shown in *Table 1. To measure lexical density we used metrics shown in Table 2. To statistically verify the difference in variability between the two types of translations, we used F-tests. By conducting t-test, we determined if the observed difference in lexical diversity and lexical density between human translations and machine translation is statistically significant.*

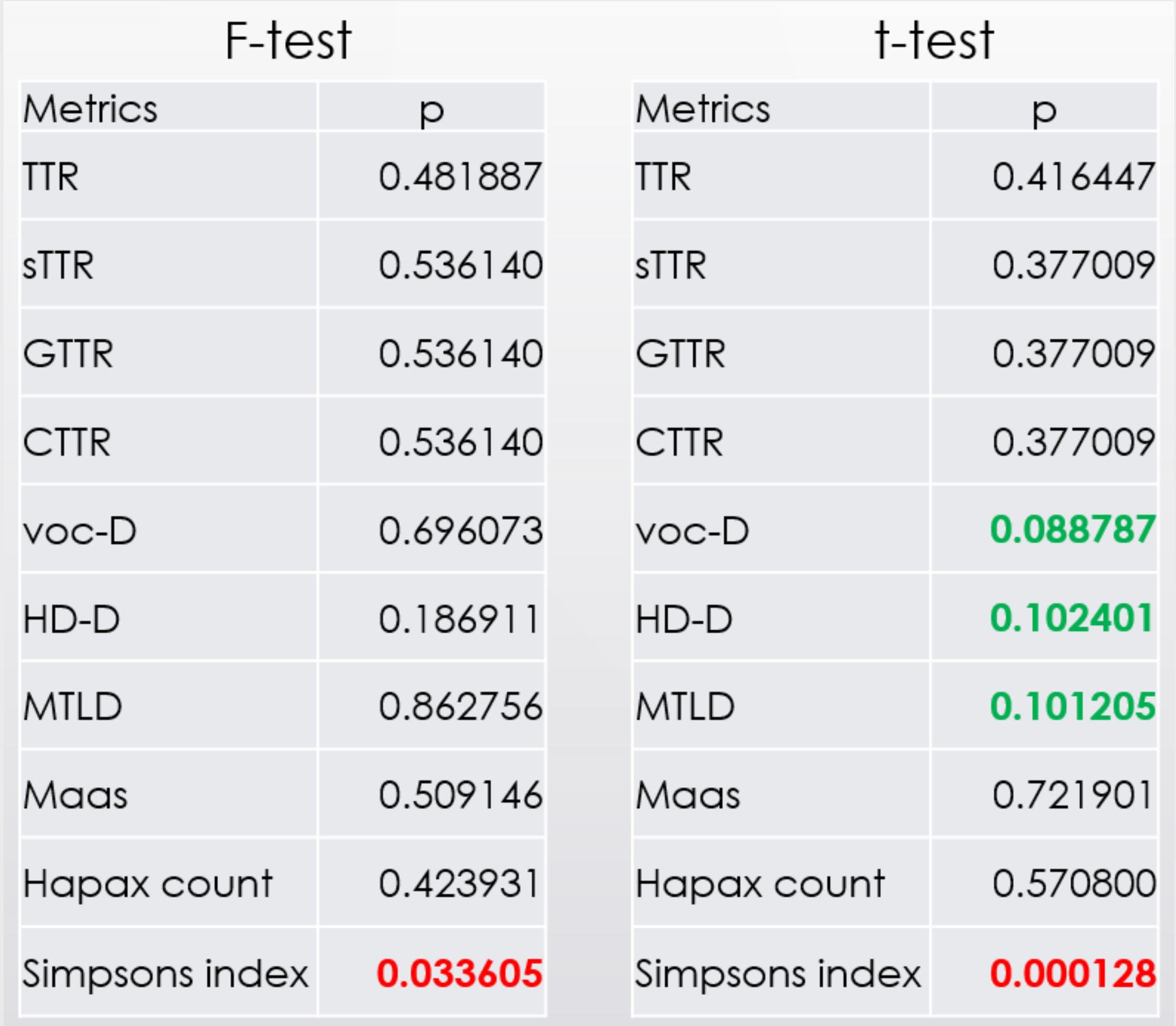
RELATED LITERATURE

Munková et al. Evaluation of Machine Translation Quality through the Metrics of Error Rate and Accuracy.

McCarthy, Jarvis MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment.

Table 1

lexical diversity



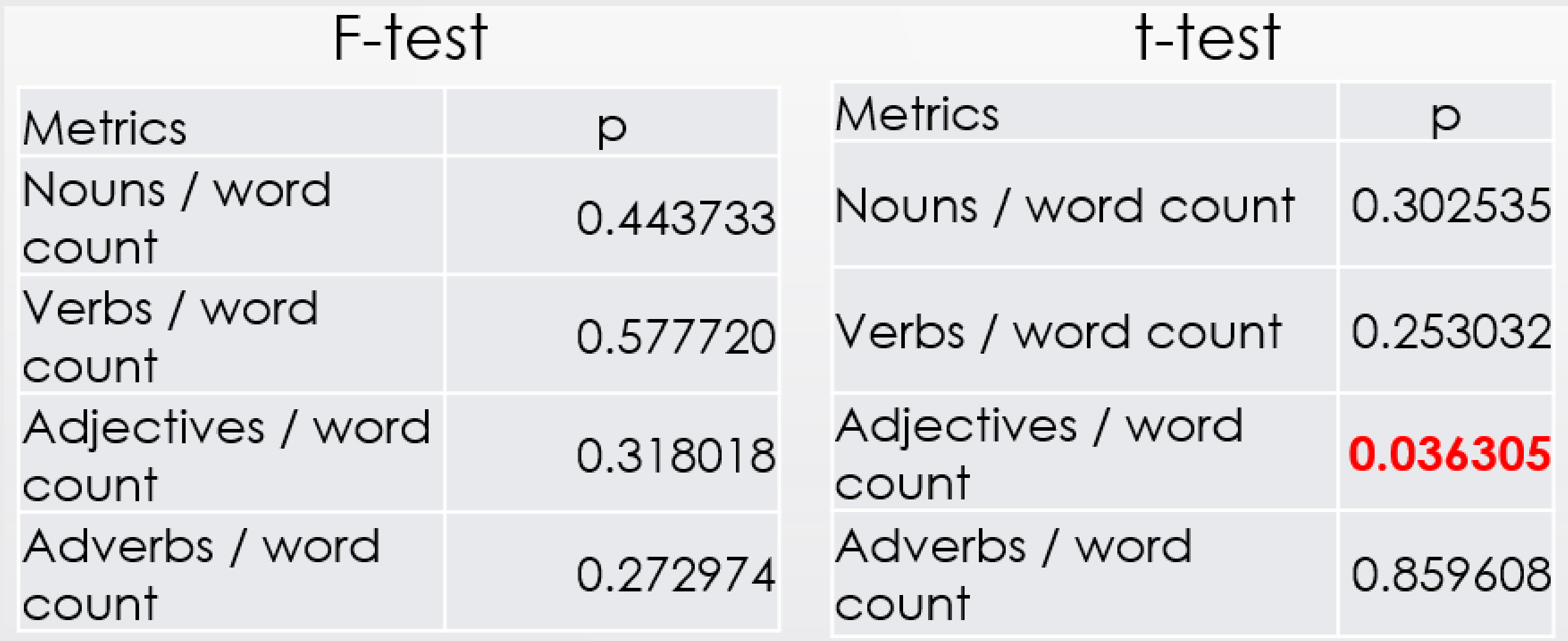
F-test		t-test	
Metrics	p	Metrics	p
TTR	0.481887	TTR	0.416447
sTTR	0.536140	sTTR	0.377009
GTTR	0.536140	GTTR	0.377009
CTTR	0.536140	CTTR	0.377009
voc-D	0.696073	voc-D	**0.088787**
HD-D	0.186911	HD-D	**0.102401**
MTLD	0.862756	MTLD	**0.101205**
Maas	0.509146	Maas	0.721901
Hapax count	0.423931	Hapax count	0.570800
Simpsons index	**0.033605**	Simpsons index	**0.000128**

RESULTS / FINDINGS

The results for lexical diversity of human and machine translations show that there was statistically significant difference in results, indicating of greater lexical diversity in human translation for Simpsons index metrics. The values HD-D, MTLD, and voc-D metrics, where the p-value ranged from 0.102460 to 0.08, indicating a large difference in results, but still not statistically significant. However, we should consider the fact that there is a high probability that with these metrics the difference would turned out to be statistically significant if we had a larger dataset. Overall, the results imply that human translations possess a higher level of lexical diversity compared to machine translations, potentially reflecting the creativity and adaptability of human translators in capturing the nuances of the source text.

Table 2

lexical density



F-test		t-test	
Metrics	p	Metrics	p
Nouns / word count	0.443733	Nouns / word count	0.302535
Verbs / word count	0.577720	Verbs / word count	0.253032
Adjectives / word count	0.318018	Adjectives / word count	**0.036305**
Adverbs / word count	0.272974	Adverbs / word count	0.859608

RESULTS / FINDINGS

We measured the lexical density of translations using the ratios of contextual words such as nouns, adjectives, adverbs and verbs to all words. Statistically significant differences were not confirmed for the metrics: ratio of nouns to all words, ratio of verbs to all words, ratio of adverbs to all words, and ratio of contextual words to all words. The *p* values were within 0.25 for the above metrics. Statistically significant differences were observed using a t-test for the metric of the ratio of adjectives to all words, where the *p* value was 0.036. We did not detect differences in variability using the F-test for the metrics mentioned above. The overall results for lexical density showed that a higher number of adjectives were used in the human translation than in the machine translation, which may indicate that the human translation tries to provide a more detailed and descriptive text compared to the machine translation. Our results suggest that there may be some differences in the lexical density of human and machine translations in terms of the proportion of adjectives, but the differences in nouns, verbs and adverbs are not statistically significant

CONCLUSION

The introduction of machine translation has marked significant progress in natural language processing. However, the quality of machine translation in comparison to human translation remains a subject of ongoing debate. This study aimed to compare the quality of machine and human translation in terms of lexical diversity and density using NLP methods. The results indicate a statistically significant difference in lexical diversity between human and machine translations, particularly for the Simpson's index metric. The metrics HD-D, MTLD, and voc-D exhibited larger differences, but the statistical significance was not established. It is suggested that with a larger dataset, these metrics may also reveal significant differences favoring human translations. Overall, the study emphasizes the importance of considering multiple metrics when evaluating translation quality. Despite the advancements in machine translation, human translation remains essential for producing high-quality translations that accurately capture the intended meaning. The findings underscore the limitations of machine translation in replicating the lexical diversity and density of human translations. The implications of this research are crucial for the field of natural language processing. Considering the limitations of this study, such as focusing on a specific language combination and relying on specific libraries that may not be fully compatible with Slovak language, further research should explore ways to improve machine translation while recognizing the importance of human involvement in the translation process.This research highlights the ongoing need for research and development in the field of language translation to achieve more precise and effective communication between cultures and languages.