Evaluation of the effectiveness of lemmatization on different languages

Student: Sako Ranj Abubakr

Supervisor: Dr. Salah Ismaeel Yahya

Koya university

Faculty of engineering

Software department

2022-2023

Contents

Introduction

Artificial intelligence and machine learning have transformed the way we interact with information. One area where these technologies have shown promise is in natural language processing, which involves the use of computational techniques to analyze and understand human language. One important technique in natural language processing is lemmatization, which involves reducing words to their base form or lemma. While lemmatization has been studied extensively in English, much less is known about its effectiveness in other languages. This research study aims to evaluate the effectiveness of lemmatization on different languages.

Methodological Approach

The proposed research study aims to evaluate the effectiveness of lemmatization on different languages through a quantitative analysis of text data. The study will focus on three languages: English, Arabic, and Spanish. The research question identified in this study is: What is the impact of lemmatization on the accuracy and efficiency of text analysis in English, Arabic, and Spanish?

Defining Research Problem

The research problem identified in this study is the lack of understanding of the effectiveness of lemmatization on different languages. While lemmatization has been studied extensively in English, much less is known about its effectiveness in other languages. This lack of understanding hinders the development of effective natural language processing tools for non-English languages.

Study Population and Sample

The study population for this research project will consist of students who have used ChatGPT for their coursework in higher education in different languages. The study will target universities located in different regions across the world to ensure a diverse range of languages are represented.

The sample size for this research project will be determined using purposive sampling technique. Participants will be selected based on the following criteria:

1. Participants must have used ChatGPT for their coursework in higher education.
2. Participants must be willing to participate in semi-structured interviews.
3. Participants should represent a diverse range of academic majors, levels of study, gender, and age.
4. Participants must be proficient in the language of study.

To ensure a variety of academic fields and levels of study, participants will be selected from different faculties and departments of the university. The sample size will be determined by reaching data saturation, where no new themes or insights are emerging from the data collected. The estimated sample size for this research project is 30 participants.

Participants will be recruited through email invitations sent to their university email accounts. The email will contain information about the research project, the purpose of the study, the expected duration of the interview, and the voluntary nature of participation. Participants who are interested in participating will be required to respond to the email, and they will be provided with a consent form to sign before the interview.

The study will be conducted in compliance with ethical considerations and participants' confidentiality will be ensured throughout the study. The interviews will be conducted in a private and confidential setting, and all data collected will be stored securely and anonymously.

Sampling Criteria

The selection of appropriate participants for the study is crucial to ensure the validity and reliability of the research findings. The following criteria will be used to guide the sampling process:

1. Participants must have used a lemmatization tool in their academic work or research for at least one semester in their respective language.

2. Participants must be willing to participate in semi-structured interviews and have a sufficient level of proficiency in the language they are studying.

3. Participants should represent a diverse range of academic fields, levels of study, age, and gender.

4. Participants should come from different universities or institutions and different regions where the target language is spoken.

To achieve the required sample size, purposive sampling will be used to recruit participants. This method will ensure that participants meet the necessary criteria and represent a diverse range of experiences and perspectives. The recruitment process will involve contacting potential participants through email, social media, or face-to-face communication. A consent form will be provided to all participants, outlining the nature of the study, their rights as participants, and the possible risks and

benefits of participating. Participants who agree to participate will be contacted to schedule a semi-structured interview at a convenient time and location.

The sample size for the study will be determined by the principle of data saturation, which refers to the point where new data no longer contributes to the research objectives. The data collection process will continue until data saturation is reached. This means that the sample size will not be predetermined, but rather determined by the point where no new information or insights are emerging from the data collected.

The study will ensure that ethical considerations are addressed in the sampling process. All participants will be informed about the nature of the research, their rights as participants, and the possible risks and benefits of participating. Confidentiality and anonymity will be maintained throughout the study to protect the participants' privacy and ensure that they feel comfortable sharing their experiences and perspectives.

Research Design

The research design for this study involves the use of a quantitative research approach. The study will involve the analysis of text data in three languages: English, Arabic, and Spanish. The effectiveness of lemmatization will be evaluated by comparing the accuracy and efficiency of text analysis with and without lemmatization.

Sampling Criteria The sampling criteria for this study are as follows:

1. Text data must be available in English, Arabic, or Spanish.
2. Text data must be of sufficient length and complexity to allow for meaningful analysis.
3. Text data must be representative of the language in question, including variations in dialect and style.
4. Text data must be publicly available or obtainable through legal means.

Data Collection

Data Collection Instruments

Several data collection instruments will be developed to gather the necessary data for the study:

1. Semi-structured interviews: Semi-structured interviews will be conducted with students who have used ChatGPT in their coursework for higher education. The interviews will explore the perceived impact of ChatGPT on their learning experiences. The interview questions will be developed based on the research questions and objectives of the study. Sample questions include:

- What were the advantages and drawbacks of using ChatGPT in your learning?
- How did ChatGPT impact your engagement and motivation in learning?
- How did ChatGPT affect your language proficiency and accuracy?
- Did ChatGPT enhance or hinder your critical thinking and problem-solving skills?
- How did ChatGPT affect your communication and interaction with classmates and instructors?
- Do you think ChatGPT should be integrated into the curriculum of your program?

2. Survey questionnaire: A survey questionnaire will be distributed to a larger sample of students who have used ChatGPT in their coursework for higher education. The survey will collect quantitative data on the perceived effectiveness of ChatGPT in improving language proficiency, accuracy, and fluency, as well as its impact on engagement, motivation, critical thinking, and problem-solving skills.

The survey questionnaire will consist of closed-ended questions with multiple response options, as well as open-ended questions for participants to provide additional comments and feedback. The survey questions will be developed based on the research questions and objectives of the study, as well as relevant literature on lemmatization and language learning.

3. Language proficiency test: A language proficiency test will be administered to the participants to assess their language skills before and after using ChatGPT in their coursework. The language proficiency test will be based on internationally recognized standards, such as the Common European Framework of Reference for Languages (CEFR), and will assess the participants' reading, writing, listening, and speaking skills. The language proficiency test will provide objective data on the effectiveness of ChatGPT in improving language proficiency.

The data collection instruments will be pilot-tested with a small sample of students to ensure their validity, reliability, and comprehensibility. The feedback and suggestions from the pilot test participants will be used to refine the data collection instruments.

Data Collection Instruments

The data collection instrument for this study will be a custom-built software tool that allows for the comparison of text analysis with and without lemmatization. The tool will be designed to accept input in English, Arabic, and Spanish, and will output metrics related to accuracy and efficiency.

Data Collection Procedure

The data collection procedure for this study will involve the following steps:

1. Identification of suitable text data in English, Arabic, and Spanish.
2. Pre-processing of text data to remove any irrelevant or redundant information.
3. Analysis of text data with and without lemmatization using the custom-built software tool.
4. Collection of metrics related to accuracy and efficiency for each analysis.
5. Statistical analysis of the data to determine the effectiveness of lemmatization in each language.
6. Validity and Reliability To ensure the validity and reliability of the study, several measures will be taken. Firstly, the custom-built software tool will be thoroughly tested to ensure accurate and consistent results. Secondly, the text data used in the study will be carefully selected to ensure that it is representative of the language in question. Finally, the statistical analysis of the data will be conducted using established methods and techniques to ensure that the results are robust and reliable.

Validity and Reliability

The proposed study will utilize a qualitative research approach, specifically thematic analysis, to investigate the perceived impact of lemmatization on different languages. To ensure the validity and reliability of the study, several measures will be taken.

Firstly, the research design and data collection methods will be carefully selected to ensure that the data collected is relevant and sufficient to address the research questions. The research design will also take into account potential biases and limitations, such as researcher bias or participant bias.

Secondly, to ensure the reliability of the study, a detailed description of the research methodology will be provided to enable other researchers to replicate the study. The research methodology will also be reviewed and evaluated by experts in the field to ensure its rigor and appropriateness.

Thirdly, the thematic analysis process will be carried out by two independent researchers who will analyze the same data set separately. The inter-rater reliability of the analysis will be calculated using a statistical measure, such as Cohen's kappa coefficient. Any discrepancies in the analysis will be discussed and resolved through consensus.

Finally, the results of the study will be triangulated by using multiple data sources and methods. This will involve collecting data from different sources, such as interviews and surveys, and comparing and contrasting the findings to ensure that they are consistent and support the research conclusions.

By implementing these measures, the study aims to ensure the validity and reliability of the research findings and enhance the credibility and trustworthiness of the study.

Ethical Considerations

1. Informed Consent: No human subjects will be involved in this study, so informed consent is not required.
2. Confidentiality: The text data used in this study will be publicly available or obtainable through legal means, so no confidentiality concerns arise.

3. Data Protection: All text data used in this study will be obtained and used in accordance with relevant data protection laws and regulations.
4. ChatGPT
5. Limitations Like any research study, this proposed investigation has limitations. The first limitation is the sampling strategy. Although purposive sampling will be used to recruit a diverse group of participants, the study will only include students from universities in the Kurdish region of Iraq, which may limit the generalizability of the findings. Another limitation is the reliance on selfreported data from participants, which may be subject to bias or error. Additionally, the study only focuses on the perceived impact of ChatGPT from the perspective of students, and does not include the perspective of instructors or administrators.

Future Directions This study provides insight into the perceived impact of ChatGPT on higher education from the perspective of students. Future research could expand on these findings by including a larger and more diverse sample, including instructors and administrators, and using a mixed-methods approach to further examine the impact of ChatGPT on higher education. Additionally, future research could investigate the effectiveness of different natural language processing techniques such as lemmatization on the performance of ChatGPT in different languages.

Contribution to Knowledge This study contributes to the knowledge on the perceived impact of ChatGPT on higher education by providing insight into the experiences and perceptions of students who have used ChatGPT in their coursework. By exploring the advantages and drawbacks of using ChatGPT in education, this study may inform the development and integration of natural language processing tools into education systems. Furthermore, this study may inform the development of language-specific tools and techniques to improve the performance and accuracy of natural language processing tools in different languages.

Data Analyses

The qualitative data gathered from the semi-structured interviews will be analyzed using thematic analysis. Thematic analysis is a method of identifying, analyzing, and reporting patterns or themes within qualitative data [25]. The aim is to identify and interpret patterns of meaning across the data set, and to develop a deeper understanding of the research question.

The data analysis process will involve the following steps:

1. Familiarization with the data: The data collected will be transcribed verbatim and read through multiple times to become familiar with the content and context.

2. Coding: Initial codes will be generated by systematically examining the data and highlighting words, phrases, or passages that are relevant to the research question. The codes will be descriptive, inductive, and grounded in the data.

3. Theme generation: The initial codes will be grouped into broader themes by looking for connections and patterns in the data. Themes will be reviewed, refined, and renamed as necessary.

4. Reviewing and defining themes: The themes will be reviewed and refined to ensure they accurately reflect the data and answer the research question.

5. Interpretation: The themes will be interpreted in relation to the research question, and implications for higher education and lemmatization will be discussed.

To enhance the validity and reliability of the analysis, several measures will be taken. Two researchers will independently code the data to ensure inter-coder reliability. Regular meetings will be held to discuss emerging themes, discrepancies in coding, and to ensure a consensus is reached. An audit trail will be kept to document the analytical process and ensure transparency.

The analysis will be conducted using NVivo 12, a qualitative data analysis software that facilitates coding, organization, and retrieval of data. NVivo 12 will enable the efficient and systematic analysis of the data set, and support the identification of themes and patterns across the data.

The results of the analysis will be presented in a descriptive and analytical manner, supported by quotes and examples from the data set. The findings will be discussed in relation to the research question and objectives, and implications for higher education and lemmatization will be explored.

Conclusion

This proposed study aims to investigate the perceived impact of ChatGPT on higher education by examining the experiences and perspectives of students who have used ChatGPT in their coursework. The study will use semi-structured interviews to gather data from participants, which will be analyzed using thematic analysis. The findings of this study may inform the development and integration of natural language processing tools into education systems, as well as the development of language-specific techniques to improve the accuracy of these tools in different languages.

References:

1. Abdulsahib, H. K. &. A.-G. B. M., 2019. Mobile-assisted language learning (MALL) in Iraq: A systematic review of recent research. Journal of Educational Technology.

2. Al-Qaysi, N. M., 2020. The Impact of Using Artificial Intelligence Applications in Teaching and Learning: A Review Study. Journal of Educational and Psychological Studies.

3. Babbie, E., 2016. The practice of social research. 14th ed. Cengage Learning.

4. Creswell, J. W. & Creswell, J. D., 2017. Research design: qualitative, quantitative, and mixed methods approaches. 5th ed. Sage Publications.

5. Denzin, N. K. & Lincoln, Y. S., 2011. The SAGE handbook of qualitative research. 4th ed. Sage Publications.

6. Guest, G., Bunce, A. & Johnson, L., 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. Field Methods, 18(1), pp. 59-82.

7. Krippendorff, K., 2013. Content analysis: An introduction to its methodology. 3rd ed. Sage Publications.

8. Patton, M. Q., 2015. Qualitative research & evaluation methods. 4th ed. Sage Publications.

9. Saldaña, J., 2015. The coding manual for qualitative researchers. 2nd ed. Sage Publications.

10. Seale, C., 2018. Researching society and culture. 4th ed. Sage Publications.

11. Smith, J. A., Flowers, P. & Larkin, M., 2009. Interpretative phenomenological analysis: Theory, method and research. Sage Publications.

12. Tong, A., Sainsbury, P. & Craig, J., 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. International Journal for Quality in Health Care, 19(6), pp. 349-357.