

الگوی نمونه برای چکیده مبسوط
(چکیده توسعه یافته Extended Abstract)

عنوان: دسته‌بندی خودکار قصدها در یک چت‌بات وظیفه‌محور در حوزه آموزش.

نام نویسندگان: الهام صالحی؛ دانشجوی دکتری زبانشناسی؛ پژوهشکده زبانشناسی؛ پژوهشگاه علوم انسانی و مطالعات فرهنگی؛ تهران؛ ایران؛

مسعود قیومی؛ دکتری زبانشناسی رایانشی؛ دانشیار؛ پژوهشکده زبانشناسی؛ پژوهشگاه علوم انسانی و مطالعات فرهنگی؛ تهران؛ ایران؛ پدیدآور رابط؛ M.ghayoomi@ihcs.ac.ir

آتوسا رستم‌بیگ تفرشی؛ دکتری زبانشناسی؛ دانشیار؛ پژوهشکده زبانشناسی؛ پژوهشگاه علوم انسانی و مطالعات فرهنگی؛ تهران؛ ایران؛

مقدمه: زبان طبیعی یکی از مهم‌ترین ابزارهای تعامل اجتماعی انسان است و گفت‌وگو نقشی اساسی در انتقال دانش، شکل‌گیری روابط اجتماعی و تداوم فرهنگ ایفا می‌کند. در زبان‌شناسی، تحلیل مکالمه و نظریه کنش گفتاری نشان می‌دهند که درک معنا در گفت‌وگو وابسته به تشخیص قصد ارتباطی گوینده است. با پیشرفت هوش مصنوعی، تلاش شده است این فرایند در سامانه‌های مکالمه‌محور مانند چت‌بات‌ها شبیه‌سازی شود. در این مسیر، معناشناسی توزیعی و مدل‌های بردارسازی واژگان امکان‌بازنمایی عددی معنا را فراهم کرده‌اند. این پژوهش با بهره‌گیری از این رویکردها به طراحی یک چت‌بات آموزشی فارسی و مقایسه الگوریتم‌های یادگیری ماشین برای تشخیص قصد کاربران می‌پردازد.

روش‌شناسی: در پژوهش حاضر، به‌منظور تشخیص قصد در یک چت‌بات آموزشی زبان فارسی، ابتدا یک پیکره تخصصی طراحی و گردآوری شد. از آنجاکه برای زبان فارسی در حوزه تشخیص قصد و پرکردن جای خالی، داده‌سنگجه مناسبی در حوزه آموزش زبان فارسی در دسترس نبود، مجموعه‌ای از جملات بر اساس تحلیل محتوای کتاب‌های آموزش زبان فارسی برای غیرفارسی‌زبانان و با مشورت متخصصان این حوزه تولید شد. داده‌های گردآوری‌شده شامل ۳۶۰۷ جمله با ۱۵۳۰۵۴ واژه بود که در چهار دسته معنایی هم‌معنایی، تضاد معنایی، باهم‌آیی و اصطلاح‌سازمان‌دهی شدند. پس از آماده‌سازی داده‌ها، مرحله پیش‌پردازش شامل اصلاح نویسه‌گذاری، فاصله‌گذاری، علائم نگارشی و یک‌دست‌سازی متون انجام گرفت و سپس برای مشخص کردن مرز عناصر معنایی از الگوی برچسب‌گذاری IOB استفاده شد. در گام بعد، جملات با بهره‌گیری از مدل‌های بردارسازی ایستای وردتووک و فست‌تکست به بردارهای ۳۰۰ بعدی تبدیل شدند تا امکان استفاده از آن‌ها در

الگوریتم‌های یادگیری ماشین فراهم شود. برای طبقه‌بندی قصد کاربران، چندین الگوریتم شامل ماشین بردار پشتیبان، جنگل تصادفی، درخت تصمیم، K-نزدیک‌ترین همسایه و دو مدل شبکه عصبی پرسپترون تک‌لایه و چندلایه به کار گرفته شد. همچنین یک شبکه عصبی پیش‌خور با دو لایه پنهان و تابع فعال‌ساز ReLU طراحی گردید. در نهایت، عملکرد مدل‌ها با استفاده از اعتبارسنجی متقابل پنج‌تایی و شاخص‌های دقت، صحت، فراخوانی و امتیاز اف‌وان ارزیابی شد.

یافته‌های اصلی: نتایج ارزیابی الگوریتم‌ها نشان می‌دهد که نوع روش بردارسازی تأثیر تعیین‌کننده‌ای بر عملکرد مدل‌های یادگیری ماشین در تشخیص قصد دارد. در بردارسازی مبتنی بر فست‌تکست، به دلیل بهره‌گیری از اطلاعات زیرواژگانی، اغلب الگوریتم‌ها عملکرد بسیار مطلوبی از خود نشان داده‌اند. در این میان، مدل پرسپترون چندلایه با دستیابی به دقت حدود ۹۹/۴۵ درصد بهترین نتیجه کل پژوهش را ثبت کرده است و نشان می‌دهد که شبکه‌های عصبی توانایی بالایی در استخراج الگوهای غیرخطی و روابط معنایی پیچیده در فضای برداری دارند. همچنین الگوریتم‌های K-نزدیک‌ترین همسایه و جنگل تصادفی نیز با دقتی نزدیک به ۹۹ درصد عملکرد بسیار بالایی ارائه کرده‌اند. در مقابل، در بردارسازی وردتووک کاهش عملکرد برخی مدل‌ها مشاهده می‌شود؛ به‌ویژه در ماشین بردار پشتیبان و مدل‌های شبکه عصبی که افت دقت قابل‌توجهی را تجربه کرده‌اند. تحلیل خطاها نیز نشان داد که بیشترین میزان اشتباه در تمایز میان روابط هم‌معنایی و تضاد معنایی رخ داده است، زیرا این دو رابطه در فضای برداری شباهت ساختاری دارند و مرزبندی میان آن‌ها برای برخی الگوریتم‌ها دشوار است. با این حال، نتایج ماتریس‌های درهم‌ریختگی نشان می‌دهد که مدل پرسپترون چندلایه در ترکیب با فست‌تکست توانسته است تقریباً تمامی نمونه‌ها را به‌درستی طبقه‌بندی کند و از نظر پایداری و قابلیت تعمیم، بهترین عملکرد را در میان روش‌های مورد بررسی ارائه دهد.

بحث و نتیجه‌گیری:

یافته‌های این پژوهش چند نکته برجسته و نسبتاً نوآورانه را در زمینه تشخیص قصد در زبان فارسی آشکار می‌سازد. نخست، نتایج نشان داد که نوع بازنمایی معنایی نقش تعیین‌کننده‌ای در عملکرد مدل‌های تشخیص قصد دارد؛ به‌گونه‌ای که بردارسازی زیرواژه‌محور فست‌تکست توانست ساختارهای واژگانی و صرفی زبان فارسی را مؤثرتر از وردتووک مدل‌سازی کند و در نتیجه عملکرد بسیاری از الگوریتم‌ها به‌طور چشمگیری بهبود یابد. مهم‌تر از آن، ترکیب این بازنمایی با مدل پرسپترون چندلایه به دقتی نزدیک به ۹۹/۴۵ درصد انجامید که نشان می‌دهد شبکه‌های عصبی غیرخطی قادرند الگوهای معنایی ظریف میان مقاصد زبانی را با دقت بسیار بالا استخراج کنند. این یافته در ارتباط مستقیم با پرسش اصلی پژوهش درباره تأثیر نوع بردارسازی و معماری مدل بر تشخیص قصد قرار دارد و نشان می‌دهد که تعامل میان بازنمایی معنایی غنی و مدل‌های یادگیری عمیق می‌تواند به بهبود چشمگیر عملکرد سامانه‌های مکالمه‌محور فارسی منجر شود. از سوی دیگر، تحلیل خطاها نشان داد که بیشترین دشواری مدل‌ها در تفکیک روابط هم‌معنایی و تضاد معنایی رخ می‌دهد که بیانگر

پیچیدگی این روابط در فضای معنایی زبان است. افزون بر این، طراحی یک پیکره تخصصی در حوزه آموزش زبان فارسی برای تشخیص قصد نیز از دستاوردهای مهم پژوهش محسوب می‌شود که می‌تواند به عنوان زیرساختی ارزشمند برای توسعه سامانه‌های درک زبان طبیعی و چت‌بات‌های آموزشی فارسی در پژوهش‌های آینده مورد استفاده قرار گیرد.

کلیدواژه‌ها: (تشخیص قصد، الگوریتم یادگیری ماشین، سامانه مکالمه‌محور، بردارسازی، تحلیل گفتمان)

Suggested format/template for extended abstract

Title: Automatic Intent Classification in a Task-Oriented Chatbot in Education.

Names of Authors: Elham Salehi; PhD Candidate; Faculty of Linguistics; Institute for Humanities and Cultural Studies;

Masood Ghayoomi; PhD in Computational Linguistics; Associate Professor; Faculty of Linguistics; Institute for Humanities and Cultural Studies; m.ghayoomi@ihcs.ac.ir

Atoosa Rostambeik Tafreshi; PhD in Linguistics; Associate Professor; Faculty of Linguistics; Institute for Humanities and Cultural Studies;

Introduction: Natural language is one of the most important tools of humans for communication, and conversation plays a fundamental role in knowledge transfer, the formation of social relationships, and the continuity of culture. In linguistics, conversation analysis and speech act theory show that understanding meaning in a dialogue depends on recognizing the speaker's communicative intent. With the advancement of artificial intelligence, efforts have been made to simulate this process in conversational systems such as chatbots. In this context, distributional semantics and word embedding models have made it possible to represent meaning in numerical format. Building on these approaches, the present study designs a Persian educational chatbot and compares several machine learning algorithms for detecting users' intents.

Methodology: To investigate the problem of intent detection in a Persian educational chatbot in the current study, a specialized corpus is first designed and compiled. Since no suitable benchmark dataset is available for Persian in the areas of intent detection and slot filling, a set of sentences is created based on the content analysis of textbooks used for teaching Persian to non-native speakers as well as consultation with experts in this field. The collected data consists of 3,607 sentences containing 153,054 words, which are organized into four semantic categories: synonymy, antonymy, collocation, and idiom. After data preparation, a preprocessing stage is carried out that included correcting coding issues of characters, spacing, and punctuation, as well as normalizing the texts. The IOB labeling scheme is then applied to identify the boundaries of semantic elements. In the next step, sentences are converted into 300-dimensional vectors using the static word-embedding models Word2Vec and FastText to be used by machine learning algorithms. To classify user intents, several algorithms are employed, including Support Vector Machines, Random Forest, Decision Trees, K-nearest Neighbors, and two neural network models, namely a single-layer perceptron and a multi-layer perceptron. In addition, a feed-forward neural network with two hidden layers and the ReLU

activation function are designed. Finally, the performance of the models is evaluated using five-fold cross-validation and the metrics of precision, recall, accuracy, and F1-score.

Main findings: The evaluation results show that the type of embedding method has a decisive impact on the performance of machine learning models in intent detection. In FastText-based embeddings, most algorithms demonstrated very strong performance due to the use of subword information. Among them, the multilayer perceptron achieves the best result among the models, reaching an accuracy of about 99.45%, indicating that neural networks have a high capability to capture nonlinear patterns and complex semantic relations in the vector space. In addition, the K-nearest Neighbors and Random Forest algorithms also show very high performance, with accuracies close to 99%. In contrast, when Word2Vec embeddings are used, a noticeable drop in accuracy is observed for some models, particularly for the Support Vector Machine and the neural network models. Error analysis further reveals that most misclassifications occur in distinguishing synonymy and antonymy, as these two relations share structural similarities in the vector space, making their boundaries difficult for some algorithms to separate them. Nevertheless, the confusion matrix results indicate that the multilayer perceptron combined with FastText is able to classify almost all samples correctly and achieved the best overall performance among the evaluated methods in terms of stability and generalization.

Discussion and conclusions: The findings of this study reveal several notable and relatively innovative points in the field of intent detection for Persian. First, the results show that the type of semantic representation plays a decisive role in the performance of intent detection models. In particular, the subword-based FastText embedding model is able to represent the lexical and morphological structures of Persian more effectively than Word2Vec, thereby significantly improves the performance of many algorithms. More importantly, combining this representation with the multilayer perceptron model lead to an accuracy of nearly 99.45%, indicating that nonlinear neural networks is capable of extracting subtle semantic patterns among linguistic intents with very high precision. This finding is directly related to the main research question concerning the effect of embedding type and model architecture on intent detection. It shows that the interaction between rich semantic representations and deep learning models can substantially improve the performance of Persian conversational systems. In addition, the error analysis shows that the greatest difficulty for the models lay in distinguishing between synonymy and antonymy, which reflects the complexity of these relations in semantic space of this language. Moreover, the development of a specialized corpus for intent detection in the domain of Persian language education is another important

contribution of this study, which can serve as a valuable infrastructure for future research on natural language understanding systems and Persian educational chatbots.

Keywords: (Intent detection, machine learning algorithm, task-oriented dialogue system, vectors, dialogue analysis)