NATURAL LANGUAGE PROCESSING AND ITS MAIN APPROACHES

Suyunova Mohinur
PhD student of National University of Uzbekistan.

Abstract: The article explores the relevance of NLP (or Computational Linguistics) to language learning, namely written and spoken communication. In addition, it further focuses on characterizing the techniques of NLP and uses of it for language learning. Together with presenting the background necessary to explain main principles of NLP, the paper also investigates the use of NLP in detecting language faults by students when they are learning the language as a second language.

Keywords: Statistical approaches, classical approaches, deep learning approaches, intelligent language tutoring systems, NLP algorithms, mal-rule technique, meta-rule

Аннотация: В статье исследуется значимость НЛП (или компьютерной лингвистики) для изучения языка, а именно письменного и устного общения. Кроме того, основное внимание уделяется характеристике техник НЛП и их использованию для изучения языка. Наряду с представлением исходной информации, необходимой для объяснения основных принципов НЛП, в статье также исследуется использование НЛП для выявления языковых ошибок у студентов, когда они изучают язык как второй язык.

Ключевые слова: статистические подходы, классические подходы, подходы глубокого обучения, интеллектуальные системы обучения языку, алгоритмы НЛП, техника неправильных правил, метаправило.

The automatic processing of human languages is known as natural language processing, or NLP (also known as computational linguistics). Since NLP is a broad, multidisciplinary field that is still relatively young, there are numerous definitions in use that are used by various practitioners. Any competent individual would include the following definition in their list:

In order to achieve human-like language processing for a variety of activities or applications, a theoretically motivated spectrum of computing techniques known as "natural language processing" is used to analyze and represent naturally occurring texts at one or more levels of linguistic analysis.

NLP, then, is a branch of computer science and linguistics that studies how computers and human (natural) languages interact. Furthermore, it is heavily influenced by developments in artificial intelligence (AI) and machine learning (ML). The natural language processing (NLP) approaches are designed to enable computers to comprehend and respond to commands in natural language. It should be mentioned that spoken and written language include the two categories of natural language processing.

Speech is the primary mode of human communication; hence written languages serve a secondary role in most activities. Written language is easier to understand than spoken language, which often contains noise and ambiguity. Natural language processing (NLP) is considered a challenging computer science topic due of its ambiguity [Abdurakhmonova, N. 2019,2021,2022,2023].

Natural language processing research dates back to the late 1940s. Machine translation (MT) was one of the first computer-based applications for natural language. According to Cambria and White (2014), NLP research has progressed from the time of punch cards and batch processing, when analyzing a sentence may take up to 7 minutes, to the present day, where millions of webpages can be analyzed in under a second.

The 'levels of language' technique is the most informative way to convey what actually happens within a Natural Language Processing system. People employ these levels to obtain meaning from written or spoken words. This is because language processing is mostly based on formal models or representations of knowledge relevant to various levels. Furthermore, language processing applications differ from data processing systems in that they make use of linguistic expertise. Natural language processing analysis is divided into the following levels: phonology, morphology, lexicon, syntactic, semantic, discourse, and pragmatics. The appendix contains the meanings of all levels.

There are several applications available that include both theory and implementations. Natural language processing can be applied to any text-based application. The appendix includes an overview of the most commonly used NLP applications. The majority of these challenges can be codified into five key tasks: classification, matching, translation, structured prediction, and sequential decision making.

According to Liddy (2001), natural language processing techniques can be classified into four categories: Symbolic approaches conduct in-depth analyses of linguistic phenomena and are based on the explicit description of language facts using well-understood knowledge representation schemes and algorithms. Statistical approaches use mathematical tools and vast text corpora to create approximate models of linguistic processes. These models are based on actual examples and do not require additional linguistic or world information.

Connectionist and statistical techniques both construct broad models of linguistic phenomena. Connectionism, also known as "parallel distributed processing," "neural networks," or "neuro-computing," differs from other statistical methods in that it combines statistical learning with representational theories, allowing for transformation, inference, and manipulation of logic formulae.

A rising number of researchers use a combination of data-driven and knowledge-driven methodologies, known as hybrid approaches. The following demonstrates that the approaches have both similarities and differences. For example, each technique has distinct assumptions, philosophical grounds, and sources of proof. There are two sorts of text classification methods: classical and deep learning.

Classical approaches rely on manual feature engineering and rules in combination with statistical algorithms. The manually designing features of data instances into feature vectors can be done in several ways. Studies have shown that the most effective surface features in hate speech detection are bag of words, word and character n-grams. In terms of classifiers, the most popular algorithm used, is the Support Vector Machine. Algorithms like Naive Bayes, Logistic Regression and Random Forest are also used for classification task [Abdurakhmonova, N. 2019,2021,2022,2023].

Deep learning approaches use neural networks to automatically learn several layers of features from input data. By the early 2000s, developments in computer hardware and optimization and training techniques have enabled the construction of ever larger and deeper networks, giving

rise to the modern term deep learning. Traditionally, NLP techniques rely on linear models trained on sparse feature vectors. However, non-linear neural networks with dense inputs have recently demonstrated success. The most commonly utilized networks are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), sometimes known as Long Short-Term Memory networks (LSTM).

Intelligent language tutoring systems (ILTSs) use natural language processing (NLP) to provide personalized feedback to learners through workbook-style exercises. Examples include the E-Tutor (Heift, 2010), Robo-Sensei (Nagata, 2009), TAGARELA (Amaral & Meurers, 2011), i-tutor (Choi, 2016), and FeedBook (Meurers et al., in press). NLP analysis can help optimize material sequencing and update learner models (Schulze, 2011). The analysis typically focuses on linguistic faults committed by the learner, although feedback can also highlight correctly employed forms, meaning, or appropriateness of a learner's response to an exercise.

What drives the adoption of NLP in tutoring systems? To provide feedback and track students' abilities in a learner model, an ILTS must first gather information about their abilities. The approach depends on the activities and learner reactions, including language production and system interaction. Language assessment (Bachman & Palmer, 1996) and task-based language teaching (Ellis, 2009) both emphasize the importance of the interaction between the activity and the learner's response. It is also crucial for determining the system analysis requirements of various activity types (Quixal & Meurers, 2016).

For many types of language learning activities, however, extensionally describing a direct and complete mapping between potential learner input and desired feedback is not possible. Nagata (2009, pp. 563-4) illustrates this with an exercise from her Japanese instructor ROBO-SENSEI. The learner reads a short communicative context and is instructed to write a sentence in Japanese based on the system's English translation. The learner response in this activity is directly dependent on the input provided by the exercise (a direct response in the framework of Bachman & Palmer, 1996), so a short, seven-word statement can be defined as the goal response.

After considering lexical, orthographic, and word order alternatives, there are 6,048 correct phrases that the learner can enter. Even limiting ill-formed patterns to erroneous particle and conjugation choices results in about a million phrases [Abdurakhmonova, N. 2019,2021,2022,2023]. Explicitly establishing a mapping between a million predicted replies and their related feedback is clearly impossible. Note that the explosion of viable learner answers depicted by Nagata is already a problem for direct response in a limited activity, where the meaning to be represented was fixed and only form variety was expected. Beyond analyzing language as a system (e.g., system-referenced tests in language assessment; Baker, 1989), performance-referenced analysis examines the ability to use language appropriately to complete a task.

To analyze learner input for various language activities that require significant variation in form, meaning, or task-appropriateness, NLP algorithms and resources should be used to abstract away from specific strings to more general properties. NLP analysis can provide feedback, update learner models, and sequence instruction based on a small set of language features and categories, rather than the enormous number of string instances.

Errors in native language texts have led to a shift from theory-driven, rule-based NLP to data-driven, statistical, and machine-learning approaches in the 1980s and 1990s. This shift is driven by the need to develop robust NLP that works in suboptimal conditions, such as noise and unknown forms and patterns. The purpose of using NLP in ILTS differs from other NLP domains.

NLP is designed to handle errors and unexpected input to produce results, such as syntactic analysis from a parser or translation from a machine translation system.

The basic purpose of NLP in ILTS is to provide feedback to learners by identifying language and response characteristics that differ from intended targets. The NLP abstraction aims to characterize faults, rather than ignoring them due to processing robustness [Abdurakhmonova, N. 2019,2021,2022,2023].

Writers' aids, such as standard spell and grammar checkers (Dickinson, 2006), share the ILTS focus on error detection, but they are based on assumptions about common errors committed by native speakers, which do not apply to language learners. Rimrott and Heift (2008) found that many L2 misspellings are multiple-edit errors and cannot be repaired by spell checkers meant for native writers. Tschichold (1999) also points out that typical writers' aids are not always useful for language learners because they require more scaffolding than a list of options from which to pick.

Language licensing systems rely on formal grammars of the language to be licensed, which can be articulated in one of two ways (Johnson, 1994). In a validity-based approach, recognizing a string involves identifying valid derivations from a collection of rules known as a grammar. Adding more rules to the language allows for licensing of more string kinds, whereas removing rules prevents licensing of any type. In a satisfiability-based setup, a grammar is a set of constraints, and a string is considered licensed if its model meets all of the constraints in the grammar.

There are two methods for examining a string to diagnose learner faults, similar to the two forms of formal grammars. The mal-rule technique takes the validity-based approach and using typical parsing algorithms. Starting with a conventional native language grammar, rules are added to license strings used by language learners but not in the native language. These rules are known as mal-rules and are used to license learner errors (Sleeman, 1982; Matthews, 1992). Meta-rules can capture generalizations across rules, such as errors in subject-verb agreement, which can occur in several rules involving subjects and finite verbs (Weischedel & Sondheimer, 1983). The mal-rule strategy is effective when errors correspond to the local tree of a single grammar rule. Otherwise, additional rules must be considered, making it substantially more difficult to detect errors and govern the interaction of mal-rules with ordinary rules. To restrict the search space created by rule interaction, the use of mal-rules might be limited. In basic cases, mal-rules are created after ordinary grammar analysis fails. However, this just limits the search space for well-formed strings. If parsing fails, it remains unclear which mal-rules should be added [Abdurakhmonova, N. 2019,2021,2022,2023].

Recent research on grammatical error correction (GEC) focuses on recognizing and repairing faults in written language as a translation problem, as opposed to statistical approaches that capture insights based on rules. The goal is to use statistical machine translation (Junczys-Dowmunt & Grundkiewicz, 2016) or neural network approaches (Chollampatt & Ng, 2018) to convert ill-formed text to well-formed text. However, these methods have limited relevance for research and applications that rely on linguistic rules and conceptualization.

Constraint relaxation, also known as the second group of language licensing approaches (Kwasny & Sondheimer, 1981), is an option for satisfiability-based grammars or rule-based grammars with complex categories that can be relaxed. Boyd (2012) suggests that a conflict detection system can diagnose learner errors when parsing is seen as a broad constraint satisfaction

problem. Constraint relaxation removes certain requirements from grammar, such as enforcing agreement. This allows for greater string licensing.

This presupposes that an error may be mapped to a specific constraint to be relaxed—that is, the domains of the learner error and the constraint in the grammar must be tightly related. Constraints can be associated with weights to control the likelihood of analysis (Foth, Menzel, & Schröder, 2005). This raises the question of how flexible control can be informed by the ranking of errors likely to occur for a specific learner given a task. Other proposals incorporate constraint relaxation and parts of mal-rules [Abdurakhmonova, N. 2019,2021,2022,2023].

Reuer (2003) combines a constraint relaxation technique with a modified parsing algorithm to license strings with inserted or omitted words. This approach incorporates meta-rule-like generalizations into the parsing process.

NLP contributes to language acquisition by exposing learners to actual native language and providing possibilities for interaction. This area of research involves work on finding and improving authentic texts for learners to read, as well as the automatic development of exercises and examinations based on such materials.

What are the specific learner language properties that need to be identified? Historically, readability has been the primary factor in text selection, with several algorithms devised for this purpose (DuBay, 2004). Current machine learning methods, informed by a broader range of linguistic characteristics, are significantly more accurate than traditional measures based on shallow features such as average sentence and word lengths (Xia, Kochmar, & Briscoe, 2016; Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017; Weiss & Meurers, 2018), including some commercial systems (Nelson, Perfetti, Liben, & Liben, 2012).

Beyond readability, SLA research has revealed that awareness of language categories and forms is an important component of successful second language acquisition (see Lightbown & Spada, 1999), and a variety of linguistic properties, including morphological, syntactic, semantic, and pragmatic information, have been identified as relevant for language awareness (see Schmidt, 1995, p. 30). In answer to this demand, the FLAIR system (Chinkina & Meurers, 2016) provides linguistically aware Web searches, allowing language learners' input to be systematically enriched with the types of language patterns to be learned next.

In conclusion, although NLP has had limited impact on real-life language teaching and SLA, it offers numerous opportunities for developing applications to support language learning and research. Interdisciplinary collaboration between SLA and NLP is critical for establishing accurate annotation systems and analysis algorithms that uncover key properties for analyzing learner language.

References:

1. Chinkina, M., & Meurers, D. Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11thWorkshop on Innovative Use of NLP for Building Educational Applications (BEA)* (pp. 188–98). San Diego, CA: Association for Computational Linguistics. Retrieved April 4, 2019 from https://www.semanticscholar.org/paper/Linguistically-Aware-Information-Retrieval%3A-Input-Chinkina-Meurers/d1e6e66b181b5912c101a19d345e56a0c5c28bba. ,2016.

2. Chollampatt, S., & Ng, H. T. A multilayer convolutional encoder–decoder neural network for grammatical error correction. In *Thirty-Second AAAI Conference on Arti!cial Intelligence (AAAI)* (pp. 5755–62). New Orleans, LA: AAAI, 2018.
3. DuBay, W. H. *The principles of readability*. CostaMesa, CA: Impact Information. Retrieved April 4, 2019 from http://www.impact-information.com/impactinfo/readability02.pdf, 2004.
4. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
5. Абдурахмонова, Н., & Бойсариева, С. (2023). TABIIY TILNI QAYTA ISHLASHDA (NLP) OKKAZIONALIZMLARNING MORFEM TAHLILI. *МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА*, *6*(3).
6. Abdurakhmonova, N. Z., Ismailov, A. S., & Mengliev, D. (2022, November). Developing NLP Tool for Linguistic Analysis of Turkic Languages. In *2022 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)* (pp. 1790-1793). IEEE.
7. Mengliev, D. B., Abdurakhmonova, N., Hayitbayeva, D., & Barakhnin, V. B. (2023, November). Automating the transition from dialectal to literary forms in Uzbek language texts: an algorithmic perspective. In *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)* (pp. 1440-1443). IEEE.
8. San'atbek, M., Mirsaid, A., & Nilufar, A. (2018). Modeling WordNet type thesaurus for Uzbek language semantic dictionary. *International Journal of Systems Engineering*, *2*(1), 26.
9. Foth, K., Menzel, W., & Schröder. Robust parsing with weighted constraints. *Natural Language Engineering*, *11*(1), 1–25, 2005.
10. Johnson,M. Two ways of formalizing grammars. *Linguistics and Philosophy*, *17*(3), 221–48, 1994.
11. Lightbown, P. M., & Spada, N. *How languages are learned*. Oxford, England: Oxford University Press, 1999.
12. Reuer,V. Error recognition and feedback with lexical functional grammar. *CALICO Journal*, *20*(3), 497–512, 2003.
13. Schmidt, R. Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu: University of Hawai'i Press, 1995.