

Zpracování přirozeného jazyka (NLP)

Zpracování přirozeného jazyka (Natural Language Processing - NLP) je klíčová a extrémně rychle se rozvíjející disciplína na pomezí informatiky, umělé inteligence (AI) a počítačové lingvistiky. Jejím cílem je naučit počítače číst, analyzovat, chápat a generovat lidský jazyk (ať už psaný text, nebo mluvené slovo) způsobem, který je smysluplný a užitečný.

Zatímco tradiční programování vyžaduje striktní a jednoznačnou syntaxi (jako C++ nebo Python), lidský jazyk je plný nejednoznačností, dvojsmyslů, sarkasmu, idiomů a gramatických výjimek. NLP je mostem, který převádí tento „chaotický“ lidský projev do strukturovaných dat, kterým dokáže algoritmus porozumět.

Dvě hlavní poddisciplíny NLP

Proces komunikace se dělí na porozumění a tvorbu. Stejně tak se NLP historicky dělí na dvě hlavní větve:

- **NLU (Natural Language Understanding - Porozumění jazyku):** Zaměřuje se na čtení s porozuměním. Počítač analyzuje text a extrahuje z něj význam, záměr uživatele (intent) a entity. Zjišťuje odpověď na otázku: „*Co se mi tento text snaží říct?*“
- **NLG (Natural Language Generation - Generování jazyka):** Zaměřuje se na tvorbu textu. Na základě strukturovaných dat nebo předchozího kontextu počítač formuluje plynulé, gramaticky správné a logické věty. Názorným příkladem NLG je odpověď od ChatGPT.

Základní techniky a kroky zpracování (NLP Pipeline)

Než může starší analytický model nebo vyhledávač z textu něco vyvodit, musí surový text projít sekvencí úprav, takzvanou **NLP Pipeline**. Zde jsou nejčastější kroky:

- **Tokenizace (Tokenization):** Rozsekání dlouhého textu na nejmenší smysluplné jednotky – takzvané tokeny. Obvykle to znamená rozdělení věty na jednotlivá slova a interpunkční znaménka.
- **Odstranění stop-slov (Stop words removal):** Odfiltrování slov, která nemají téměř žádný sémantický význam a vyskytují se v textu příliš často (např. spojky a předložky jako „a“, „ale“, „se“, „v“, „na“). Tím se ušetří výpočetní výkon.
- **Lematizace a Stemming:** Převod slov na jejich základní tvar. *Stemming* hrubě oseká předpony a přípony (např. „rybářům“ → „ryb“), zatímco sofistikovanější *Lematizace* bere v potaz slovník a převede slovo na jeho kořen (např. „byl“ → „být“).
- **Určování slovních druhů (POS Tagging):** Algoritmus u každého tokenu určí, zda se jedná o podstatné jméno, sloveso, přídavné jméno atd. To je klíčové pro pochopení kontextu věty.
- **Rozpoznávání jmenných entit (NER - Named Entity Recognition):** Identifikace a kategorizace klíčových prvků v textu. Systém najde v textu jména osob, názvy firem, lokace, data nebo finanční částky (např. „Brno“ → [LOKACE], „Apple“ → [ORGANIZACE]).

Evoluce přístupů k NLP

Technologie NLP prošla od svých počátků obrovskou revolucí. Změnil se způsob, jakým počítače jazyk chápou.

1. Pravidlové systémy (50. - 80. léta)

Systémy založené na tvrdě nakódovaných gramatických pravidlech (if-then-else) a rozsáhlých slovnících, které tvořili lingvisté.

- *Příklad:* První chatbot ELIZA (1966), který fungoval jen na hledání klíčových slov a jejich mechanickém obracení do otázek.
- *Nevýhoda:* Lidský jazyk je příliš složitý, výjimky nelze pokrýt pravidly.

2. Statistické NLP (90. léta - 2010)

Nástup strojového učení (Machine Learning). Namísto psaní pravidel lingvisty se do počítače nahrály obrovské databáze textů (korpora). Algoritmy (např. Naivní Bayes, Skryté Markovovy modely) počítaly pravděpodobnost, s jakou se slova vyskytují vedle sebe. Tímto způsobem fungovaly první úspěšné spam filtry.

3. Hluboké učení a Word Embeddings (2010 - 2017)

Počítače začaly využívat neuronové sítě a reprezentovat slova jako složité vektory v mnohorozměrném matematickém prostoru (technologie Word2Vec). Počítač sice nechápe význam slova „Král“, ale ví, že z matematického hlediska k němu má blízko „Muž“ a „Trůn“. Umožnilo to řešit matematické rovnice typu: $Král - Muž + Žena = Královna$.

4. Transformery a Velké jazykové modely (LLM) (2017 - současnost)

Zlomový okamžik přišel v roce 2017 s architekturou **Transformer** (vytvořenou společností Google). Tento algoritmus zavedl mechanismus „**Self-Attention**“ (**Pozornost**). Prohlíží celou větu najednou (nikoliv slovo po slovu) a určuje, jaké slovo souvisí s jakýmkoli jiným slovem v textu, i když jsou daleko od sebe. Na této technologii stojí moderní giganti jako rodina modelů GPT (OpenAI) nebo BERT (Google).

Praktické využití NLP

S prvky NLP se dnes setkáváme prakticky neustále:

- **Strojový překlad (Machine Translation):** Nástroje jako Google Translate nebo DeepL, které plynule překládají celé věty při zachování kontextu.
- **Analýza sentimentu:** Firmy automaticky skenují miliony tweetů a uživatelských recenzí, aby

zjistily, zda se o jejich nové službě mluví pozitivně, negativně, nebo neutrálně.

- **Extrakce informací a sumarizace:** Lékařský software, který přečte 500stránkový chorobopis a vytáhne z něj do úhledné tabulky pouze podané léky a naměřený krevní tlak, nebo vygeneruje stručné jednostránkové shrnutí.
- **Hlasoví asistenti:** Siri, Alexa a Google Assistant využívají NLU pro pochopení mluveného příkazu a NLG pro vygenerování hlasové odpovědi.

Srovnání NLP přístupů

Metoda	Doba největší slávy	Jak se systém učí?	Největší slabina	Typické nasazení
Pravidlová	80. léta	Lingvista manuálně píše regex a gramatická IF/THEN pravidla	Extrémně nepružné, nezvládá slang a překlady	Extrakce formátovaných dat (např. rodných čísel)
Statistická	Nulové roky	Počítá četnost a pravděpodobnost slov ve velkém textu (TF-IDF)	Chybí hlubší chápání kontextu a synonym	Detekce SPAMu v e-mailu
Transformery (LLM)	2017 a dále	Pozornost (Self-Attention) a trénink na téměř celém internetu	Hardwarově extrémně náročné, občasné „halucinace“ (vymyšlení si)	ChatGPT, DeepL, Copilot

From:

<https://www.serviceit.cz/> - IT ENCYKLOPEDIE

Permanent link:

https://www.serviceit.cz/doku.php?id=zpracovani_jazyka

Last update: **2026/06/06 12:05**

