



Tomáš Mikolov



„Když jsem jako student v roce 2007 generoval text z neuronových jazykových modelů, věděl jsem, že pozoruji něco, co přede mnou nikdo neviděl. Text byl mnohem plynulejší než z n-gramových modelů a věděl jsem, že tohle je budoucnost.” -

Tomáš Mikolov

1 Úvod: Revoluce ve zpracování přirozeného jazyka

Tomáš Mikolov patří mezi nejvlivnější postavy v historii umělé inteligence 21. století. Jeho algoritmus Word2Vec, publikovaný v roce 2013, fundamentálně změnil způsob, jakým počítače reprezentují a rozumí lidskému jazyku. S více než 194 000 citacemi na Google Scholar se řadí mezi nejcitovanější vědce v oblasti strojového učení a jeho práce položila základy pro všechny současné velké jazykové modely včetně GPT, BERT a Claude.

1.1 Stav zpracování jazyka před rokem 2013

Před Mikolovým průlomem bylo zpracování přirozeného jazyka (NLP) založeno na statistických metodách, které sémantickou podobnost slov ignorovaly. Slova byla reprezentována pomocí *one-hot encoding* – každé slovo jako vektor s jednou jedničkou v potenciálně 100 000-dimenzionálním prostoru. Metody jako TF-IDF a Bag-of-Words neumožňovaly modelovat vztahy mezi slovy typu „král” a „královna” nebo „Paříž” a „Francie”.

Geoffrey Hinton již v roce 1986 představil koncept distribuovaných reprezentací v knize *Parallel Distributed Processing*, ale praktické využití přišlo až o dvě dekády později. Yoshua Bengio v roce 2003 publikoval průkopnickou práci *A Neural Probabilistic Language Model*, která jako první ukázala, že neuronové sítě mohou učit hustá slovní vnoření (word embeddings) zachycující sémantiku. Bengioův model však byl výpočetně velmi náročný – obsahoval tři vrstvy včetně komplexní softmax vrstvy přes celý vokabulář.

Ronan Collobert pak v roce 2008 ukázal v práci *Natural Language Processing (Almost) from Scratch*, že předtrénovaná slovní vnoření zlepšují výkonnost na downstream úlohách jako rozpoznávání pojmenovaných entit nebo syntaktická analýza. Přesto však žádný z těchto přístupů nebyl dostatečně škálovatelný pro masivní nasazení v praxi.

2 Životní dráha českého průkopníka AI

Tomáš Mikolov se narodil v roce 1984 v Šumperku. Na Fakultě informačních technologií VUT v Brně získal v roce 2007 titul Ing. s diplomovou prací zaměřenou na jazykové modelování pro rozpoznávání řeči v češtině. Již během

magisterského studia experimentoval s neuronovými jazykovými modely pro generování textu – tehdy naprostá novinka v oboru.

V letech 2007–2012 pokračoval v doktorském studiu na VUT pod vedením profesora Jana Černockého. Svou disertační práci *Statistical Language Models Based on Neural Networks* obhájil 2. října 2012. Během doktorátu hostoval na Johns Hopkins University, Université de Montréal (kde spolupracoval s Yoshuou Bengiem) a Microsoft Research.

Klíčový průlom přišel v období 2012–2014, kdy Mikolov pracoval jako Research Scientist v Google Brain. Zde vyvinul své nejslavnější dílo – Word2Vec. V roce 2014 přestoupil do Facebook AI Research (FAIR), kde vedl vývoj systému fastText. Po šesti letech v Silicon Valley se v roce 2020 vrátil do České republiky, kde působil jako Senior Research Scientist v CIIRC ČVUT Praha. V roce 2025 spoluzaložil startup BottleCap AI zaměřený na efektivní velké jazykové modely.

Za svou práci získal řadu prestižních ocenění včetně Ceny Neuron (2018) za „největší zlepšení v oblasti zpracování přirozeného jazyka za poslední desetiletí“, NeurIPS Test of Time Award (2023) za paper o Word2Vec a Stříbrnou medaili předsedy Senátu ČR (2024).

3 Technické přínosy: Od RNNLM k Word2Vec

3.1 Recurrent Neural Network Language Model (2010)

Mikolovým prvním průlomem byla práce *Recurrent neural network based language model*, publikovaná na konferenci INTERSPEECH 2010. Tato práce ukázala, že rekurentní neuronové sítě dokáží modelovat jazykovou strukturu podstatně lépe než dosavadní n-gramové modely. RNNLM toolkit dosáhl 50% snížení perplexity oproti state-of-the-art backoff modelům a 18% snížení word error rate na Wall Street Journal datasetu.

Jak sám Mikolov později uvedl: „*Navzdory tomu, že word2vec je můj nejcitovanější paper, nikdy jsem si nemyslel, že je to můj nejvlivnější projekt. RNNLM byl podle mě přinejmenším stejně revoluční jako AlexNet.*” RNNLM toolkit byl prvním nástrojem umožňujícím praktický trénink neuronových jazykových modelů na velkých korpusech.

3.2 Word2Vec: Democratizace word embeddings

V lednu 2013 Mikolov publikoval paper *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781), který představil algoritmus Word2Vec. V říjnu téhož roku následoval druhý paper *Distributed Representations of Words and Phrases and their Compositionality*, publikovaný na prestižní konferenci NeurIPS. Tyto dva články patří s více než 33 000 a 34 000 citacemi k nejcitovanějším pracím v historii strojového učení.

Word2Vec představil dvě architektury:

- **Continuous Bag-of-Words (CBOW)**: predikuje cílové slovo z kontextových slov v okolí
- **Skip-gram**: predikuje kontextová slova z cílového slova

Klíčovou inovací bylo dramatické zjednodušení architektury. Mikolov odstranil skrytou vrstvu z Bengiova modelu a nahradil výpočetně náročný softmax tzv. *negative sampling* technikou – místo klasifikace přes celý vokabulář se model učí binární klasifikaci na několika negativních vzorcích. Tento trik umožnil trénink na miliardách slov během hodin místo týdnů.

Word2Vec ukázal fascinující vlastnost – vektory slov zachycují lingvistické analogie. Slavný příklad: $\vec{král} - \vec{muž} + \vec{žena} \approx \vec{královna}$. Podobně fungují vztahy jako „Paříž : Francie = Berlín : Německo” nebo „velký : větší = malý : menší”.

Zajímavostí je, že původní Word2Vec paper byl zamítnut na konferenci ICLR 2013. Kód byl open-sourcován až v srpnu 2013 po interních diskusích v Google. Tento open-source release způsobil obrovskou vlnu adopce v akademické i průmyslové sféře.

3.3 fastText: Vnoření s podstrukturou

V roce 2016–2017 Mikolov spolu s týmem ve Facebook AI Research vyvinul fastText, publikovaný v článku *Enriching Word Vectors with Subword Information*. fastText rozšířil Word2Vec o reprezentaci slov pomocí character n-gramů (úseků 3–6 znaků). Tato inovace umožnila generování vektorů pro neznámá slova (out-of-vocabulary) a lepší zachycení morfologie, což je klíčové pro flektivní jazyky jako čeština.

fastText také dramaticky zrychlil trénink – model bylo možné natrénovat na více než miliardě slov za méně než 10 minut na běžném CPU. Facebook

AI Research vydal předtrénované modely pro 157 (později 294) jazyků, což učinilo kvalitní word embeddings dostupnými pro prakticky jakýkoliv jazyk.

4 Dopad na vývoj moderních jazykových modelů

4.1 Od statických k kontextualizovaným vnořením

Word2Vec představoval statická vnoření – každé slovo mělo jedno vektorové vnoření bez ohledu na kontext. V roce 2018 přišel model ELMo (Embeddings from Language Models), který zavedl kontextualizovaná vnoření závislá na okolí slova. Přesto však moderní transformerové modely jako BERT (2019) a GPT používají word embeddings jako svou první vrstvu – koncepčně jde o přímé pokračování Mikolovovy práce.

Jak poznamenal Sebastian Ruder, expert na NLP: „*Word2Vec byl katalyzátorem pro 'neural revolution' v NLP. Ukázal, že jednoduché neuronové architektury trénované na masivním množství textu dokáží automaticky objevovat lingvistickou strukturu. Tento insight nastavil scénu pro vše, co následovalo.*”

4.2 Průmyslové aplikace

Technika embeddingu inspirovaná Word2Vec se rozšířila daleko za hranice NLP:

- **Airbnb** používá embedding listingů pro personalizované vyhledávání s nárůstem CTR o 21% a 99% konverzí
- **Yahoo** aplikoval embedding reklam s nárůstem CTR o 9%
- **Spotify** doporučuje skladby pomocí embedding vektorů – více než 700 milionů skladeb měsíčně
- **Alibaba** používá embeddingy pro doporučování produktů
- Časopis **Nature** publikoval v roce 2019 studii o využití embeddingů pro analýzu 3,3 milionu vědeckých abstraktů k objevování nových materiálů

Koncepty byly také adaptovány mimo text:

- **Gene2Vec**: analýza biologických sekvencí
- **Node2Vec**: vnoření grafových struktur pro sociální sítě
- **Image2Vec**: reprezentace obrázků pro vyhledávání

4.3 Citační vliv

Podle Google Scholar má Mikolov celkem přes 194 000 citací s h-indexem 49 a i10-indexem 85. Jeho tři nejcitovanější práce jsou právě papery o Word2Vec a RNNLM. V roce 2023 obdržel NeurIPS Test of Time Award za paper „Distributed Representations of Words and Phrases“, což je prestižní ocenění udělované pracím s trvalým dopadem po 10 letech od publikace.

5 Závěr: Místo v historii AI

Tomáš Mikolov patří do úzké skupiny vědců, jejichž práce fundamentálně změnila celý obor. Jeho největším přínosem nebyla specifická technika, ale **demonstrate principu**: jednoduché neuronové architektury trénované na masivním množství dat dokáží automaticky objevovat komplexní strukturu jazyka.

Word2Vec demokratizoval přístup ke kvalitním slovním reprezentacím. Před rokem 2013 bylo získání dobrých word embeddings náročné – vyžadovalo to expertní znalosti, výpočetní zdroje a dlouhý čas. Po open-source release Word2Vec v srpnu 2013 mohl kdokoli stáhnout předtrénovaný model a použít ho během minut.

Mikolov také jako jeden z prvních ukázal paradigma *pretrain-then-finetune*, které je dnes základem všech velkých jazykových modelů. Myšlenka, že předtrénování na obrovském množství neoznačených dat vytvoří užitečné reprezentace použitelné pro specifické úlohy, je dnes klíčová pro GPT-4, Claude, BERT a další systémy.

Z historické perspektivy můžeme sledovat jasnou linii: Hinton (1986) položil teoretické základy distribuovaných reprezentací → Bengio (2003) ukázal, že neuronové sítě dokáží učit word embeddings → Collobert (2008) demonstroval transfer learning → Mikolov (2013) to udělal **praktickým a škálovatelným**.

Jak sám Mikolov nedávno uvedl: „*Co když ve skutečnosti nedosažení umělé inteligence je největší existenční hrozbou pro lidstvo?*” Tento filozofický postoj odráží jeho přesvědčení o transformační síle AI. Jeho práce na Word2Vec a neuronových jazykových modelech byla jedním z klíčových kroků na cestě k dnešním velkým jazykovým modelům, které začínají prokazovat překvapivě obecné schopnosti v porozumění a generování lidského jazyka.

Reference

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed Representations of Words and Phrases and their Compositionality*. NeurIPS 2013, pp. 3111-3119
3. Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). *Recurrent neural network based language model*. INTERSPEECH 2010
4. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. TACL, Vol. 5, pp. 135-146
5. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). *A Neural Probabilistic Language Model*. JMLR, 3, pp. 1137-1155
6. Collobert, R., & Weston, J. (2008). *A Unified Architecture for Natural Language Processing*. ICML 2008
7. VUT Brno. *Statistical Language Models Based on Neural Networks*. PhD Thesis Database. <https://www.fit.vut.cz/study/phd-thesis/283/.en>
8. IEEE Signal Processing Society (2021). *Industry Leaders in Signal Processing and Machine Learning: Tomas Mikolov*. <https://signalprocessingsociety.org/newsletter>
9. Google Open Source Blog (2013). *Learning the meaning behind words*. <https://opensource.googleblog.com/2013/08/>
10. ELLIS Unit Prague (2023). *The NeurIPS Test of Time Award for Tomáš Mikolov*. <https://ellis.ciirc.cvut.cz/>
11. Nadace Neuron. *Tomáš Mikolov - Cena Neuron*. <https://www.nadaceneuron.cz/en/person/tom-mikolov>
12. Guthrie, D. et al. (2018). *Real-time Personalization using Embeddings for Search Ranking at Airbnb*. KDD 2018
13. Peters, M. et al. (2018). *Deep contextualized word representations*. NAACL 2018

14. Devlin, J. et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL 2019
15. Hinton, G.E. (1986). *Distributed Representations in Parallel Distributed Processing*. MIT Press
16. Ruder, S. *On word embeddings - Part 1*. <https://www.ruder.io/word-embeddings-1/>
17. Rare Technologies Podcast. *Tomas Mikolov on Word2Vec and AI Research*. <https://rare-technologies.com/rrp-1-tomas-mikolov/>
18. Towards Data Science. *Uncovering the Pioneering Journey of Word2Vec - Interview with Tomas Mikolov*. <https://towardsdatascience.com/uncovering-the-pioneering-journey-of-word2vec/>