Daten oder keine Daten, das ist die Frage: lässt sich MÜ ohne parallele Daten trainieren?

Prof. Dr. Josef van Genabith
josef.van_genabith@dfki.de

Overview

- MT wird ziemlich gut - human parity?
- Überwachte ML/MT: parallele Daten
- Unüberwachte MT: mono-linguale Daten
- Selbstüberwachte MT: Vergleichbare Daten
- Grenzen & Fazit

We simplify and cut corners …

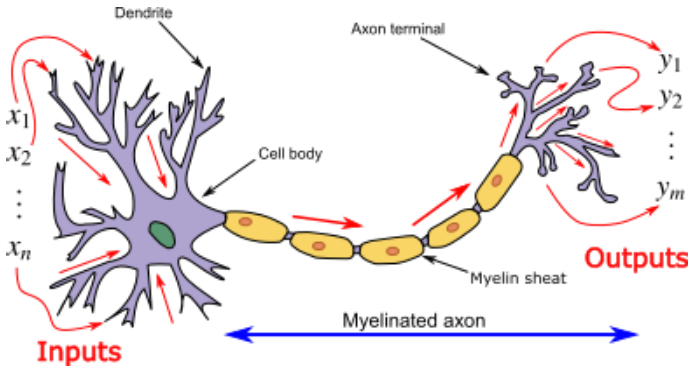
MT und Human Parity?

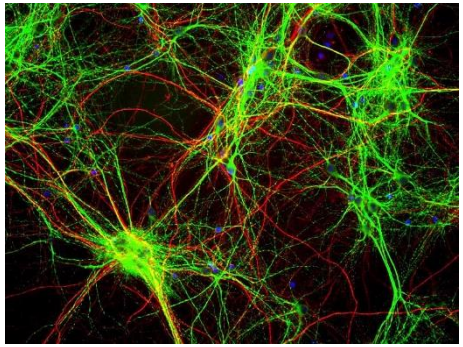
English→German

Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	HUMAN
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
86.6	0.094	Microsoft-WMT19-sent-level
87.3	0.081	JHU
84.4	0.077	Helsinki-NLP
84.2	0.038	online-Y
83.7	0.010	lmu-ctx-tf-single
84.1	0.001	PROMT-NMT
82.8	−0.072	online-A
82.7	−0.119	online-G
80.3	−0.129	UdS-DFKI
82.4	−0.132	TartuNLP-c
76.3	−0.400	online-X
43.3	−1.769	en-de-task

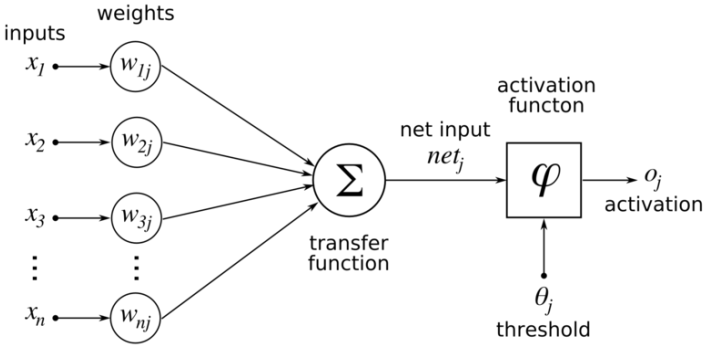
WMT 2019, Florence, Italy

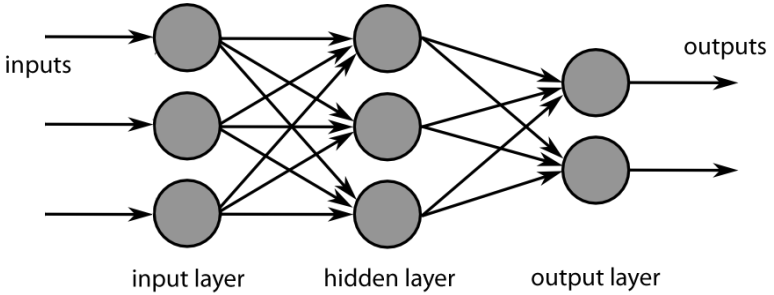
Warum ist die MT so gut: Künstliche Neuronale Netzwerke





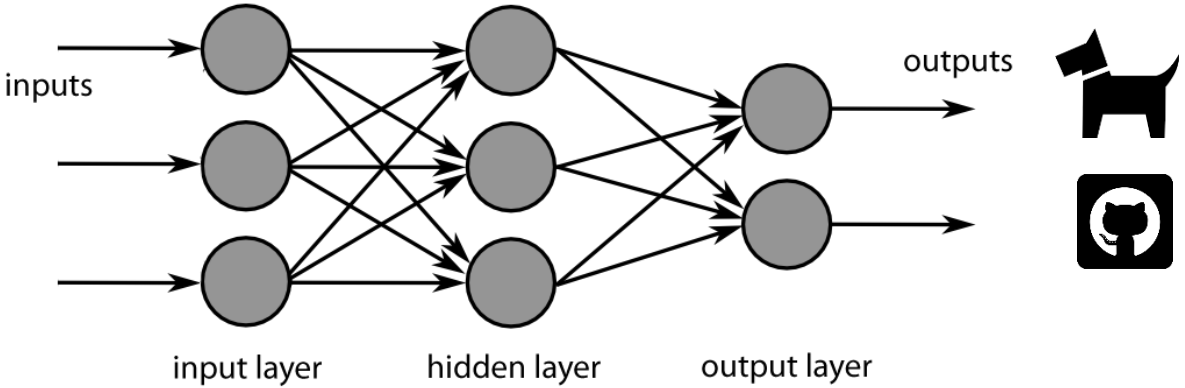
Sources:
Wikimedia





Neuronale Netzwerke & Überwachtes Lernen

$$\begin{bmatrix} 7 \\ 22 \\ 4 \\ 112 \\ 34 \\ \vdots \\ 8 \end{bmatrix}$$



Sources:
Wikimedia
pixy.org

Neuronale Netzwerke & Überwachtes Lernen

$$\begin{bmatrix} 7 \\ 22 \\ 4 \\ 112 \\ 34 \\ \vdots \\ 8 \end{bmatrix}$$

inputs

Back-propagation

Error!

outputs

0.3 1.0

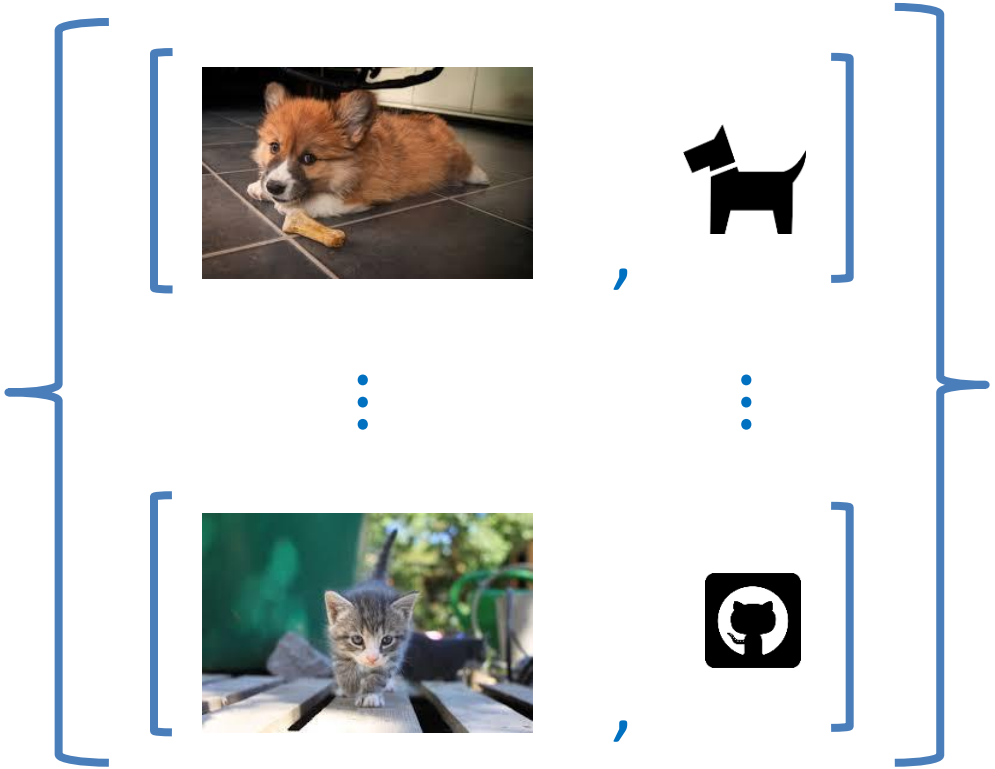
0.7 0.0

input layer hidden layer output layer

Forward pass

Sources:
Wikimedia
pixy.org

Neuronale Netzwerke & Überwachtes Lernen



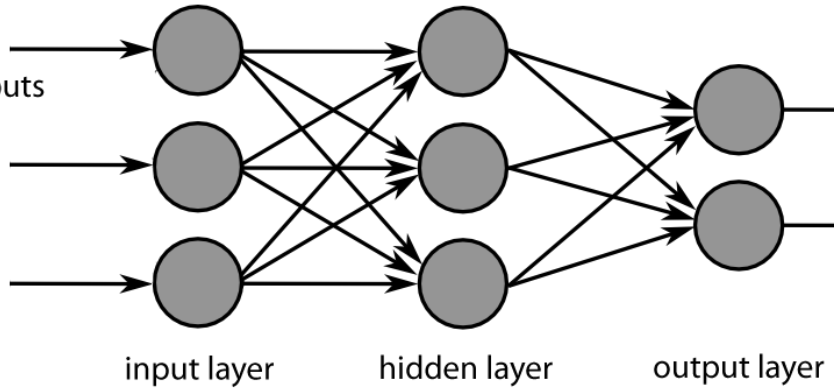
Supervised ML
Labelled training data
[data , label]
Label = supervision signal

Neuronale Netzwerke & Überwachtes Lernen

Is $[5, 2 \ldots.]$

everything $[8, 11 \ldots.]$

OK $[33, 1, \ldots.]$

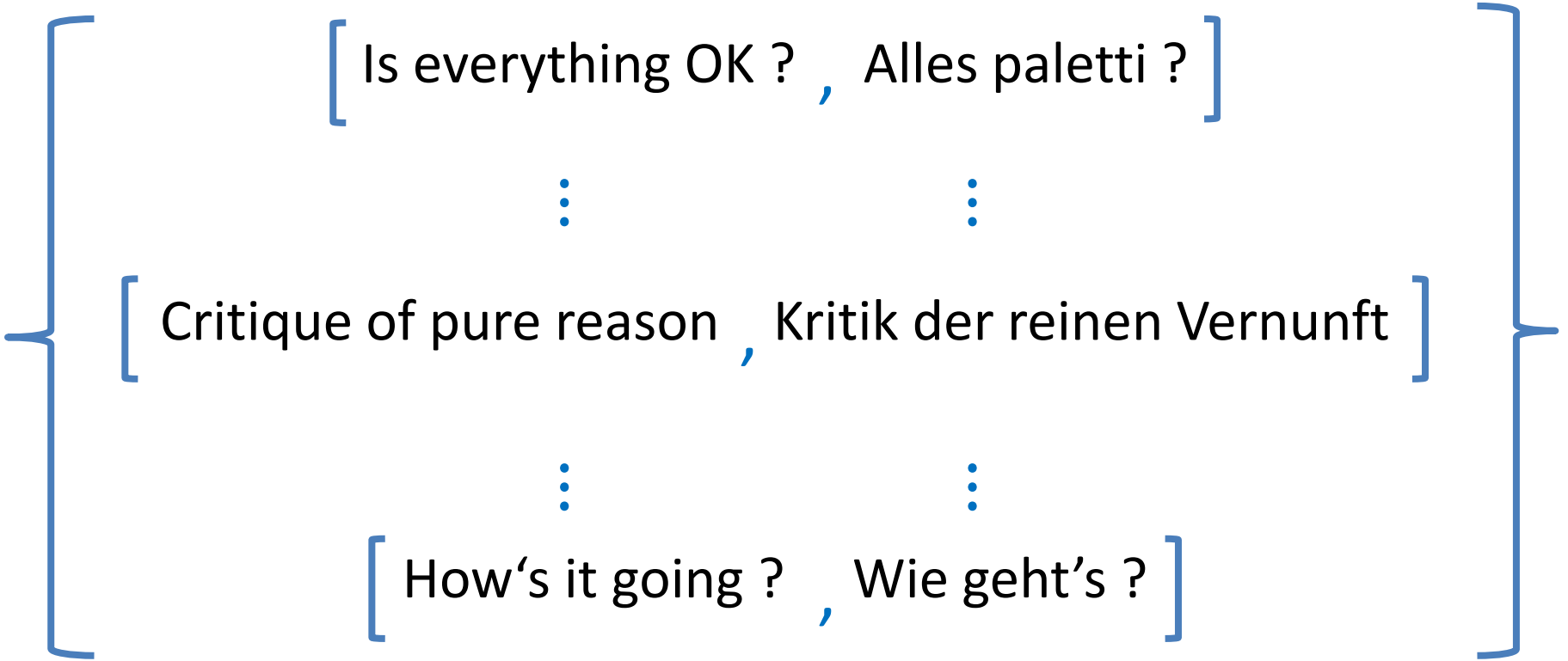


$[2, \ldots.]$ Alles

$[7, \ldots.]$ paletti

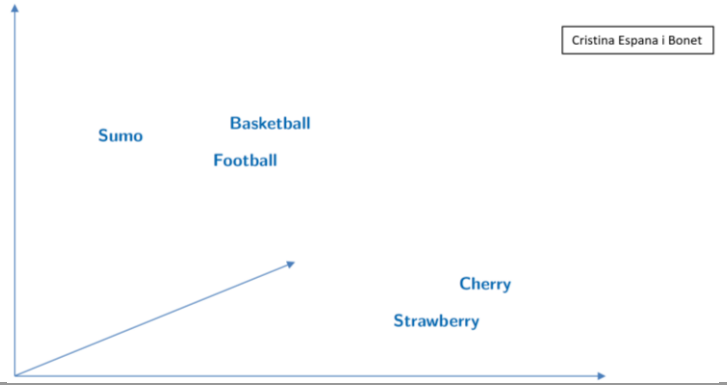
Sources:
Wikimedia

Neuronale Netzwerke & Überwachtes Lernen



Word Embeddings

- Wie bekommen wir Wörter in ein NN?
- Wörter als Zahlen/Vektoren: $[33, 1, \ldots, \ldots]$
- Ein Vektor = Punkt in mehrdimensionalem Raum
- Mit Wörtern rechnen: king - male + female ≈ queen
- Feingranulare Ähnlichkeit
- Sofa, Couch, Sessel
- Algebra ⇔ Zebra



Das Problem:

- ~ 7000 Sprachen
- $n \times (n-1)$ directe MT Systeme
- $48{,}993{,}000$ directe MT Systeme
- $(n-1)+(n-1)$ MT Systeme mit einem Pivot
- $13{,}998$ MT Systeme mit einem Pivot
- Wir haben parallele Trainingsdaten für < 100 Sprachpaare
- Wie können wir den restlichen /meisten! Sprachen helfen???

Source: pixabay

Unüberwachte (N)MT

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre.
- Guillaume Lample, Alexis Conneau, Myle Ott, Ludovic Denoyer, and Marc’Aurelio Ranzato.

Unüberwachte MT

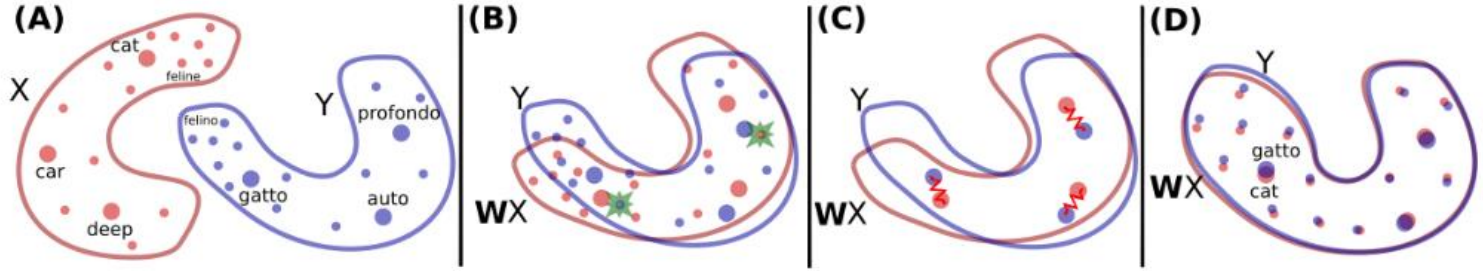
- Für die meisten Sprachen haben wir keine bilingualen Daten …
- Aber: jede Menge von mono-lingualen Daten in der Ausgangs- und Zielsprache

- => Wie können wir Übersetzung aus mono-lingualen Daten lernen???

Unüberwachte MT: wie funktioniert das?

- Berechne word embeddings für die mono-lingualen Daten
- Aligniere sie (Conneau et al. 2018)

https://github.com/flairNLP/flair/issues/852



- Wortübersetzungen => wortbasierte SMT:

cat : gatto
car : auto
deep: ….

Unüberwachte MT: wie funktioniert das?

One way of doing this (super-simplified):

- Train strong E and F Language Models (LMs) on mono-lingual data
- Use the word-based translation dictionary and the LMs in two word-based unsupervised SMT systems: E>F and F>E
- Use E>F to (back-) translate some E to F: E>F(E) = F‘ creating some dodgy F‘: synthetic F‘ to E
- Use F>E to (back-) translate some F to E: F>E(F) = E‘ creating some dodgy E‘: synthetic E‘ to F
- Train better “supervised“ PB-SMT/NMT on F‘ to E and E‘ to F
- Loop …

Unüberwachte MT: Resultate für WMT 14

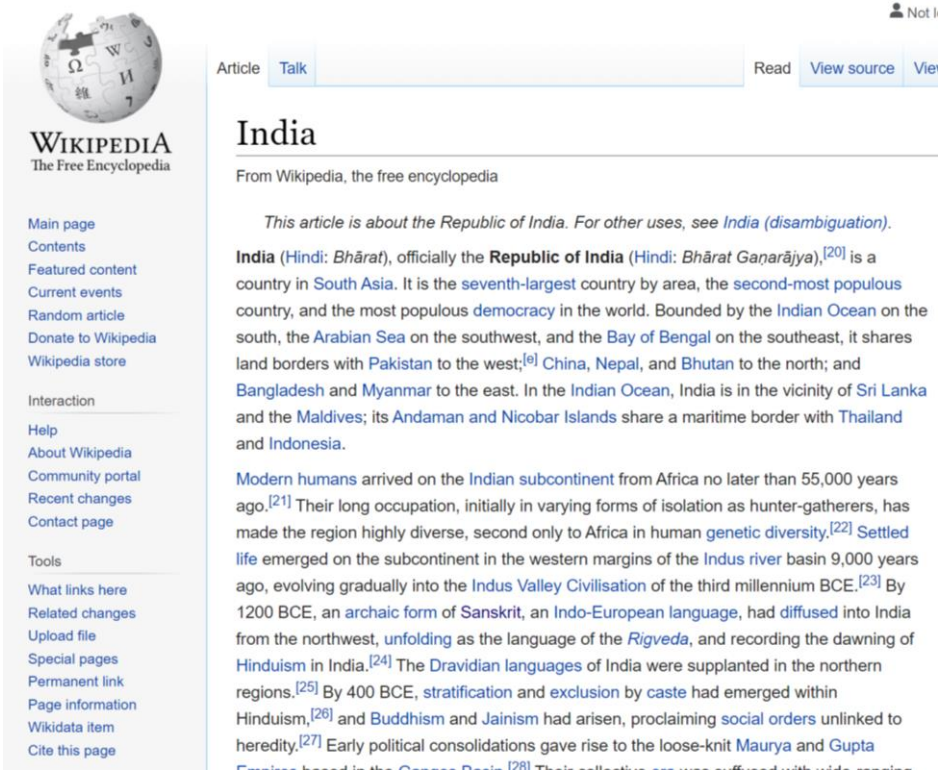
		FR-EN	EN-FR	DE-EN	EN-DE
Unsup.	Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
	+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
	Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
	Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5
	detok. SacreBLEU	33.2	33.6	26.4	21.2
Supervised	WMT winner	35.0	35.8	29.0	20.6
	Original transformer (Vaswani et al., NIPS'17)*	-	41.0	-	28.4
	Large scale back-translation (Edunov et al., EMNLP'18)	-	43.8	-	33.8

From: Mikel Artexte, Unsupervised Machine Translation, slides Recent Advances in Machine Translation RAMT2021, NITS, Assam, India

Selbst-überwachte NMT

- Dana Ruiter, Cristina Espana Bonet and Josef van Genabith.

Wikipedia: vergleichbare Daten ≠ Übersetzungen



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Not l

Article | Talk | Read | View source | Vie

India

From Wikipedia, the free encyclopedia

This article is about the Republic of India. For other uses, see India (disambiguation).

India (Hindi: *Bhārat*), officially the **Republic of India** (Hindi: *Bhārat Gaṇarājya*),[20] is a country in South Asia. It is the seventh-largest country by area, the second-most populous country, and the most populous democracy in the world. Bounded by the Indian Ocean on the south, the Arabian Sea on the southwest, and the Bay of Bengal on the southeast, it shares land borders with Pakistan to the west;[e] China, Nepal, and Bhutan to the north; and Bangladesh and Myanmar to the east. In the Indian Ocean, India is in the vicinity of Sri Lanka and the Maldives; its Andaman and Nicobar Islands share a maritime border with Thailand and Indonesia.

Modern humans arrived on the Indian subcontinent from Africa no later than 55,000 years ago.[21] Their long occupation, initially in varying forms of isolation as hunter-gatherers, has made the region highly diverse, second only to Africa in human genetic diversity.[22] Settled life emerged on the subcontinent in the western margins of the Indus river basin 9,000 years ago, evolving gradually into the Indus Valley Civilisation of the third millennium BCE.[23] By 1200 BCE, an archaic form of Sanskrit, an Indo-European language, had diffused into India from the northwest, unfolding as the language of the *Rigveda*, and recording the dawning of Hinduism in India.[24] The Dravidian languages of India were supplanted in the northern regions.[25] By 400 BCE, stratification and exclusion by caste had emerged within Hinduism,[26] and Buddhism and Jainism had arisen, proclaiming social orders unlinked to heredity.[27] Early political consolidations gave rise to the loose-knit Maurya and Gupta

WIKIPEDIA
Die freie Enzyklopädie

Hauptseite
Themenportale
Zufälliger Artikel

Mitmachen
Artikel verbessern
Neuen Artikel anlegen
Autorenportal
Hilfe
Letzte Änderungen
Kontakt
Spenden

Werkzeuge
Links auf diese Seite
Änderungen an verlinkten Seiten
Spezialseiten
Permanenter Link
Seiteninformationen
Wikidata-Datenobjekt
Artikel zitieren

In anderen Projekten

Nicht angemel

Artikel | Diskussion | Lesen | Bearbeiten | Quelltext bearbeiten

Indien

Dieser Artikel behandelt das Land – zu anderen Bedeutungen siehe Indien (Beg

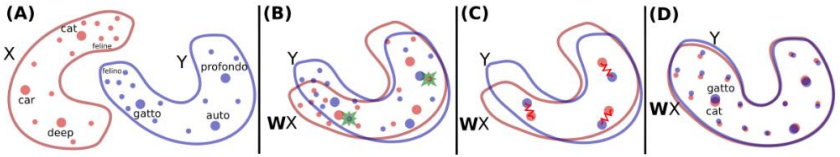
Indien (Aussprache [ˈɪndi̯ən]) ist ein Staat in Südasien, der den größten Teil des indischen Subkontinents umfasst. Indien ist eine Bundesrepublik, die von 29 Bundesstaaten gebildet wird und außerdem sieben bundesunmittelbare Gebiete umfasst. Der Eigenname der Republik lautet in den beiden landesweit gültigen Amtssprachen **Bharat Ganarajya** (Hindi) und **Republic of India** (Englisch). Die moderne demokratische und säkulare indische Republik besteht seit 1949 und sei 1950 gilt die Verfassung Indiens.

Der Himalaya bildet die natürliche Nordgrenze Indiens, im Süden umschließt der Indische Ozean das Staatsgebiet. Indien grenzt an Pakistan, das chinesische Autonome Gebiet Tibet, Nepal, Bhutan, Myanmar (Birma) und Bangladesch. Weitere Nachbarstaaten im Indischen Ozean sind Sri Lanka und die Malediven. Hinsichtlich der Landesfläche ist Indien das flächenmäßig siebtgrößte Land der Erde.

Das Gebiet Indiens ist mindestens seit der bronzezeitlichen Indus-Hochkultur zivilisiert. Der indische Staat ist mit über 1,37 Milliarden Einwohnern (Ende 2018)[6] nach der Volksrepublik China (1,4 Mrd. Ende 2018)[7] das zweitbevölkerungsreichste Land der Erde und somit die bevölkerungsreichste Demokratie der Welt.[8] Bei gleichbleibend hohem Bevölkerungswachstum könnte Indien schon im Jahr 2020 China überholen. Durch fortschreitende

Selbst-überwachte NMT: wie funktioniert das?

Super vereinfacht:



- Lernen word embeddings L1, L2
- Alignieren
- Multilinguale NMT {L1,L2}>{L1,L2} initialisiert mit word embeddings
- Für all Sätze s_{L1} und s_{L2}, SSNMT vergleicht ihre embeddings:
  - Wenn sie ähnlich sind, nutze sie als Trainingsdaten für SSNMT
  - Ansonsten, schau dir das nächste Paar an

Selbst-überwachte NMT: Resultate WMT 2014

Reference	Corpus, *en*+*fr* sent. (in millions)	BLEU *en2fr*	BLEU *fr2en*
Unsupervised NMT			
Artetxe et al. (2018b)	NCr13, 99+32	15.1	15.6
Lample et al. (2018a)	WMT, 16+16	15.1	14.3
Yang et al. (2018)	WMT, 16+16	17.0	15.6
Lample et al. (2018b)	NCr17, 358+69	**25.1**	**24.2**
Our approach			
Self-supervised NMT	**WP, 12+8**	**29.2**	**27.4**
Unsupervised NMT+SMT			
Artetxe et al. (2018a)	NCr13, 99+32	26.2	25.9
Lample et al. (2018b)	NCr17,358+69	27.6	27.7

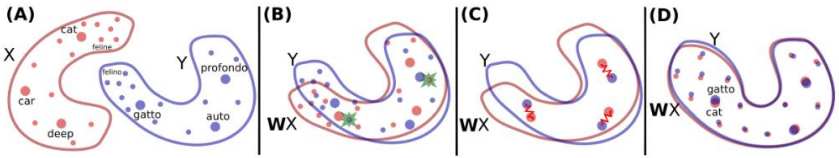
Selbst-überwachte NMT: Resultate WMT 2014

Reference	Corpus, *en+fr* sent. (in millions)	BLEU *en2fr*	BLEU *fr2en*
Unsupervised NMT			
Artetxe et al. (2018b)	NCr13, 99+32	15.1	15.6
Lample et al. (2018a)	WMT, 16+16	15.1	14.3
Yang et al. (2018)	WMT, 16+16	17.0	15.6
Lample et al. (2018b)	NCr17, 358+69	**25.1**	**24.2**
Our approach			
Self-supervised NMT	**WP, 12+8**	**29.2**	**27.4**
Unsupervised NMT+SMT			
Artetxe et al. (2018a)	NCr13, 99+32	26.2	25.9
Lample et al. (2018b)	NCr17,358+69	27.6	27.7
Ren et al. (2019)	NCr, 50+50	29.5	28.9
Artetxe et al. (2019)	NCr13, 99+32	**36.2**	**33.5**

Fazit

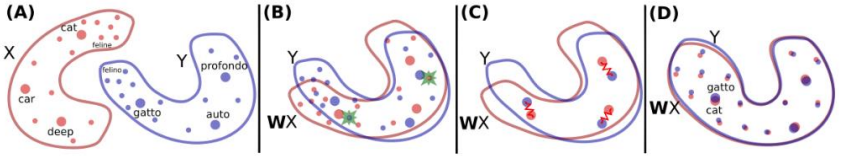
- Überwachte MT braucht parallele Daten
- Oft gibt es die aber nicht
- ☹ ☹ ☹
- Mono-linguale and vergleichbare Daten
- Unüberwachte (N)MT und selbst-überwachte NMT kann aus solchen Daten lernen
- ☺ ☺ ☺

Fazit



- Alles paletti?
- Nein …
- Un- und selbst-überwachte (N)MT basiert stark auf ⤴
- Das funktioniert nicht so gut wenn
  - L1 und L2 unteschiedliche Skripte haben
  - Große Domainunterschiede zwischen L1 und L2 Daten gibt
- Bislang gute (very impressive!) Resultate für Sprachpaare wie EN- FR wo man eigentlich UN- and SS-(N)MT gar nicht braucht …

Fazit



- Un- und selbst-überwachte (N)MT
- Muss sich noch für low-/no-resource Sprachpaare beweisen
- Area of very active research

Fazit

- Un- und selbst-überwachte (N)MT jetzt schon super gut für
  - Domain Adaptation
  - Stil Adaptation (formal, einfach, …)

- Und dafür gibt es oft keine parallelen Trainingsdaten …

ELRC https://lr-coordination.eu/

Data or no Data, that is the Question: Learning MT without Translation Data?

Unüberwachte (N)MT

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. Unsupervised statistical machine translation. **EMNLP 2018**
- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018c. Unsupervised neural machine translation. **ICLR 2018**
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. **ICLR 2018**
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. **EMNLP 2018**
- Many, many more since …!

Self-supervised NMT

- Dana Ruiter, Cristina Espana Bonet and Josef van Genabith. Self-supervised Neural Machine Translation. **ACL-2019**
- Dana Ruiter, Josef van Genabith and Cristina Espana Bonet. Self-Induced Curriculum Learning in Self-Supervised Neural Machine Translation. **EMNLP-2020**.

Translation Examples (Artetxe et al. 2019)

Source	Reference	Artetxe et al. (2018b)	Proposed system
D'autres révélations ont fait état de documents divulgués par Snowden selon lesquels la NSA avait intercepté des données et des communications émanant du téléphone portable de la chancelière allemande Angela Merkel et de ceux de 34 autres chefs d'État.	Other revelations cited documents leaked by Snowden that the NSA monitored German Chancellor Angela Merkel's cellphone and those of up to 34 other world leaders.	Other disclosures have reported documents disclosed by Snowden suggested the NSA had intercepted communications and data from the mobile phone of German Chancellor Angela Merkel and those of 32 other heads of state.	Other revelations have pointed to documents disclosed by Snowden that the NSA had intercepted data and communications emanating from German Chancellor Angela Merkel's mobile phone and those of 34 other heads of state.
La NHTSA n'a pas pu examiner la lettre d'information aux propriétaires en raison de l'arrêt de 16 jours des activités gouvernementales, ce qui a ralenti la croissance des ventes de véhicules en octobre.	NHTSA could not review the owner notification letter due to the 16-day government shutdown, which tempered auto sales growth in October.	The NHTSA could not consider the letter of information to owners because of halting 16-day government activities, which slowed the growth in vehicle sales in October.	NHTSA said it could not examine the letter of information to owners because of the 16-day halt in government operations, which slowed vehicle sales growth in October.
Le M23 est né d'une mutinerie, en avril 2012, d'anciens rebelles, essentiellement tutsi, intégrés dans l'armée en 2009 après un accord de paix.	The M23 was born of an April 2012 mutiny by former rebels, principally Tutsis who were integrated into the army in 2009 following a peace agreement.	M23 began as a mutiny in April 2012, former rebels, mainly Tutsi integrated into the national army in 2009 after a peace deal.	The M23 was born into a mutiny in April 2012, of former rebels, mostly Tutsi, embedded in the army in 2009 after a peace deal.
Tunks a déclaré au Sunday Telegraph de Sydney que toute la famille était «extrêmement préoccupée» du bien-être de sa fille et voulait qu'elle rentre en Australie.	Tunks told Sydney's Sunday Telegraph the whole family was "extremely concerned" about his daughter's welfare and wanted her back in Australia.	Tunks told The Times of London from Sydney that the whole family was "extremely concerned" of the welfare of her daughter and wanted it to go in Australia.	Tunks told the Sunday Telegraph in Sydney that the whole family was "extremely concerned" about her daughter's well-being and wanted her to go into Australia.

Table 4: Randomly chosen translation examples from French→English newstest2014 in comparison of those reported by Artetxe et al. (2018b).